BIOINFORMATICS DISCOVERY NOTE Vol. 27 no. 7 2011, pages 895–898
doi:10.1093/bioinformatics/btr067

Genome analysis Advance Access publication February 11, 2011

Bases adjacent to mononucleotide repeats show an increased single nucleotide polymorphism frequency in the human genome

K. J. Siddle*,†, J. A. Goodship, B. Keavney and M. F. Santibanez-Koref
Institute of Human Genetics, Newcastle University, Newcastle upon Tyne, NE1 3BZ, UK
Associate Editor: John Quackenbush

ABSTRACT

Mononucleotide repeats (MNRs) are abundant in eukaryotic genomes and exhibit a high degree of length variability due to insertion and deletion events. However, the relationship between these repeats and mutation rates in surrounding sequences has not been systematically investigated. We have analyzed the frequency of single nucleotide polymorphisms (SNPs) at positions close to and within MNRs in the human genome. Overall, we find a 2- to 4-fold increase in the SNP frequency at positions immediately adjacent to the boundaries of MNRs, relative to that at more distant bases. This relationship exhibits a strong asymmetry between 3′ and 5′ ends of repeat tracts and is dependent upon the repeat motif, length and orientation of surrounding repeats. Our analysis suggests that the incorporation or exclusion of bases adjacent to the boundary of the repeat through substitutions, in which these nucleotides mutate towards or away from the base present within the repeat, respectively, may be another mechanism by which MNRs expand and contract in the human genome.

Contact: kjsiddle@pasteur.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 24, 2010; revised on January 11, 2011; accepted on February 1, 2011

1 INTRODUCTION

Tandem repeats of short sequence motifs are common in the genomes of eukaryotic species (Ellegren, 2004). Lander *et al*. (2001) estimated that microsatellites with repeat motifs between 1 and 6 bp account for at least 3% of the human genome. Mononucleotide repeats (MNRs), repeats of a single base, are the most abundant class of microsatellites in the human genome (Toth *et al*., 2000). However, this class of repeats has been frequently overlooked in studies of microsatellite evolution and variation.

One of the most salient characteristics of short tandem repeats is the high variability in length exhibited by these sequences within populations due to the insertion or deletion of repeat units, most likely reflecting difficulties in replication and repair (Levinson and Gutman, 1987; Wierdl *et al*., 1997). This variability has led to the widespread use of these sequences as polymorphic markers (Kong *et al*., 2002; Kopelman *et al*., 2009). For MNRs, this type of instability is an important characteristic of some tumors, in particular, mismatch repair deficient cancers of the gut and ovary (Buhard *et al*., 2006; Kawaguchi *et al*., 2009).

Alterations in the rate and spectrum of point mutations have also been reported in sequences surrounding microsatellites. Several studies considering small numbers of di- and trinucleotide repeats have found evidence of an increase in the mutation rate near the repeat (Brohede and Ellegren, 1999; Djian *et al*., 1996; Vowles and Amos, 2004). However, there is a lack of consensus regarding this trend with other publications proposing that the frequency of mutations decreases toward microsatellites (Varela and Amos, 2010), or remains stable (Karhu *et al*., 2000).

In this article, we have investigated the relationship between MNRs and mutation rates in the human genome. Using data from the dbSNP database, we took the number of single nucleotide polymorphisms (SNPs) in the region surrounding MNRs as a measure of the propensity of a particular site to be affected by mutations. We hypothesized that any variation in the frequency of SNPs is likely to be strongest at positions closest to the repeat, as homopolymeric regions are more likely to be subject to errors during replication which could result in the misincorporation of a base at adjacent positions (Brohede and Ellegren, 1999). The boundaries of MNRs can also be more easily and clearly defined than those of repeats with longer motifs since sources of ambiguity such as adjacent truncated motifs do not occur. Therefore, by using MNRs we were able to focus in particular on positions immediately adjacent to the repeat in order to discern, with greater accuracy, the extent to which any change in mutation rate is specific to the bases at the boundary between repetitive and unique sequence.

2 METHODS

2.1 Datasets

All MNRs of ≥3 bp were extracted from the hg18 assembly of the human reference genome (http://genome.ucsc.edu) using a perl script. Only repeats in autosomes were included in the present analysis. Mutations occurring close to MNR tracts were identified using the dbSNP130 database (http://genome.ucsc.edu). All SNPs for which the method used to validate the variant was known (8.5 million) were included in the analysis.

