Genet. Res., Camb. (2000), **76**, *pp.* 227–236. *With* 3 *figures.* Printed in the United Kingdom © 2000 Cambridge University Press

Insertions, substitutions, and the origin of microsatellites

YONG ZHU, JOAN E. STRASSMANN AND DAVID C. QUELLER*
Department of Ecology and Evolutionary Biology, Rice University, PO Box 1892, Houston, TX 77251–1892, USA

(*Received 11 October 1999 and in revised form 26 March and 14 June 2000*)

Summary

This paper uses data from the Human Gene Mutation Database to contrast two hypotheses for the origin of short DNA repeats: substitutions and insertions that duplicate adjacent sequences. Because substitutions are much more common than insertions, they are the dominant source of new 2-repeat loci. Insertions are rarer, but over 70% of the 2–4 base insertion mutations are duplications of adjacent sequences, and over half of these generate new repeat regions. Insertions contribute fewer new repeat loci than substitutions, but their relative importance increases rapidly with repeat number so that all new 4–5-repeat mutations come from insertions, as do all 3-repeat mutations of tetranucleotide repeats. This suggests that the process of repeat duplication that dominates microsatellite evolution at high repeat numbers is also important very early in microsatellite evolution. This result sheds light on the puzzle of the origin of short tandem repeats. It also suggests that most short insertion mutations derive from a slippage-like process during replication.

1. Introduction

Microsatellites are tandem repeats of DNA motifs two to five bases long, common in the genomes of eukaryotes and some prokaryotes (Weber, 1990; Field & Wills, 1996). Because of their high levels of polymorphism in numbers of repeats, they have been widely used as markers in studies of kinship, population structure and genetic mapping (e.g. Queller *et al.*, 1993; Estoup *et al.*, 1995; Weissenbach *et al.*, 1992). They are also implicated in a number of human genetic disorders (Sutherland & Richards, 1995). However, we do not yet have a clear understanding of exactly how the strings of repeats that make up microsatellites originate.

Studies of microsatellite mutation and evolution have focused on established microsatellites with multiple repeats. The number of repeats usually increases or decreases by a single repeat unit, though sometimes more (Levinson & Gutman, 1987*a*; Valdes *et al.*, 1993; Kruglyak *et al.*, 1998). The mechanism appears to involve slippage during DNA replication (Schlötterer & Tautz, 1992; observed rates may also reflect efficiency of repair mechanisms (Wierdl *et al.*, 1997)). Slippage is thought to depend on mispairing of tandem repeats during DNA replication (Levinson & Gutman, 1987*b*), so it may not occur when there are few tandem repeats. Three lines of evidence seem to support this possibility. First, direct studies of slippage mutations show that they are more common in loci with longer repeats (Brinkmann *et al.*, 1998). Second, loci with fewer than 5 repeats are rarely polymorphic, as expected if they incur few mutations, and polymorphism levels increase with number of repeats (Weber, 1990; Strassmann *et al.*, 1997; Zhu *et al.*, 2000). Variant repeat units interrupting a string of repeats reduce slippage rates (Petes *et al.*, 1997). Third, while longer strings of repeats occur more often than expected by chance, as expected from high slippage rates, this was reported not to be true for very short repeat sequences below a threshold of about 8 nucleotides (Rose & Falush, 1998; but see Pupko & Graur, 1999).

If slippage is dependent on possession of a few repeats then some process other than slippage must

* Corresponding author. Tel: +1 (713) 348 5220. Fax: +1 (713) 348 5232. E-mail: queller@rice.edu

https://doi.org/10.1017/S001667230000478X

account for the origin and early evolution of repeat loci below this threshold. A reasonable hypothesis is that some threshold number of repeats must be acquired through other kinds of random mutations, such as substitutions, before slippage can occur (Levinson & Gutman, 1987*b*; Stephan & Cho, 1994; Messier *et al*., 1996; Rose & Falush 1998).

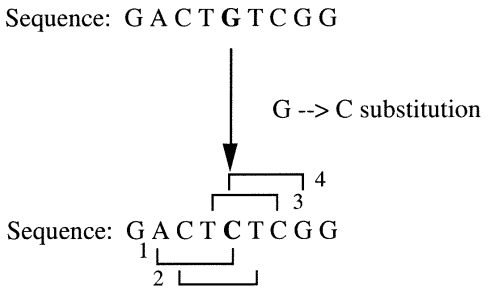
In a phylogenetic study of three wasp microsatellite loci, we observed that short insertions in the flanking, non-repeat regions had a high likelihood of being duplications of adjacent bases (GenBank accession numbers in Zhu *et al*., 2000). This suggested that microsatellites might evolve by a slippage-like mechanism from the very beginning, starting with the duplication of a few bases to form a 2-repeat proto-microsatellite. However, the number of insertions in this wasp dataset was too small to draw any general conclusions, and some events inferred to be insertions could really be deletions if the phylogeny was not correct. Therefore we turned to a different dataset to explore the origins of tandem repeats: the Human Gene Mutation Database (Krawczak & Cooper, 1997).

2. Methods

The mutations compiled in the Human Gene Mutation Database are located in the coding regions of human nuclear genes and cause inherited diseases (Krawczak & Cooper, 1997). This database has two clear advantages for evaluating mutations generating new proto-microsatellites of only two repeats. First, because the wild-type sequence is known (the normal, non-mutated, disease-free state), one can easily distinguish insertions from deletions. Second, the database is large and contains numerous mutations in many genes. At the time of our survey, the database included 88 two-base insertions, 35 three-base insertions, 63 four-base insertions and 9070 substitutions.

The database search engine requires that you first specify a gene of interest, and then specify insertion or substitution mutations. Our procedures differed somewhat for insertions and substitutions, but both involved finding the mutations, finding their flanking sequences, and checking for formation of new repeats.

We examined all the genes in the database to see whether they had any insertion mutations of 2–4 basepairs. We conducted an exhaustive search by using wildcard searches of gene names (e.g. 'gl*' would pull out all genes beginning with the letters gl). The database search engine picked up a maximum of 40 genes from a given search, so we made our abbreviations of gene names more specific if we obtained 40 genes, thus ensuring that we accessed every gene in the database at the time of the study. Because there were so many substitutions in the database, we did not do an exhaustive search, but instead used a sample: the first 1000 we encountered.



Four possible combinations including substitution site for checking dinucleotide repeats:

1. ACT**C**
2. CT**C**T created repeat
3. T**C**TC created repeat
4. **C**TCG

Fig. 1. Example of the different possible reading frames that need to be evaluated to determine whether a substitution creates a new dinucleotide repeat. In this example there are two repeat motifs, 'CT' and 'TC', but we counted only one because they are different ways representing the same repeat locus. There are also six and eight such frames for checking any potential trinucleotide and tetranucleotide repeats, respectively (not shown).

