Downloaded from http://rsbl.royalsocietypublishing.org/ on April 20, 2018

biology letters
Molecular evolution

Biol. Lett. (2008) **4**, 399–401
doi:10.1098/rsbl.2008.0209
Published online 3 June 2008

Body temperature predicts maximum microsatellite length in mammals

William Amos[1,*] and Andrew Clarke[2]

[1]*Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK*
[2]*British Antarctic Survey, NERC, High Cross, Madingley Road, Cambridge CB3 0ET, UK*
**Author for correspondence* (*w.amos@zoo.cam.ac.uk*).

A long-standing mystery in genome evolution is why short tandem repeats vary so much in length and frequency. Here, we test the hypothesis that body temperature acts to influence the rate and nature of slippage-based mutations. Using the data from both 28 species where genome sequencing is advanced and 76 species from which marker loci have been published, we show that in mammals, maximum repeat number is inversely correlated with body temperature, with warmer-blooded species having shorter 'long' microsatellites. Our results support a model of microsatellite evolution in which maximum length is limited by a temperature-dependent stability threshold.

Keywords: microsatellite; mutation; genome

1. INTRODUCTION

Microsatellites are abundant components of higher eukaryotic genomes and used widely as genetic markers. Most mutations occur by slippage (Schlötterer & Tautz 1992), resulting in the gain or loss of single repeat units (Ellegren 2000), though with occasional larger 'jump' mutations (Di Rienzo *et al.* 1994) and point mutations that may disrupt the repeat array (Jin *et al.* 1996). This apparently simple mode of evolution has prompted intensive study, yet key aspects remain unresolved. One of these relates to the upper length boundary: very long microsatellites are rare or absent, implying a mechanism capable of constraining repeat number (Garza *et al.* 1995).

Two models have emerged as alternative mechanisms capable of explaining the upper length boundary. Most simply, length may not be constrained *per se*, but the time taken to become long could allow point mutations to disrupt the repeat array, thereby reducing the maximum number of pure repeats (Kruglyak *et al.* 1998). The alternative model invokes some kind of stability threshold. Microsatellites often exhibit directional biases such that expansion mutations outnumber contractions or vice versa. Early studies reported mainly an expansion bias (Amos *et al.* 1996; Primmer *et al.* 1996), but it now appears that the direction of bias switches from expansion to contraction when an allele becomes 'long' (Xu *et al.* 2000). Such a pattern can be viewed as a stability threshold capable of restricting maximum repeat number by bias reversal.

One intriguing aspect of microsatellite length is that, for any given repeat motif, the mean and maximum repeat number varies greatly between taxa. Looking for general patterns, it has been noted that microsatellite length increases from birds through mammals to fishes and reptiles, suggesting a testable hypothesis that maximum length is influenced by a body temperature-dependent stability threshold (Amos 1999). However, these vertebrate groups have contrasting physiologies that may impact on microsatellite length for reasons other than body temperature, making the underlying causative factor unclear. For a more rigorous test of this hypothesis, it would be preferable to ask whether the same pattern exists within a single taxonomic group. Consequently, we decided to exploit the publication of large tracts of genomic sequences, including several complete or advanced genome assemblies, in order to ask whether a correlation exists between maximum repeat number and body temperature in mammals.

2. MATERIAL AND METHODS

The maximum repeat number was estimated in two ways. First, large tracts of sequences were downloaded for 28 diverse mammals in which genome sequencing is advanced or complete. For each of these, we analysed a target minimum of approximately 250 Mb of sequence, either as the largest available complete chromosome or as a large block of concatenated contigs (for details, see electronic supplementary material). From these, we extracted a maximum of 20 000 microsatellites containing five or more pure repeats, repeating the process in turn for four different motifs: $(AT)_n$; $(AC)_n$; $(AAT)_n$; and $(AAC)_n$. The maximum repeat number was estimated using extrapolation in log–log plots of repeat number against frequency (figure 1), performed 'blind' and by eye due to frequent nonlinearities that preclude linear algebraic extrapolation. The species list and data sources are summarized in table 1 of electronic supplementary material.

