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, 1987a; 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, 1987b), 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
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; 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 protomicrosatellite. 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.

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

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).
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 \((7/19) \times 35 = 12.89\) new 2-repetition 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 \((59/1000) \times 9070 = 535.13\) (59 substitutions of that type, 1000 of 9070 substitutions assessed). The overall fraction of 2-repeat trinucleotides from insertions was therefore \(12.89/(12.89 + 535.13) = 0.0035\). 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.2a). 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.

---

**Table 1. Dinucleotide insertions and their surrounding sequences**

<table>
<thead>
<tr>
<th>Gene Symbol/name</th>
<th>Nucleotides affected</th>
<th>Duplication of adjacent dinucleotide? (number duplicated)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>p67-phox</td>
<td>CCTcCTTGG</td>
<td>Yes (1)</td>
<td>Nuno et al. (1995)</td>
</tr>
<tr>
<td>beta-spectrin</td>
<td>CGAaAGAGGTG</td>
<td>Yes (2)</td>
<td>Tse et al. (1991)</td>
</tr>
<tr>
<td>Hb Agmana</td>
<td>AACGtGTTGCA</td>
<td>Yes (2)</td>
<td>Ristaldi et al. (1990)</td>
</tr>
<tr>
<td>Na-Cl cotransporter</td>
<td>TtgCTCTG</td>
<td>No</td>
<td>Mastroianni et al. (1996)</td>
</tr>
<tr>
<td>TSC2</td>
<td>GCCGAGGAG</td>
<td>No</td>
<td>Au et al. (1998)</td>
</tr>
<tr>
<td>NF1</td>
<td>TATaucGTTTCG</td>
<td>No</td>
<td>Colman et al. (1997)</td>
</tr>
<tr>
<td>DHA-PA</td>
<td>CAttGTATAT</td>
<td>No</td>
<td>Ofman et al. (1998)</td>
</tr>
<tr>
<td>PPO</td>
<td>GGAGagCCCTA</td>
<td>Yes (1)</td>
<td>Lam et al. (1997)</td>
</tr>
<tr>
<td>CF</td>
<td>CATCtcTCATTCC</td>
<td>Yes (2)</td>
<td>Iannuzzi et al. (1991)</td>
</tr>
<tr>
<td>hMSH2</td>
<td>GAcaTTTAC</td>
<td>No</td>
<td>Maliaka et al. (1996)</td>
</tr>
<tr>
<td>NF1</td>
<td>AGTttACTG</td>
<td>No</td>
<td>Ainsworth et al. (1993)</td>
</tr>
<tr>
<td>SOD-1</td>
<td>TGAAtAGAA</td>
<td>No</td>
<td>Orrell et al. (1997)</td>
</tr>
<tr>
<td>DSS</td>
<td>CATacGCT</td>
<td>Yes (1)</td>
<td>Rautenstrauss et al. (1994)</td>
</tr>
<tr>
<td>APC</td>
<td>CATATA</td>
<td>Yes (1)</td>
<td>Paffenholtz et al. (1994)</td>
</tr>
<tr>
<td>C4A</td>
<td>GGCCTcAGTC</td>
<td>Yes (1)</td>
<td>Barba et al. (1993)</td>
</tr>
<tr>
<td>APC</td>
<td>ATTTuA</td>
<td>No</td>
<td>Mandl et al. (1994)</td>
</tr>
<tr>
<td>COL3A1</td>
<td>ATTGTGTC</td>
<td>No</td>
<td>Richards et al. (1994)</td>
</tr>
<tr>
<td>IDUA</td>
<td>TCCaCTTC</td>
<td>No</td>
<td>Bunge et al. (1994)</td>
</tr>
<tr>
<td>NF2</td>
<td>AGGAGaGTCTT</td>
<td>Yes (2)</td>
<td>Mautner et al. (1996)</td>
</tr>
<tr>
<td>PAX6</td>
<td>GCcGcGTGC</td>
<td>No</td>
<td>Davis &amp; Cowell (1993)</td>
</tr>
<tr>
<td>MS2H</td>
<td>ATAtgTGATACGA</td>
<td>Yes (1)</td>
<td>Nystrom-Lahti et al. (1996)</td>
</tr>
<tr>
<td>MATA1</td>
<td>GACTgCTAA</td>
<td>No</td>
<td>Chamberlin et al. (1996)</td>
</tr>
<tr>
<td>hMLH1</td>
<td>GTGcGcACC</td>
<td>Yes (1)</td>
<td>Wijnen et al. (1996)</td>
</tr>
<tr>
<td>CFTR</td>
<td>GGATATATatatATTC</td>
<td>Yes (4)</td>
<td>White et al. (1990)</td>
</tr>
<tr>
<td>PAX6</td>
<td>TACTggATCCA</td>
<td>Yes (1)</td>
<td>Jordan et al. (1992)</td>
</tr>
<tr>
<td>ZIC3</td>
<td>GGGCttGAGA</td>
<td>No</td>
<td>Geppia et al. (1997)</td>
</tr>
<tr>
<td>G6Pase</td>
<td>CATCatATAGT</td>
<td>Yes (2)</td>
<td>Lei et al. (1993)</td>
</tr>
<tr>
<td>CD40</td>
<td>CGTCtccCGAC</td>
<td>Yes (2)</td>
<td>Macchi et al. (1995)</td>
</tr>
<tr>
<td>PDH</td>
<td>AGttTTTTCC</td>
<td>No</td>
<td>Chun et al. (1993)</td>
</tr>
<tr>
<td>RBP</td>
<td>CAGAgtaTG</td>
<td>Yes (1)</td>
<td>Kohmann et al. (1994)</td>
</tr>
<tr>
<td>HBB</td>
<td>ACTGttTGACA</td>
<td>Yes (2)</td>
<td>Ristaldi et al. (1990)</td>
</tr>
</tbody>
</table>
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 repeats.
The origin of microsatellites

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. 2b). 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. 2b).

We could determine the relative numbers of short repeat loci generated by substitutions versus insertions 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, 1987b), 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
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)$_{10}$; 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
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


Foster, J. W., Dominguez-Steglich, M. A., Guioli, S., Kowk, https://doi.org/10.1017/S001667230000478X Published online by Cambridge University Press


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.