For substitutions, the database included sufficient adjacent sequences for us to check to see whether repeats were created. However, the entries for insertion mutations did not include flanking sequences, so we sought them in the original publications. We used all insertion mutations of 2–4 bases for which we could readily find the reference and identify the appropriate sequence: 31 dinucleotides, 19 trinucleotides and 25 tetranucleotides.

To determine whether substitutions created repeats, we searched the sequences using a simple computer program that identified repeats in any of the 18 possible windows. There are four such windows for dinucleotide repeats, six for trinucleotide repeats and eight for tetranucleotide repeats (Fig. 1). For insertions, we examined the sequences by eye. Insertions of multiple identical bases, such as TT, were also sometimes duplications of an adjacent sequence, but were not counted, as they generated mononucleotide repeats. We counted new repeats as duplications of the adjacent sequence only if the full insertion was duplicated. For example, an ATC insertion was counted as generating new repeats if it was adjacent to another ATC, but not if it created a dinucleotide repeat by being adjacent to another TC. Because we were interested in generation of repeats by

Table 1. *Dinucleotide insertions and their surrounding sequences*

	Gene Symbol/name	Nucleotides affected	Duplication of adjacent dinucleotide? (number duplicated)	References
1	*p67-phox*	CCTTctCTTGG	Yes (1)	Nunoi *et al.* (1995)
2	*beta-spectrin*	CGAagAGAGGTG	Yes (2)	Tse *et al.* (1991)
3	*Hb Agnana*	AACAgtGTGTCACG	Yes (2)	Ristaldi *et al.* (1990)
4	*Na-Cl cotrasporter*	TTgtCTCTG	No	Mastroianni *et al.* (1996)
5	*TSC2*	GCGTatGAGC	No	Au *et al.* (1998)
6	*NF1*	TATaaGCTTCG	No	Colman *et al.* (1997)
7	*DHAPAT*	CAttGTTAT	No	Ofman *et al.* (1998)
8	*PPO*	GGAGagCCCTA	Yes (1)	Lam *et al.* (1997)
9	*CF*	CATCtcTCATTC	Yes (2)	Iannuzzi *et al.* (1991)
10	*hMSH2*	GACtaTTTAC	No	Maliaka *et al.* (1996)
11	*NF1*	AGTTttACTG	No	Ainsworth *et al.* (1993)
12	*SOD-1*	TGAAttAGAA	No	Orrell *et al.* (1997)
13	*DSS*	CATAcgCGT	Yes (1)	Rautenstrauss *et al.* (1994)
14	*APC*	CATAtaGT	Yes (1)	Paffenholz *et al.* (1994)
15	*C4A*	GGCTCtcAGTC	Yes (1)	Barba *et al.* (1993)
16	*APC*	ATTTttA	No	Mandl *et al.* (1994)
17	*COL3A1*	AAttTGTC	No	Richards *et al.* (1994)
18	*IDUA*	TCcaCTTC	No	Bunge *et al.* (1994)
19	*NF2*	AGGAGAgaTCTT	Yes (2)	Mautner *et al.* (1996)
20	*PAX6*	GCCccGTGC	No	Davis & Cowell (1993)
21	*MSH2*	ATAtgTGTACGA	Yes (1)	Nystrom-Lahti *et al.* (1996)
22	*MATA1*	GACTtgCTAA	No	Chamberlin *et al.* (1996)
23	*hMLH1*	GTGCgcACC	Yes (1)	Wijnen *et al.* (1996)
24	*CFTR*	GGATATATatATTC	Yes (4)	White *et al.* (1990)
25	*PAX6*	TACTgaGATCCA	Yes (1)	Jordan *et al.* (1992)
26	*ZIC3*	GGGCttGAGA	No	Gebbia *et al.* (1997)
27	*G6Pase*	CATCatATATGT	Yes (2)	Lei *et al.* (1993)
28	*CD40*	CGTCTCtcCGAC	Yes (2)	Macchi *et al.* (1995)
29	*PDH*	AGttTTTTCC	No	Chun *et al.* (1993)
30	*RB1*	CAGAgaTGT	Yes (1)	Lohmann *et al.* (1994)
31	*HBB*	ACTGtgTGACA	Yes (2)	Ristaldi *et al.* (1990)

duplication of an adjacent sequence, we did not count any insertions that created repeats not due to duplication, as in line 11 of Table 2, where a CGG insertion creates two GCG repeats. Thus, throughout this paper, repeats arising from insertion should be read as repeats arising by an insertion which duplicates an adjacent sequence.

We estimated the total number of each type of mutation from the number of those mutations we found, multiplied by a scaling factor to account for the part of the database not checked. For example, we checked 19 of 35 trinucleotide insertions and found that seven of them generated a new two-repeat microsatellite. So, for this example we estimated that there were $(7/19) \times 35 = 12{\cdot}89$ new 2-repeat trinucleotides from 3 base insertions. Other trinucleotide insertions generated 3 to 5 repeat microsatellites. Continuing the example but for substitutions, 59 of the 1000 substitutions we checked generated new 2-repeat trinucleotide microsatellites. Thus $(59/1000) \times 9070 = 535{\cdot}13$ (59 substitutions of that type, 1000 of 9070 substitutions assessed). The overall fraction of 2-repeat trinucleotides from insertions was therefore $12{\cdot}89/(12{\cdot}89 + 535{\cdot}13) = 0{\cdot}0235$. Similar logic was applied to calculating dinucleotide and tetranucleotide motif repeats of 2 to 5 repeats.

3. Results

Over 70% of the insertions that we examined were duplications of adjacent bases (Tables 1–3; Fig. 2*a*). Specifically, 55% of the dinucleotide insertions 68% of the trinucleotide insertions, and 92% of the tetranucleotide insertions copied adjacent sequences (excluding copies of mononucleotide runs). Some of the duplications were of already existing short repeat sequences of 2–4 units, but over half of the 2- and 3-base duplications, and nearly all the 4-base duplications, had no pre-existing repeat structure. These are identified by '1' in column 4 of Tables 1–3, because there was only one pre-existing copy of the duplicated motif, going to a 2-copy proto-microsatellite after the insertion. We found that 29% of all dinucleotide insertions generated a new 2-copy repeat, 23% added a third repeat to an existing run of 2 repeats, and 3% added a fourth or fifth repeat.