2.2 Analysis

To focus specifically on local differences in SNP frequency around MNRs, we examined 10 bp either side of each repeat tract as well as the first and last three bases of the repeat itself. The distribution of SNPs was then analyzed according to repeat type, repeat length, base change and orientation.

3 RESULTS AND DISCUSSION

A total of 176.9 million autosomal MNRs ≥3 bp in length were identified. As T and G repeats are the same as A and C repeats,

*To whom correspondence should be addressed.
†Present Address: Institut Pasteur, Human Evolutionary Genetics, Department of Genomes and Genetics, F-75015 Paris, France.

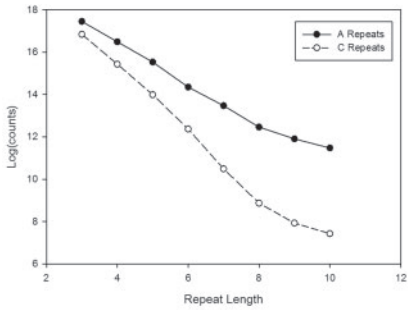

Fig. 1. MNR counts according to length and repeat base in the human genome. Counts are given on a natural log scale for repeats of between 3 and 10 bp.

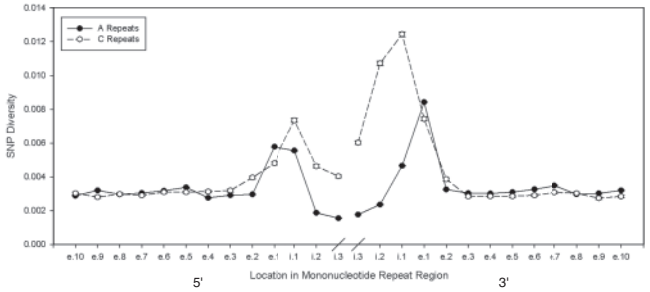

Fig. 2. SNP diversity at positions surrounding polyA and polyC repeats of length >5 bp. Positions are given from the base 10 bp from the 5′ end of the repeat to the base 10 bp from the 3′ end of the repeat. Error bars show the standard errors of the estimates; however, due to the large sample size, these intervals are very tight. SNP diversity surrounding A, T, C and G repeats prior to standardization are given in Supplementary Figure 1.

respectively, but in the opposite orientation, it is common practice to standardize MNRs into polyA (including A and T repeats) and polyC (including C and G repeats) (Katti *et al*., 2001; Kofler *et al*., 2007). All MNRs analyzed were standardized according to this convention.

Over 122 million polyA and 54 million polyC repeats were extracted from the human genome (Supplementary Table 1). The frequency of repeats decreased with increasing length with polyC repeats showing a faster rate of decline, and therefore shorter mean and maximum lengths, than polyA repeat tracts (Fig. 1).

3.1 Relationship between diversity and position

We initially assessed the SNP diversity at positions surrounding MNRs of length >5 bp, as this is slightly below the proposed threshold for slippage mutations in MNRs (Lai and Sun, 2003; Rose and Falush, 1998). We designated positions from 1 to 10 bp outside the MNR as e.1–e.10 and the first and last three bases inside the repeat as i.1–i.3, where the number reflects the distance from the MNR boundary. These designations can occur either 5′ or 3′ of the MNR, where the strand used is the one on which the MNR is encoded as a polyA or a polyC tract.

We observed an increased diversity at positions immediately adjacent to the MNR boundary compared with more distant positions (Fig. 2). However, differences in the relationship between MNRs and SNP frequency between polyA and polyC repeats suggests that this relationship is further influenced by the repeat motif. In polyA repeats, this increase was largest at the first positions outside the microsatellite, while polyC tracts showed the greatest excess of mutations at the first and final bases of the repeat itself. Looking beyond the repeat boundaries, the diversity at bases between 3 and 10 bp from the repeat remained virtually constant and was similar for both classes of repeat ($\sim 3 \times 10^{-3}$). Within the repeat, focusing only on the first and last three bases, diversity for polyA repeats fell below that of external, non-boundary positions, consistent with previous observations from CA microsatellites (Brohede and Ellegren, 1999). In contrast, in polyC repeats, diversity remained elevated relative to that of more distant positions, though it did fall at positions further inside the repeat.

We also observed a marked 5′ to 3′ asymmetry at the boundaries of repeat tracts (Fig. 2). This indicated that the 3′ end of a repeat—on the strand where the MNR is encoded as an A or C tract—has a higher frequency of mutations relative to the 5′ end (polyA 5′ = 2×, 3′ = 2.8×; polyC 5′ = 2.4×, 3′ = 4.1×). There was further a larger effect at the boundaries of polyC repeats compared with polyA repeats based on a χ^2 test ($P < 1 \times 10^{-16}$). This asymmetry is reminiscent of observations made in dinucleotide repeats and certain minisatellites (Varela *et al*., 2008; Vowles and Amos, 2004).