Given the limited number of species for which sufficiently large quantities of bulk sequence are available, we also analysed microsatellites developed as genetic markers. As a rule, during marker development, people tend to select from among the clones they generate to maximize both repeat number and repeat purity, two traits that correlate positively with variability. Consequently, marker microsatellites are crudely representative of the longest repeat tracts carried by a species. We therefore used BLAST to search GenBank for $(AC)_{7+}$ and then selected up to a maximum of eight loci at random in which the maximum number of pure repeats was determined. Given the stochasticity associated with choosing one single locus to represent a species, we used the mean repeat number across all markers examined as our measure of maximum length. By so doing, we obtain an estimate of the length of the longest markers, rather than of the maximum possible length. Searches were conducted for 596 species for which good body temperature data are available in the literature (Clarke & Rothery 2008), yielding usable maximum length estimates for a total of 78 species. These data are summarized in table 2 of electronic supplementary material.

Phylogenetic non-independence is not expected to be a significant problem because most species are evolutionarily distant from all others. However, to control for this possibility, we repeated all regressions using a comparative analysis by independent contrasts (CAICs; Purvis & Rambaut 1995). Mammalian phylogenies were constructed by reference to many sources but particularly to two web-based resources, the Tree of Life (http://tolweb.org/tree/) and Wikipedia, and phylogenies of the rodents (Huchon *et al.* 2002), marsupials (Cardillo *et al.* 2004) and carnivores (Bininda-Edwards *et al.* 1999). Since a good distance matrix is not available for all species, we assumed stepwise evolution. The application of CAIC is conservative because body temperature evolves unpredictably and is often shared owing not to recent shared ancestry but to shared ecology, such as body size or flight, or shared physiology, such as the low body temperature characteristic of groups such as marsupials and edentates.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsbl.2008.0209 or via http://journals.royalsociety.org.

Received 14 April 2008
Accepted 16 May 2008

This journal is © 2008 The Royal Society

Downloaded from http://rsbl.royalsocietypublishing.org/ on April 20, 2018

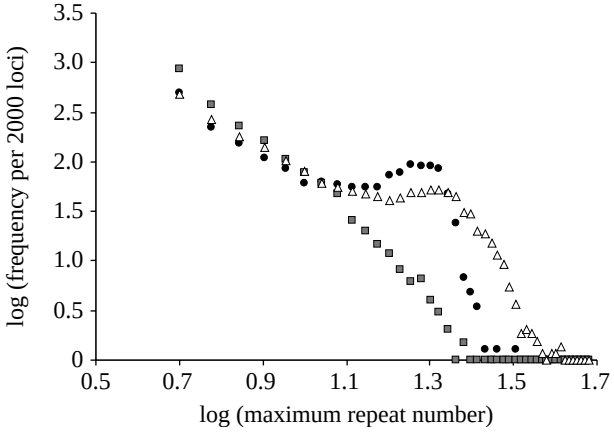

Figure 1. Extrapolation of maximum microsatellite repeat number. Examples of the log–log extrapolation of maximum repeat number are given for the bushbaby (grey squares, extrapolated maximum = 1.41), European hedgehog (open triangles, extrapolated maximum = 1.6) and guinea-pig (filled circles, extrapolated maximum = 1.46). These illustrate unhumped, medium humped and strongly humped distributions, respectively.

3. RESULTS

Both the bulk sequence data and cloned microsatellites reveal negative correlations between body temperature and microsatellite length. Bulk genome sequences are represented by $(AT)_n$ (figure 2*a*), by some way the strongest trend ($R^2 = 0.44$, $n = 28$, $p = 0.00011$). Both $(AAT)_n$ ($R^2 = 0.19$, $n = 25$, $p = 0.027$) and $(AC)_n$ ($R^2 = 0.156$, $n = 28$, $p = 0.037$) reveal marginally significant, negative trends, while $(AAC)_n$ is negative but non-significant ($R^2 = 0.057$, $n = 25$, $p = 0.25$). For cloned microsatellites, there is understandably more scatter but the overall trend is significant ($R^2 = 0.09$, $n = 78$, $p = 0.007$; figure 2*b*). These relationships appear due to body temperature rather than metabolic rate because in a multiple regression with both body temperature and body size (a strong correlate of the metabolic rate), only temperature is significant. Repeating all regressions using the method of independent contrasts to control for shared ancestry, we obtain similar though somewhat less significant results (genomic AT, $R^2 = 0.31$, $n = 28$, $p = 0.0019$; cloned AT, $R^2 = 0.1$, $n = 28$, $p = 0.0066$; all other comparisons not significant).