https://doi.org/10.1017/S001667230000478X

Table 2. *Trinucleotide insertions and their surrounding sequences*

	Gene symbol/name	Nucleotides affected	Duplication of adjacent trinucleotide? (number duplicated)	References
1	*NF2*	GATttgTTGGTG	Yes (1)	Ruttledge *et al.* (1996)
2	*CLCN5*	CGAGACCaccGGGATAGGC	Yes (1)	Lloyd *et al.* (1997)
3	*vWF*	GGACATGatgATGGA	Yes (2)	Ribba *et al.* (1991)
4	*vWF*	GTCCCgcgGCGGCGT	Yes (2)	Gaucher *et al.* (1994)
5	*ALDP*	GAGGtggTGGTGGTGGCC	Yes (3)	Feigenbaum *et al.* (1996)
6	*XPCC*	AGTggtGGTGAG	Yes (1)	Li *et al.* (1993)
7	*CYP11B1*	ATGCTGCTGCTGCTGctgCACCAT	Yes (4)	Geley *et al.* (1996)
8	*Fibrillin*	CAACCAccaAGCAAC	Yes (1)	Milewicz & Duvic. (1994)
9	*C1-inhibitor*	ACTGtgtGGGTGGAG	No	Siddique *et al.* (1993)
10	*VHL*	TAACGtctTCTTCTA	Yes (2)	Glavac *et al.* (1996)
11	*LCAT*	CCGcggCGC	No	Gotoda *et al.* (1991)
12	*AAP*	AGGCGGCGgcgGCGGCC	Yes (3)	Holmes *et al.* (1987)
13	*DHPT*	CCGCTA ctaCCAA	Yes (1)	Howells *et al.* (1990)
14	*SPTA*	AGTTGttgCTGCGG	Yes (1)	Roux *et al.* (1989)
15	*HPRT1*	ATG gcaCAG ACT	No	Sege-Peterson *et al.* (1992)
16	*ALDP*	AAG aatGGG	No	Krasemann *et al.* (1996)
17	*PKLR*	TGC agcATC	No	Lenzner *et al.* (1994)
18	*CFTR*	CTC ctaCTA CAC	Yes (1)	Dörk *et al.* (1997)
19	*HBA*	CCCCgaaACCA	No	Moo-Penn *et al.* (1989)

Table 3. *Tetranucleotide insertions and their surrounding sequences*

	Gene symbol/name	Nucleotides affected	Duplication of adjacent tetranucleotide? (number duplicated)	References
1	*CD40*	TCATAAAtaaaCTT	Yes (1)	Macchi *et al.* (1995)
2	*APC*	AattcTG	No	Olschwang *et al.* (1993)
3	*LDH-B*	TGGACATTcattCTTA	Yes (1)	Maekawa *et al.* (1994)
4	*CYP17*	ACCCctacCTACGG	Yes (1)	Kagimoto *et al.* (1989)
5	*PDH*	TGactaACTAACCG	Yes (1)	Chun *et al.* (1993)
6	*E1*	GTTTAAGTaagtCAGT	Yes (1)	Naito *et al.* (1994)
7	*RB1*	GTATTGTTTGtttgCACT	Yes (1)	Lohmann *et al.* (1994)
8	*ND*	CGTAGGtaggAA	Yes (1)	Berger *et al.* (1992)
9	*Androgen receptor*	GAAGcctaCCTATG	Yes (1)	Batch *et al.* (1992)
10	*TPO*	AGACGGCCggccGCGC	Yes (1)	Abramowicz *et al.* (1992)
11	*Beta-hexosaminidase*	TATCctatCTATAT	Yes (1)	Myerowitz & Costigan (1988)
12	*DAX-1*	GGATggatGACG	Yes (1)	Habiby *et al.* (1996)
13	*AT*	AGTACCGaccgCTGT	Yes (1)	Emmerich *et al.* (1994)
14	*APRT*	CGAAagccAGCCTACT	Yes (1)	Kamatani *et al.* (1992)
15	*TSC2*	CCTACTtactCCCT	Yes (1)	Yates *et al.* (1997)
16	*Transglutaminase*	AGTACGACcgacG	Yes (1)	Bichakjian *et al.* (1998)
17	*LDL*	TACaagaAAGAATT	Yes (1)	Lehrman *et al.* (1985)
18	*FBN1*	ACAacttACTTATT	Yes (1)	Dietz *et al.* (1993)
19	*APECED*	ACAGGcaggCAGGCC	Yes (2)	Aaltonen *et al.* (1997)
20	*SRY*	TTgtagGTAGCT	Yes (1)	Foster *et al.* (1994)
21	*HLA*	ATGACTGactgGG	Yes (1)	Pshezhetsky *et al.* (1997)
22	*PEX*	TGTtagtGAGAA	No	Chang *et al.* (1997)
23	*TTP*	CCAgtaaGTAAGA	Yes (1)	Ouahchi *et al.* (1995)
24	*F9*	CTGGATTgattAAGG	Yes (1)	Bottema *et al.* (1989)
25	*HEXA*	TATATCtatcCTAT	Yes (1)	Myerowitz & Costigan (1988)

Thirty-seven per cent of all trinucleotide insertions generated a new 2-copy repeat, 16% added a third repeat to an existing run of 2 repeats, and 16% added a fourth or fifth repeat. Finally, 88% of all tetranucleotide insertions generated a new, 2-copy repeat and 4% added a third repeat to an existing run of 2

https://doi.org/10.1017/S001667230000478X

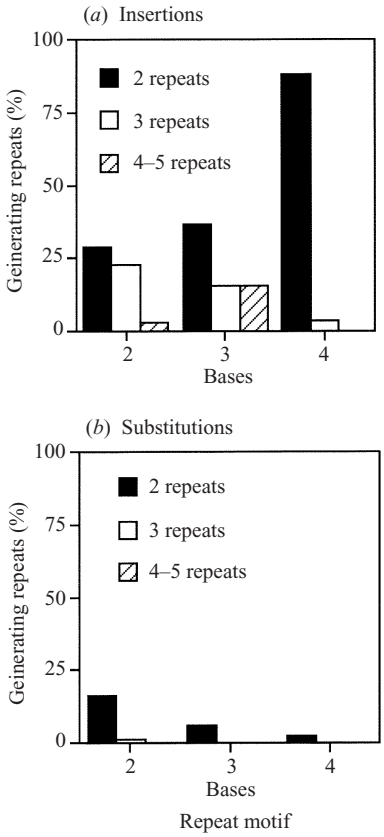


Fig. 2. Percentage of (*a*) insertion mutations and (*b*) substitution mutations generating microsatellites of 2–5 tandem repeats. Total repeats generated is obtained by adding together the bars for repeats of each specified length.

repeats. Thus insertions are generally copies of adjacent sequence, and generate proto-microsatellites, or short microsatellites.

New proto-microsatellites were also generated by substitutions. The total number of substitutions in the database was much larger than the number of 2–4 basepair insertions (9070 vs 186). However, a relatively low percentage of these substitutions generated new repeats and very few generated runs of more than 2 repeats (Fig. 2*b*). We found that 16%, of substitutions generated a new 2-copy dinucleotide repeat. One per cent of substitutions added a third repeat to an existing run of 2 dinucleotide repeats and no substitutions generated a longer run of dinucleotide repeats. Six per cent of substitutions generated a new 2-copy trinucleotide repeat, 0·2% of substitutions added a third repeat to an existing run of 2 trinucleotide repeats, and no substitutions generated a longer run of trinucleotide repeats. Three per cent of substitutions generated a new 2-copy repeat of a tetranucleotide repeat and no substitutions generated any longer run of tetranucleotide repeats (Fig. 2*b*).