Repeat tracts can be difficult to sequence (Lunter *et al*., 2008; Shinde *et al*., 2003). It is, therefore, possible that the excess of SNPs observed in bases adjacent to these repeats was an artifact. To assess whether such errors were likely to influence our results, we selected well-characterized SNPs, for which allele frequencies have been determined, and repeated the analysis using SNPs with a minor allele frequency (MAF) $\geqslant$0.05 (3 million). The distribution of SNPs relative to MNRs was found to be very similar to that observed using the full database (Supplementary Fig. 2). The strong similarity between these analyses indicates that conclusions drawn from the full dataset are unlikely to be the result of artifacts. We also aligned 667 submitted sequences (www.ncbi.nlm.nih.gov) corresponding to 100 SNPs, adjacent to MNRs, to the hg18 assembly of the human reference sequence. Of these, only two sequences showed ambiguities in the alignment around the location of the SNP, which may indicate the presence of an insertion or deletion event. Moreover, that we consistently observed in our results such a clear excess of SNPs only in the bases at the repeat boundary, and that polyA and polyC repeats were found to have different relationships with the mutation rate at adjacent bases suggests that sequencing or alignment issues are not likely to be influencing the results.

To show that our findings are likely to reflect mutational processes affecting all MNRs and are not influenced by constraints due to their location, for example in coding or transcribed regions, we reanalyzed our data according to the genomic regions in which these repeats are found. Repeats were annotated according to whether they; overlapped with coding regions, were within the mature transcripts of protein-coding genes but outside the coding regions, were located in introns or were found in intergenic regions. We observed that the relative differences in the mutation rate at the boundary positions of MNRs are similar irrespective of the location of the repeat (Supplementary Table 2).

3.2 Relationship between diversity and repeat length

To establish whether the relationship between MNRs and mutation rate is dependent on repeat length, SNP diversity at the boundaries of

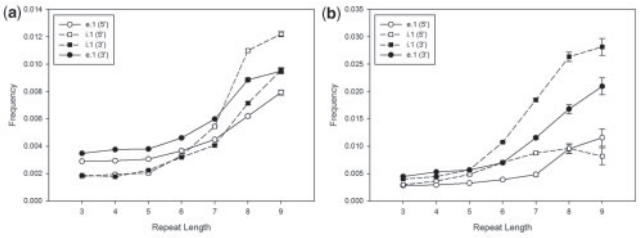

Fig. 3. Frequency of point mutations in bases at the boundary of polyA (**a**) and polyC (**b**) repeats of increasing length. Error bars show the standard errors of the estimates.

repeat tracts was analyzed according to repeat size. We included only repeats of ≤9 bp as beyond this the frequency of repeats became too low to draw reliable conclusions for the less-abundant polyC repeats.

Diversity increased with repeat length in both polyA and polyC repeats up to 9 bp at positions adjacent to the repeat boundary (Fig. 3). This trend, however, was substantially more marked in repeats of ≥5 bp suggesting that there may be a threshold below which little effect is observed. The greatest increase observed for polyA repeats was at the 5′ internal boundary (5-fold difference between repeats of 3 and 9 bp). In polyC repeats, a slower increase was observed in 5′ positions, while the internal 3′ boundary position showed the largest and most rapid increase with a 7-fold difference between the longest and shortest repeats. At the 5′ internal boundary, a slight decrease of 9 bp was observed in polyC repeats. As the error bars for this estimate overlap with those for repeats of 8 bp, this may be an artifact of the small sample size. However, it may reflect a real change in the relationship between MNRs and mutation rate at this length, although caution should still be exercised when extrapolating this to repeats >9 bp.

This is reminiscent of the relationship observed between repeat length and repeat instability. The rate of slippage during replication, one of the main causes of length variability, increases with the length of homopolymeric tracts (Sia *et al*., 1997; Wierdl *et al*., 1997), leading to an increase in repair in these sequences. It may be hypothesized, that similar processes to those influencing MNR length variation also affect the mutation rate at the boundaries of these repeats. This observation is independent of the decrease in mutation rate in regions surrounding long repeats anticipated by Kruglyak *et al.* (1998) and noted in previous studies (Santibanez-Koref *et al*., 2001) since it pertains to the frequency of SNPs in the broader sequence context of the repeat, while our observations concern positions immediately adjacent to the repeat boundary.