We could determine the relative numbers of short repeat loci generated by substitutions versus insertions

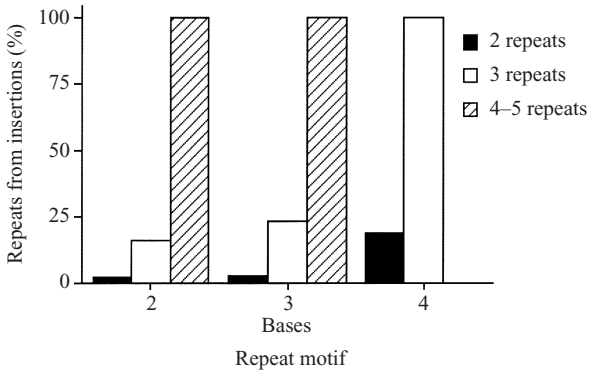


Fig. 3. Percentage of small repeat loci (2–5 repeats) in the mutation database that arise from insertions that are duplications of an adjacent sequence rather than from substitutions.

if we assume that these two are represented in the database in proportions similar to their occurrence across the genome. We estimated that insertions generated a minority of new 2-repeat loci in the database: 1·7% of dinucleotides, 2·4% of trinucleotides and 18·5% of tetranucleotides (Fig. 3). Though insertions occur less frequently than substitutions, their relative importance in generating new repeats rapidly increases with the length of the repeat. For mutations increasing the number of repeats from 2 to 3, 16·2% were insertions in the dinucleotide class, 23·4% in the trinucleotide class, as was the only recorded mutation to a third tetranucleotide repeat (Fig. 3). All recorded mutations generating a fourth or fifth repeat were insertions (Fig. 3).

4. Discussion

Over 70% of all 2–4 base insertions consist of copies of existing sequences, and generate runs of 2–5 repeats. The majority of these were not extensions of pre-existing repeats, but instead generated a short repeat region where none existed before. This result indicates that the kinds of processes that lead to expansion and polymorphism at established microsatellite loci also occur with few or no repeats. The mechanism is not clear. Slippage is generally thought to require repeats, with repeats in the new strand mispairing with other repeats on the template during DNA replication (Levinson & Gutman, 1987*b*), but this is not possible in the absence of repeats. However, we did not find evidence for two other proposed mechanisms of insertional mutation that might generate repeats: mispairing of inverted repeats (e.g. ATACC/GGTAT) (Ohshima *et al*., 1992) or symmetric elements (Cooper & Krawczak, 1993).

Messier *et al*. (1996) suggested that a there may be minimum number of repeats that must be generated

https://doi.org/10.1017/S001667230000478X

by substitution before expansion by slippage can occur. They offered support from a primate phylogenetic study of a short microsatellite sequence. However, there is an alternative reading of this history that involves only slippage events, without any enabling substitutions (Gordon, 1997). Even if the interpretation of Messier *et al.* (1996) is correct, this is a single piece of anecdotal evidence, and it can be opposed by other anecdotal phylogenetic evidence showing expansion at very low repeat numbers (e.g. $(AG)_2$ to $(AG)_3$; Primmer & Ellegren, 1998).

Both slippage mutations (Brinkmann *et al.*, 1998) and repeat number polymorphisms (Weber, 1990) are more common at higher repeat numbers. But this need not imply that slippage is either absent or unimportant at lower repeat numbers. Studies of mutation at microsatellite loci have not considered loci with few repeats, and would have to be carried out on a much larger scale to do so. For example, the study of Brinkmann *et al.* (1998) found 23 microsatellite slippage mutations in over 10000 meioses, using nine loci with mean repeat numbers ranging from 6 to 15. A study of this size would not be very useful for detecting slippage at the smallest repeat numbers if the mutation rates are one or more orders of magnitude lower.

Similar considerations apply to studies of polymorphisms. It is clear that polymorphism increases with repeat number (Weber, 1990), but few studies examine loci with very few repeats. Strassmann *et al.* (1997) confirmed this general pattern, including a few trinucleotide loci with 3–4 repeats, only one of which was slightly polymorphic. The observation of lower polymorphism supports the inference that mutation rates are lower (or repair rates higher) at low repeat numbers. However, slippage with few or no repeats could be much less frequent than slippage with many repeats, but still be frequent enough to be important in generating new microsatellites. In short, our finding that slippage (or some mutation process with the same effect) takes place even in the absence of repeats is not inconsistent with earlier studies of mutations and polymorphism.

Rose & Falush (1998) compared observed and expected numbers of microsatellites of various lengths in the yeast genome. They found that long stretches of repeats were more common than expected by chance, which they attributed to duplication by slippage. They also reported that very short stretches of repeats, at or below an 8-nucleotide threshold (2 tetranucleotide, 4 dinucleotide or 8 mononucleotide repeats), were not more common than expected by chance. This would seem to imply that slippage is not important below this threshold. However, Pupko & Graur (1999), also using the yeast genome, found that even 2-repeat microsatellites were observed more often than expected, and that the observed excess was more or less of the size expected by extrapolating from longer repeats. These results are more in line with ours, suggesting that there is no repeat number threshold, but only a continuous change in mutation rates. The reason for the discrepancy between the two yeast studies is not clear, though the two studies used different methods of calculating expected frequencies. If slippage contributes only a minority of new microsatellites, as suggested by our data (Fig. 3), then it is not surprising that small differences in assumptions may lead to differences in results. Another contributing factor may be the fact that 70% of the yeast genome is coding sequence where insertions of 1, 2, 4 or 5 bases would cause reading frameshifts and would therefore rarely persist long enough to be sampled. So 70% of the data may be essentially noise, lowering the power of these studies to detect weak effects.

Our data also come from coding regions. One limitation of using a gene database is that we will miss those microsatellites that originate from polyA tails of retroposons, a process that appears to be important in mammals (Arcot *et al.*, 1995; Nadir *et al.*, 1996), but not in birds (Primmer *et al.*, 1997). Another disadvantage of using the Human Gene Mutation Database is that the mutations in the database generally have deleterious phenotypic effects. This could lead to various biases, though the direction of such biases is not always clear. For example, dinucleotide and tetranucleotide insertions would cause frameshifts in coding regions, causing more severe effects than non-frameshift trinucleotide insertions. This could cause the frameshift mutation classes to be either relatively over-represented because they are more likely to have detectable effects, or under-represented if they often cause early lethality. Similarly, frameshift mutations may be over-represented or under-represented compared with substitutions, and this would alter our quantitative comparisons in Fig. 3. On this score, it is somewhat reassuring that the dinucleotides, which cause frameshifts, show patterns rather similar to the trinucleotides, which do not (Fig. 2, 3). Clearly it would be desirable to confirm our results with sources of data free of such biases. However, at least one important conclusion seems unlikely to be affected. It is difficult to imagine any biases in the Human Gene Mutation Database would inflate the frequency of insertions that are copies of adjacent sequences.

While these results need to be supplemented by studies of non-genic DNA, and also by studies of other organisms, our data suggest that duplication of entire repeats is important in the origin and early evolution of microsatellites. The rarity of repeat-length polymorphisms in microsatellites with few repeats does not refute slippage; it only shows that the rate is lower than the very high rates that characterize

https://doi.org/10.1017/S001667230000478X

longer microsatellites. Our data also suggest that some new 2-repeat microsatellites arise from a mutational mechanism that has the same effect as slippage, the duplication of an adjacent sequence. The importance of this process increases rapidly with repeat number, but there does not appear to be any fixed repeat-number threshold that must be surpassed before slippage can occur.

We thank Barry W. Sullender and Marek Kimmel for helpful discussion and review of the manuscript, Laura Holt for help with database searching, Yujuan Guo for computer programming and Peter D. Stenson for advice on the mutation database. This work was partially financially supported by the US National Science Foundation and the W. M. Keck Center for Computational Biology.

References

Aaltonen, J., Björses, P., Perheentupa, J., Horelli-Kuitunen, N., Palotie, A., Lee, Y. S., Francis, F., Henning, S., Thiel, C., Lethrach, H. & Yaspo, M. (1997). An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. *Nature Genetics* **17**, 399–403.

Abramowicz, M. J., Targovnik, H. M., Varela, V., Cochaux, P., Krawiec, L., Pisarev, M. A., Propato, F. V., Juvenal, G., Chester, H. A. & Vassart, G. (1992). Identification of a mutation in the coding sequence of the human thyroid peroxidase gene causing congenital goiter. *Journal of Clinical Investigation* **90**, 1200–1204.

Ainsworth, P. J., Rodenhiser, D. I & Costa, M. T. (1993). Identification and characterization of sporadic and inherited mutations in exon 31 of the neurofibromatosis (NF1) gene. *Human Genetics* **91**, 151–156.

Arcot, S. S., Wang, Z., Weber, J. L., Deininger, P. L. & Batzer, M. A. (1995). Alu repeats: a source for the genesis of primate microsatellites. *Genomics* **29**, 136–144.

Au, K. S., Rodriguez, J. A., Finch, J. L., Volcik, K. A., Roach, E. S., Delgado, M. R., Rodriguez, E. Jr. & Northrup, H. (1998). Germ-line mutational analysis of the TSC2 gene in 90 tuberous-sclerosis patients. *American Journal of Human Genetics* **62**, 286–294.

Barba, G., Rittner, C. & Schneider, P. M. (1993). Genetic basis of human complement C4A deficiency: detection of a point mutation leading to nonexpression. *Journal of Clinical Investigation* **91**, 1681–1686.

Batch, J. A., Williams, D. M., Davies, H. R., Brown, B. D., Evans, B. A., Hughes, I. A. & Patterson, M. N. (1992). Androgen receptor gene mutations identified by SSCP in fourteen subjects with androgen insensitivity syndrome. *Human Molecular Genetics* **1**, 497–503.

Berger, W., van de Pol, D., Warburg, M., Gal, A., Bleeker-Wagemakers, L., de Silva, H., Meindl, A., Meitinger, T., Cremers, F. & Ropers, H. H. (1992). Mutations in the candidate gene for Norrie disease. *Human Molecular Genetics* **1**, 461–465.

Bichakjian, C. K., Nair, R. P., Wu, W. W., Goldberg, S. & Elder, J. T. (1998). Prenatal exclusion of lamellar ichthyosis based on identication of two new mutations in the transglutaminase 1 gene. *Journal of Investigative Dermatology* **110**,179–182.

Bottema, C. D., Ketterling, R. P., Cho, H. I. & Sommer, S. S. (1989). Hemophilia B in a male with a four-base insertion that arose in the germline of his mother. *Nucleic Acids Research* **17**, 10139.

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J. & Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *American Journal of Human Genetics* **62**, 1408–1415.

Bunge, S., Kleijer, W. J., Steglich, C., Beck, M., Zuther, C. Morris, C. P., Schwinger, E., Hopwood, J. J., Scott, H. S. & Gal, A. (1994). Mucopolysaccharidosis type 1: Identification of 8 novel mutations and determination of the frequency of the two common alpha-L-iduronidase mutations (W402X and Q70X) among European patients. *Human Molecular Genetics* **3**, 861–866.

Chamberlin, M. E., Ubagai, T., Mudd, S. H., Wilson, W. G., Leonard, J. V. & Chou, J. Y. (1996). Demyelination of the brain is associated with methionine adenosyltransferase I/II deficiency. *Journal of Clinical Investigation* **98**, 1021–1027.

Chang, C. C., Lee, W. H., Moser, H., Valle, D. & Gould, S. J. (1997). Isolation of the human PEX12 gene, mutated in group 3 of the peroxisome biogenesis disorders. *Nature Genetics* **15**, 385–388.

Chun, K., MacKay, N., Petrova-Benedict, R. & Robinson, B. H. (1993). Mutations in the X-linked E1 alpha subunit of pyruvate dehydrogenase leading to deficiency of the pyruvate dehydrogenase complex. *Human Molecular Genetics* **2**, 449–454.

Colman, S. D., Abernathy, C. R., Ho, V. T. & Wallace, M. R. (1997). Four frameshift mutations in neurofibromatosis type 1 caused by small insertions. *Journal of Medical Genetics* **34,** 579–581.

Cooper, D. N. & Krawczak, M. (1993). *Human Gene Mutation*, pp. 209–217. Oxford: Bios Scientific Publishers.

Davis, A. & Cowell, J. K. (1993). Mutations in the PAX6 gene in patients with hereditary aniridia. *Human Molecular Genetics* **2**, 2093–2097.

Dietz, H. C., McIntosh, I., Sakai, L. Y., Corson, G. M., Chalberg, S. C., Pyeritz, R. E. & Francomano, C. A. (1993). Four novel FBN1 mutations: significance for mutant transcript level and EGF-like domain calcium binding in the pathogenesis of Marfan syndrome. *Genomics* **17**, 468–475.

Dörk. T., Dworniczak, B., Aulehla-Scholz, C., Wieczorek, D., Bohm, I., Mayerova A., Seydewitz, H. H., Nieschlag, E., Meschede, D., Horst, J., Pander, H., Sperling, H., Ratjen, F., Passarge, E., Schmidtke, J. & Stuhrmann, M. (1997). Distinct spectrum of CFTR gene mutations in congenital absence of vas deferens. *Human Genetics* **100**, 365–377.

Emmerich, J., Chadeuf, G., Alhenc-Gelas, M., Gouault-Heilman, M., Toulon, P., Fiessinger, J. N. & Aiach, M. (1994). Molecular basis of antithrombin type I deficiency: the first large in-frame deletion and two novel mutations in exon 6. *Thrombosis and Haemostasis* **72**, 534–539.

Estoup, A., Garnery, L., Solignac, M. & Cornuet, J. M. (1995). Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation model. *Genetics* **140**, 679–695.