To investigate the relationship between length polymorphisms and the frequency of point mutations at the MNR boundary, we repeated our analysis using only those repeats showing length variability according to the dbSNP130 database. We find evidence that repeats showing length variation are enriched for SNPs at the boundary positions: repeats containing indels represent 0.005% of repeats but 0.04–0.1% of those with substitutions in boundary positions (Supplementary Table 3). However, these repeats showed the same trends in relation to the frequency of mutations in the boundary positions for A and C repeats as monomorphic repeats. Furthermore, as these repeats represent such a small fraction of the total dataset they were not excluded from the analysis.

Table 1. Counts of base changes at 5′ and 3′ external boundary positions in polyA and polyC repeats

	PolyA repeats		PolyC repeats	
Type of mutation	Adjacent 5′	Adjacent 3′	Adjacent 5′	Adjacent 3′
Change to repeat base	30 267	43 155	1692	2079
	(71.5%)	(77.9%)	(50.6%)	(54.2%)
Change to other base	12 039	12 274	1650	1755
	(28.5%)	(22.1%)	(49.4%)	(45.7%)
Total	42 306	55 429	3342	3834

3.3 Relationship between repeat motif and base change

To investigate whether variants at external positions immediately adjacent to a repeat showed a tendency to change to the base of the nearest mononucleotide tract, we selected SNPs for which ancestral allele information was known (8.4 million). We included only tracts of length >5 bp. Compared with the proportion expected if there was no influence of the MNR (1 in 3), we observed a significant excess of changes to the base of the neighboring repeat (Table 1). While this excess was observed in both polyA and polyC repeats, it was significantly larger in the former ($P < 1 \times 10^{-16}$, χ^2 test). Furthermore, there was a 3′–5′ asymmetry in the extent of this effect with the 3′ end of the repeat showing a significantly higher proportion of SNPs changing to the repeat base ($P < 1 \times 10^{-16}$).

This is consistent with the hypothesis that slippage increases the probability of misincorporation of a base at positions adjacent to MNRs (Brohede and Ellegren, 1999). Furthermore, it may contribute to the more rapid decline in repeat frequency with increasing repeat length and lower average repeat length observed for polyC repeats compared to polyA repeats, since this would result in a tendency for polyC repeats to decrease in size, through mutations away from the repeat motif in the last and first repeat base, and to be interrupted by internal mutations. Conversely, polyA repeats exhibited a greater internal stability and would be more likely to lengthen through mutations in the bases adjacent to the MNR.

3.4 Relationship between diversity and repeat orientation

Finally, we assessed the influence of the sequence context on diversity. Findings for trinucleotide repeats have indicated that stability depends on the orientation of the repeat with respect to the direction of progress of the replication fork (Freudenreich *et al*., 1997). For regions of the genome that are preferentially replicated in one orientation, this should result in repeats in one orientation being more stable than those in the opposite orientation, leading to the clustering of repeats in the same orientation. Within such clusters, repeats in the less common orientation should be less stable. We therefore investigated whether the frequency of point mutations at positions adjacent to the MNR boundary was dependent on the orientation of the nearest repeat with the same motif. In this analysis, we included only MNRs >5 bp. We found that the orientation of the nearest repeat and diversity are not independent (Table 2). Compared with the numbers expected assuming independence, we observed for both polyA and polyC repeats an excess of mutations in the first external base when the nearest neighbor had the same orientation. Conversely, there was an excess of mutations in the last base of

Table 2. Relationship between the orientation of the nearest MNR and the number of SNPs

	PolyA repeats		PolyC repeats	
Position	Nearest tract orientation		Nearest tract orientation	
	Same	Opposite	Same	Opposite
External	432 639 (59%)	29 5948 (41%)	197 938 (55%)	160 129 (45%)
Internal	261 564 (52%)	240 190 (48%)	215 548 (53%)	189 852 (47%)
P-value*	$< 2.2 \times 10^{-16}$		$< 2.2 \times 10^{-16}$	

*Chi-squared test.

a repeat when the nearest neighbor was in the opposite orientation. Interestingly, in the absence of other factors, such a bias would result in a tendency of MNRs in a same orientation context to enlarge, while those that are in regions where MNRs show predominantly the opposite orientation would tend to contract. This indicates that our earlier observations regarding the relationship between MNRs and the frequency of SNPs is correlated with the orientation of neighboring MNRs, suggesting that, in a certain sequence context, MNRs in one orientation will be more stable than those in the opposite orientation.