Feigenbaum, V., Lombard-Platet, G., Guidoux, S., Sarde, C. O., Mandel, J. L. & Aubourg, P. (1996). Mutational and protein analysis of patients and heterozygous women with X-linked adrenoleukodystrophy. *American Journal of Human Genetics* **58**, 1135–1144.

Field, D. & Wills, C. (1996). Long, polymorphic microsatellites in simple organisms. *Proceedings of the Royal Society of London*, *Series B* **263**, 209–215.

Foster, J. W., Dominguez-Steglich, M. A., Guioli, S., Kowk,

G., Weller, P. A., Stevanovic, M., Weissenbach, J., Mansour, S., Young, I. D., Goodfellow, P. N., Brook, J. D. & Schafer, A. J. (1994). Campomelic dysplasia and autosomal sex reversal caused by mutations in an SRY-related gene. *Nature* **372**, 525–530.

Gaucher, C., Dieval, J. & Mazurier, C. (1994). Characterization of von Willebrand factor gene defects in two unrelated patients with type IIC von Willebrand disease. *Blood* **84**, 1024–1030.

Gebbia, M., Ferrero, G. B., Pilia, G., Bassi, M. T., Aylsworth, A., Penman-Splitt, M., Bird, L. M., Bamforth, J. S., Burn, J., Schlessinger, D., Nelson, D. L. & Casey, B. (1997). X-linked situs abnormalities result from mutations in ZIC3. *Nature Genetics* 17, 305–308.

Geley, S., Kapelari, K., Johrer, K., Peter, M., Glatzl, J., Vierhapper, H., Schwarz, S., Helmberg, A., Sippell, W. G., White, P. C. & Kofler, R. (1996). CYP11B1 mutations causing congenital adrenal hyperplasia due to 11 beta-hydroxylase deficiency. *Journal of Clinical Endocrinology and Metabolism* **81**, 2896–2901.

Glavac, D., Neumann, H. P. H., Wittke, C., Jaenig, H., Masek, O., Streicher, T., Pausch, F., Engelhardt, D., Plate, K. H., Höfler, H., Chen, F., Zbar, B. & Brauch, H. (1996). Mutations in the VHL tumor suppressor gene and associate lesions in families with von Hippel–Lindau disease from central Europe. *Human Genetics* **98**, 271–280.

Gordon, A. J. E. (1997). Microsatellite birth register. *Journal of Molecular Evolution* **45**, 337–338.

Gotoda, T., Yamada, N., Murase, T., Sakuma, M., Murayama, N., Shimano, H., Kozaki, K., Albers, J. J., Yazaki, Y. & Akanuma, Y. (1991). Differential phenotypic expression by three mutant alleles in familial lecithin:cholesterol acyltransferase deficiency. *Lancet* **338**, 778–781.

Habiby, R. L., Boepple, P., Nachtigall, L., Sluss, P. M., Crowley, W. F. Jr & Jameson, J. L. (1996). Adrenal hypoplasia congenita with hypogonadotropic hypogonadism: evidence that DAX-1 mutations lead to combined hypothalamic and pituitary defects in gonadotropin production. *Journal of Clinical Investigation* **98**, 1055–1062.

Holmes, W. E., Lijnen, H. R., Nelles, L., Kluft, C., Nieuwenhuis, H. K., Rijken, D. C. & Collen, D. (1987). Alpha 2-antiplasmin Enschede: alanine insertion and abolition of plasmin inhibitory activity. *Science* **238**, 209–211.

Howells, D. W., Forrest, S. M., Dahl, H. H. & Cotton, R. G. (1990). Insertion of an extra codon for threonine is a cause of dihydropteridine reductase deficiency. *American Journal of Human Genetics* **47**, 279–285.

Iannuzzi, M. C., Stern, R. C., Collins, F. S., Hon, C. T., Hidaka, N., Strong, T., Becker, L., Drumm, M. L., White, M. B., Gerrard, B. & Dean, M. (1991). Two frameshift mutations in the cystic fibrosis gene. *American Journal of Human Genetics* **48**, 227–231.

Jordan, T., Hanson, I., Zaletayev, D., Hodgson, S., Prosser, J., Seawright, A., Hastie, N., & van Heyningen, V. (1992). The human PAX6 gene is mutated in two patients with aniridia. *Nature Genetics* **1**, 328–332.

Kagimoto, K., Waterman, M. R., Kagimoto, M., Ferreira, P., Simpson, E. R. & Winter, J. S. (1989). Identification of a common molecular basis for combined 17 alpha-hydroxylase/17,20-lyase deficiency in two Mennonite families. *Human Genetics* **82**, 285–286.

Kamatani, N., Hakoda, M., Otsuka, S., Yoshikawa, H. & Kashiwazaki, S. (1992). Only three mutations account for almost all defective alleles causing adenine phosphoribosyltransferase deficiency in Japanese patients. *Journal of Clinical Investigation* 90, 130–135.

Krasemann, E. W., Meier, V., Korenke, G. C., Hunneman, D. H. & Hanefeld, F. (1996). Identification of mutations in the ALD-gene of 20 families with adrenoleukodystrophy/adrenomyeloneuropathy. *Human Genetics* **97**, 194–197.

Krawczak, M. & Cooper, D. N. (1997). Human Gene Mutation Database. *Trends in Genetics* **13**, 121–122. See http://www.uwcm.ac.uk/uwcm/mg/new_back.html

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. *Proceedings of the National Acadamy of Sciences of the USA* **95**, 10774–10778.

Lam, H., Dragan, L., Tsou, H. C., Merk, H., Peacocke, M., Goerz, G., Sassa, S., Poh-Fitzpatrick, M., Bickers, D. R. & Christiano, A. M. (1997). Molecular basis of variegate porphyria: a *de novo* insertion mutation in the protoporphyrinogen oxidase gene. *Human Genetics* **99**, 126–129.

Lehrman, M. A., Goldstein, J. L., Brown, M. S., Russell, D. W. & Schneider, W. J. (1985). Internalization-defective LDL receptors produced by genes with nonsense and frameshift mutations that truncate the cytoplasmic domain. *Cell* **41**, 735–743.

Lei, K. J., Shelly, L. L., Pan, C. J., Sidbury, J. B. & Chou, J. Y. (1993). Mutations in the glucose-6-phosphatase gene that cause glycogen storage disease type 1a. *Science* **262**, 580–583.

Lenzner, C., Nurnberg, P., Thiele, B. J., Reis, A., Brabec, V., Sakalova, A. & Jacobasch, G. (1994). Mutations in the pyruvate kinase L gene in patients with hereditary hemolytic anemia. *Blood* **83**, 2817–2822.

Levinson, G. & Gutman, G. A. (1987*a*). High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. *Nucleic Acids Research* **15**, 5323–5338.

Levinson, G. & Gutman, G. A. (1987*b*). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* **4**, 203–221.

Li, L., Bales, E. S., Peterson, C. A. & Legerski, R. J. (1993). Characterization of molecular defects in xeroderma pigmentosum group C. *Nature Genetics* **5**, 413–417.

Lloyd, S. E., Gunther, W., Pearce, S. H., Thomson, A., Bianchi, M. L., Bosio, M., Craig, I. W., Fisher, S. E., Scheinman, S. J., Wrong, O., Jentsch, T. J. & Thakker, R. V. (1997). Characterisation of renal chloride channel, CLCN5, mutations in hypercalciuric nephrolithiasis (kidney stones) disorder. *Human Molecular Genetics* **6**, 1233–1239.

Lohmann, D. R., Brandt, B., Hopping, W., Passarge, E. & Horsthemke, B. (1994). Spectrum of small length germline mutations in the RB1 gene. *Human Molecular Genetics* **3**, 2187–2193.

Macchi, P., Villa, A., Strina, A., Sacco, M. G., Morali, F., Brugnoni, D., Giliani, S., Mantuano, E., Fasth, A. & Andersson, B, Zegers, B. J. M., Cavagni, G., Reznick, I., Levy, J., Zan-Bar, I., Porat, Y., Airo, P., Plebani, A., Vezzoni, P. & Notarangelo, L. D. (1995). Characterization of nine novel mutations in the CD40 ligand gene in patients with X-linked hyper IgM syndrome of various ancestry. *American Journal of Human Genetics* **56**, 898–906.

Maekawa, M., Sudo, K., Nagura, K., Li, S. S. & Kanno, T. (1994). Population screening of lactate dehydrogenase deficiencies in Fukuoka Prefecture in Japan and molecular characterization of these three independent mutations in

the lactate dehydrogenase-B(H) gene. *Human Genetics* **93**, 74–76.

Maliaka, Y. K., Chudina, A. P., Belev, N. F., Alday, P., Bochkov, N. P. & Buerstedde, J. M. (1996). CpG dinucleotides in the hMSH2 and hMLH1 genes are hotspots for HNPCC mutations. *Human Genetics* **97**, 251–255.

Mandl, M., Paffenholz, R., Friedl, W., Caspari, R., Sengteller, M. & Propping, P. (1994). Frequency of common and novel inactivating APC mutations in 202 families with familial adenomatous polyposis. *Human Molecular Genetics* **3**, 181–184.

Mastroianni, N., Bettinelli, A., Bianchetti, M., Colussi, G., De Fusco, M., Sereni, F., Ballabio, A. & Casari, G. (1996). Novel molecular variants of the Na–Cl cotransporter gene are responsible for Gitelman syndrome. *American Journal of Human Genetics* **59**, 1019–1026.

Mautner, V. F., Baser, M. E. & Kluwe, L. (1996). Phenotypic variability in two families with novel splice-site and frameshift NF2 mutations. *Human Genetics* **98**, 203–206.

Messier, W., Li, S. H. & Stewart, C. B. (1996). The birth of microsatellites. *Nature* **381**, 483.

Milewicz, D. M. & Duvic, M. (1994). Severe neonatal Marfan syndrome resulting from a *de novo* 3-bp insertion into the fibrillin gene on chromosome 15. *American Journal of Human Genetics* **54**, 447–453.

Moo-Penn, W. F., Swan, D. C., Hine, T. K., Baine, R. M., Jue, D. L., Benson, J. M., Virshup, D. M. & Zinkham, W. H. (1989). Hb Catonsville (glutamic acid inserted between Pro-37(C2)alpha and Thr-38(C3)alpha). Non-allelic gene conversion in the globin system? *Journal of Biological Chemistry* **264**, 21454–21457.

Myerowitz, Z. & Costigan, F. C. (1988). The major defect in Ashkenazi Jews with Tay–Sachs disease is an insertion in the gene for the alpha-chain of beta-hexosaminidase. *Journal of Biological Chemistry* **263**, 18587–18589.

Nadir, E., Margalit, H., Gallily, T. & Ben-Sasson, S. A. (1996). Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. *Proceedings of National Acadamy of Sciences of the USA* **93**, 6470–6475.

Naito, E., Ito, M., Yokota, I., Matsuda, J., Yara, A. & Kuroda, Y. (1994). Pyruvate dehydrogenase deficiency caused by a four-nucleotide insertion in the E1 alpha subunit gene. *Human Molecular Genetics* **3**, 1193–1194.

Nunoi, H., Iwata, M., Tatsuzawa, S., Onoe, Y., Shimizu, S., Kanegasaki, S. & Matsuda, I. (1995). AG dinucleotide insertion in a patient with chronic granulomatous disease lacking cytosolic 67-kD protein. *Blood* **86**, 329–333.

Nystrom-Lahti, M., Wu, Y., Moisio, A. L., Hofstra, R. M., Osinga, J., Mecklin, J. P., Jarvinen, H. J., Leisti, J., Buys, C. H., de la Chapelle, A. & Peltomaki, P. (1996). DNA mismatch repair gene mutations in 55 kindreds with verified or putative hereditary non-polyposis colorectal cancer. *Human Molecular Genetics* **5**, 763–769.

Ofman, R., Hettema, E. H., Hogenhout, E. M., Caruso, U., Muijsers, A. O. & Wanders, R. J. A. (1998). Acyl-CoA: dihydroxyacetonephosphate acyltransferase: Cloning of the human cDNA and resolution of the molecular basis in rhizomelic chondrodysplasia punctata type 2. *Human Molecular Genetics* **7**, 847–853.

Ohshima, A., Inouye, S. & Inouye, M. (1992). *In vivo* duplication of genetic elements by the formation of stem-loop DNA without an RNA intermediate. *Proceedings of the National Academy of Sciences of the USA* **89**, 1016–1020.

Olschwang, S., Laurent-Puig, P., Groden, J., White, R. & Thomas, G. (1993). Germ-line mutations in the first 14 exons of the adenomatous polyposis coli (APC) gene. *American Journal of Human Genetics* **52**, 273–279.

Orrell, R. W., Habgood, J. J., Gardiner, I., King, A. W., Bowe, F. A., Hallewell, R. A., Marklund, S. L., Greenwood, J., Lane, R. J. & deBelleroche, J. (1997). Clinical and functional investigation of 10 missense mutations and a novel frameshift insertion mutation of the gene for copper–zinc superoxide dismutase in UK families with amyotrophic lateral sclerosis. *Neurology* **48**, 746-751.

Ouahchi, K., Arita, M., Kayden, H., Hentati, F., Ben Hamida, M., Sokol, R., Arai, H., Inoue, K., Mandel, J. L. & Koenig, M. (1995). Ataxia with isolated vitamin E deficiency is caused by mutations in the alpha-tocopherol transfer protein. *Nature Genetics* **9**, 141–145.

Paffenholz, R., Mandl, M., Caspari, R., Sengteller, M., Propping, P. & Friedl, W. (1994). Eleven novel germline mutations in the adenomatous polyposis coli (APC) gene. *Human Molecular Genetics* **3**, 1703–1704.