4 CONCLUSIONS

In this study, we have used information from published sources and extracted from the human reference sequence to show that there is a significant increase in the number of SNPs immediately adjacent to the boundaries of MNR tracts. SNPs located at these repeat boundaries account for 2% of all validated SNPs in the human genome. We have found no evidence to suggest that these observations are due to sequencing or alignment errors and are, therefore, confident that they represent a genuine relationship, contributing to genetic variability at repeat boundaries.

The inclusion of all MNRs in the human genome means that this study is the most comprehensive investigation to date of the distribution of variation surrounding MNRs. Our findings expand upon previous discussion of the relationship between microsatellites and the mutation rate. We have shown that there is a substantial increase in mutation rate that specifically affects the bases at the boundaries of MNRs and that this increase is modulated by repeat motif, length and orientation with respect to neighboring repeats. We have further shown a tendency for point mutations at the external boundary to mutate toward the repeat motif, suggesting that the incorporation and exclusion of adjacent bases may be a mechanism by which these repeats expand and contract in the human genome.

Funding: This work was supported by grants from the British Heart Foundation (BHF) and the National Institute for Health Research Biomedical Research Centre in Ageing and Age-related Disease awarded to the Newcastle Hospital National Health Service Foundation Trust. B.K. holds a BHF Personal Chair.

Conflict of Interest: none declared.

REFERENCES

Brohede,J. and Ellegren,H. (1999) Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. *Proc. Biol. Sci.*, **266**, 825–833.

Buhard,O. *et al*. (2006) Multipopulation analysis of polymorphisms in five mononucleotide repeats used to determine the microsatellite instability status of human tumors. *J. Clin. Oncol.*, **24**, 241–251.

Djian,P. *et al*. (1996) Codon repeats in genes associated with human diseases: fewer repeats in the genes of nonhuman primates and nucleotide substitutions concentrated at the sites of reiteration. *Proc. Natl Acad. Sci. USA*, **93**, 417–421.

Ellegren,H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435–445.

Freudenreich,C.H. *et al*. (1997) Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. *Mol. Cell. Biol.*, **17**, 2090–2098.

Karhu,A. *et al*. (2000) Rapid expansion of microsatellite sequences in pines. *Mol. Biol. Evol.*, **17**, 259–265.

Katti,M.V. *et al*. (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.*, **18**, 1161–1167.

Kawaguchi,M. *et al*. (2009) Analysis of candidate target genes for mononucleotide repeat mutation in microsatellite instability-high (MSI-H) endometrial cancer. *Int. J. Oncol.*, **35**, 977–982.

Kofler,R. *et al*. (2007) SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics*, **23**, 1683–1685.

Kong,A. *et al*. (2002) A high-resolution recombination map of the human genome. *Nat. Genet.*, **31**, 241–247.

Kopelman,N.M. *et al*. (2009) Genomic microsatellites identify shared Jewish ancestry intermediate between Middle Eastern and European populations. *BMC Genet.*, **10**, 80.

Kruglyak,S. *et al*. (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.

Lai,Y. and Sun,F. (2003) The relationship between microsatellite slippage mutation rate and the number of repeat units. *Mol. Biol. Evol.*, **20**, 2123–2131.

Lander,E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Levinson,G. and Gutman,G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.*, **4**, 203–221.

Lunter,G. *et al*. (2008) Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.*, **18**, 298–309.

Rose,O. and Falush,D. (1998) A threshold size for microsatellite expansion. *Mol. Biol. Evol.*, **15**, 613–615.

Santibanez-Koref,M.F. *et al*. (2001) A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes. *Mol. Biol. Evol.*, **18**, 2119–2123.

Shinde,D. *et al*. (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Res.*, **31**, 974–980.

Sia,E.A. *et al*. (1997) Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. *Mol. Cell. Biol.*, **17**, 2851–2858.

Toth,G. *et al*. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, **10**, 967–981.

Varela,M.A. *et al*. (2008) Heterogeneous nature and distribution of interruptions in dinucleotides may indicate the existence of biased substitutions underlying microsatellite evolution. *J. Mol. Evol.*, **66**, 575–580.

Varela,M.A. and Amos,W. (2010) Heterogeneous distribution of SNPs in the human genome: microsatellites as predictors of nucleotide diversity and divergence *Genomics*, **95**, 151–159.

Vowles,E.J. and Amos,W. (2004) Evidence for widespread convergent evolution around human microsatellites. *PLoS Biol.*, **2**, E199.

Wierdl,M. *et al*. (1997) Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics*, **146**, 769–779.