Petes, T. D., Greenwell, P. W. & Dominska, M. (1997). Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae*. *Genetics* **146**, 491-498.

Primmer, C. R. & Ellegren, H. (1998). Patterns of molecular evolution in avian microsatellites. *Molecular Biology and Evolution* **15**, 997–1008.

Primmer, C. R., Raudsepp, T., Chowdhary, B. P., Moller, A. P. & Ellegren, H. (1997). Low frequency of microsatellite in the avian genome. *Genome Research* **7**, 471–482.

Pshezhetsky, A. V., Richard, C., Michaud, L., Igdoura, S., Wang, S., Elsliger, M. A., Qu, J., Leclerc, D., Gravel, R., Dallaire, L. & Potier, M. (1997). Cloning expression and chromosomal mapping of human lysosomal sialidase and characterization of mutations in sialidosis. *Nature Genetics* **15**, 316–320.

Pupko, T. & Graur, D. (1999). Evolution of microsatellites in the yeast *Saccharomyces cerevisiae*: role of length and number of repeated units. *Journal of Molecular Evolution* **48**, 313–316.

Queller, D. C., Strassmann, J. E. & Hughes, C. R. (1993). Microsatellites and kinship. *Trends in Ecology and Evolution* 8, 285–288.

Rautenstrauss, B., Nelis, E., Grehl, H., Pfeiffer, R. A. & Van Broeckhoven, C. (1994). Identification of a *de novo* insertional mutation in P0 in a patient with a Dejerine–Sottas syndrome (DSS) phenotype. *Human Molecular Genetics* **3**, 1701–1702.

Ribba, A. S., Lavergne J. M., Bahnak, B. R., Derlon, A., Pietu, G. & Meyer, D. (1991). Duplication of a methionine within the glycoprotein Ib binding domain of von Willebrand factor detected by denaturing gradient gel electrophoresis in a patient with type IIB von Willebrand disease. *Blood* **78**, 1738–1743.

Richards, A. J., Narcisi, P., Ferguson, C., Cobben, J. M. & Pope, F. M. (1994). Two new mutations affecting the donor splice site of COL3A1 IVS37 and causing skipping of exon 37 patients with Ehlers–Danlos syndrome type IV. *Human Molecular Genetics* **3**, 1901–1902.

Ristaldi, M. S., Pirastu, M., Murru, S., Casula, L., Loudianos, G., Cao, A., Sciarratta, G. V., Agosti, S., Parodi, M. I. & Leone, D. (1990). A spontaneous mutation produced a novel elongated beta-globin chain structural variant (Hb Agnana) with a thalassemia-like phenotype. *Blood* **75**, 1378–1379.

Rose, O. & Falush, D. (1998). A threshold size for microsatellite expansion. *Molecular Biology and Evolution* **15**, 613–615.

Roux, A. F., Morle, F., Guetarni, D., Colonna, P., Sahr, K., Forget, B. G., Delaunay, J. & Godet, J. (1989).

Molecular basis of Sp alpha I/65 hereditary elliptocytosis in North Africa: insertion of a TTG triplet between codons 147 and 149 in the alpha-spectrin gene from five unrelated families. *Blood* **73**, 2196–2201.

Ruttledge, M. H., Andermann, A. A., Phelan, C. M., Claudio, J. O., Han, F., Chretien, N., Rangaratnam, S., MacCollin, M., Short, O., Parry, D., Michels, V., Riccardi, V. M., Weksberg, R., Kitamura, K., Bradburn, J. M., Hall, B. D., Propping, P. & Rouleau, G. A. (1996). Type of mutation in the neurofibromatosis type 2 gene (NF2) frequently determines severity of disease. *American Journal of Human Genetics* **59**, 331–342.

Schlötterer, C. & Tautz, D. (1992). Slippage synthesis of simple sequence DNA. *Nucleic Acids Research*. **20**, 211–215.

Sege-Peterson, K., Chambers, J., Page, T., Jones, O. W. & Nyhan, W. L. (1992). Characterization of mutations in phenotypic variants of hypoxanthine phosphoribosyltransferase deficiency. *Human Molecular Genetics* **1**, 427–432.

Siddique, Z., McPhaden, A. R. & Whaley, K. (1993). C1-inhibitor gene nucleotide insertion cause type II hereditary angio-oedema. *Human Genetics* **92**, 189–190.

Stephan, W. & Cho, S. (1994). Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. *Genetics* **136**, 333–341.

Strassmann, J. E., Barefield, K., Solís, C. R., Hughes, C. R. & Queller, D. C. (1997). Trinucleotide microsatellite loci for a social wasp, *Polistes*. *Molecular Ecology* **6**, 97-100.

Sutherland, G. R. & Richards, R. I. (1995). Simple tandem DNA repeats and human genetic disease. *Proceedings of the National Academy of Sciences USA* **92**, 3636–3641.

Tse, W. T., Gallagher, P. G., Pothier, B., Costa, F. F., Scarpa, A., Delaunay, J. & Forget, B. G. (1991). An insertional frameshift mutation of the beta-spectrin gene associated with elliptocytosis in spectrin nice (beta 220/216). *Blood* **78**, 517–523.

Valdes, A. M., Slatkin, M. & Freimer, N. B. (1993). Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* **133**, 737–749.

Weber, J. L. (1990). Informativeness of human (dC-dA)$_n$·(dG-dT)$_n$ polymorphisms. *Genomics* **7**, 524–530.

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

White, M. B., Amos, J., Hsu, J. M., Gerrard, B., Finn, P. & Dean, M. (1990). A frameshift mutation in the cystic fibrosis gene. *Nature* **344**, 665–667.

Wierdl, M., Dominska, M. & Petes, T. D. (1997). Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* **146**, 769–779.

Wijnen, J., Khan, P. M., Vasen, H., Menko, F., van der Klift, H., van den Broek, M., van Leeuwen-Cornelisse, I., Nagengast, F., Meijers-Heijboer, E. J., Lindhout, D., Griffioen, G., Cats, A., Kleibeuker, J., Varesco, L., Bertario, L., Bisgaard, M. L., Mohr, J., Kolodner, R. & Fodde, R. (1996). Majority of hMLH1 mutations responsible for hereditary nonpolyposis colorectal cancer cluster at the exonic region 15–16. *American Journal of Human Genetics* **58**, 300–307.

Yates, J. R., van Bakel, I., Sepp, T., Payne, S. J., Webb, D. W., Nevin, N. C. & Green, A. J. (1997). Female germline mosaicism in tuberous sclerosis confirmed by molecular genetic analysis. *Human Molecular Genetics* **6**, 2265–2269.

Zhu, Y., Queller, D. C. & Strassmann, J. E. (2000). A phylogenetic perspective on sequence evolution in microsatellite loci. *Journal of Molecular Evolution* **50**, 324–338.

https://doi.org/10.1017/S001667230000478X