4. DISCUSSION

We conducted two independent tests of whether the length of a species' longest microsatellites correlates with body temperature in mammals. In both the cases, negative trends are revealed, consistent with the broader picture seen across the main vertebrate groups. Allowing for phylogenetic non-independence reduces but does not eliminate the trends.

The proportion of variation explained by our regressions could be seen as modest. However, congeneric species differ by up to 4 K in body temperature, half the total range of approximately 8 K we examine, suggesting that body temperature can evolve very rapidly. When body temperature increases, a thermal stability threshold might be able to track the change quite closely by causing the rapid deletion of

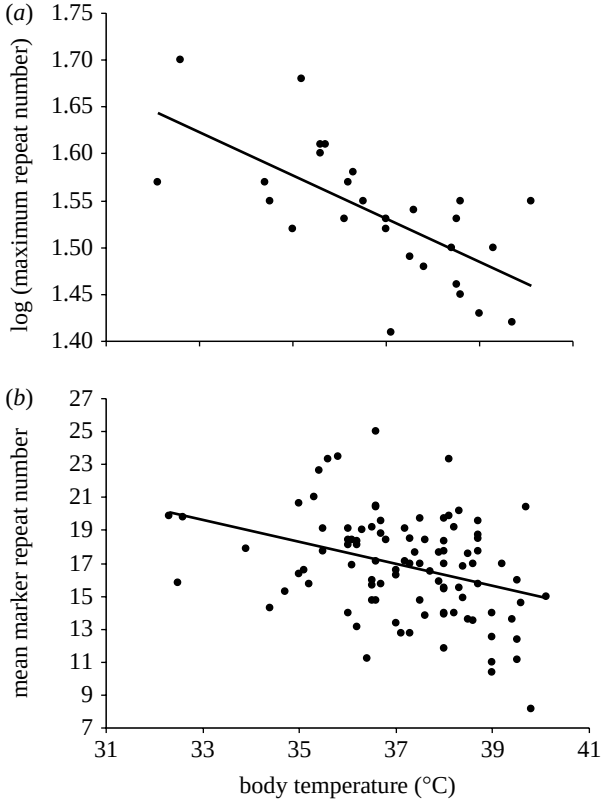

Figure 2. Relationship between mammalian body temperature and maximum microsatellite repeat number. (*a*) Correlation between maximum repeat number of $(AT)_n$ microsatellites identified in bulk genomic DNA, determined as in figure 1, and body temperature in diverse mammals ($R^2 = 0.44$, $n = 28$, $p = 0.00011$). (*b*) Correlation between mean repeat number of up to eight $(AC)_n$ microsatellite markers deposited in GenBank and body temperature for 78 mammals ($R^2 = 0.09$, $n = 78$, $p = 0.007$). Non-marker sequences such as those identified in non-coding DNA sequences were ignored, since these have not been selected for greater length.

the longest microsatellites. By contrast, if body temperature decreases, tracking is likely to be less efficient because it requires time for the longest microsatellites to expand further up to a higher boundary. In this light, it is interesting to note that the strongest relationship is seen for the shortest, least stable motif, AT, and the weakest for the longest, most stable motif, AAC. Such a pattern is consistent with the higher rates of slippage seen in shorter AT-rich motifs (Tuntiwechapikul & Satazar 2002; Kelkar *et al.* 2008), allowing more precise tracking of changes in body temperature.

Stronger regressions are also not expected because body temperature is often difficult to define. Mammals exhibit a circadian variability in body temperature, the amplitude of which is inversely proportional to body size (Aschoff 1982). Some taxa (e.g. tenrecs) have notoriously 'loose' thermoregulation, while mole rats have abandoned endothermy altogether and several mammalian lineages, such as bats, squirrels and rodents, include members who hibernate, dropping their body temperatures dramatically during cold-season torpor. Here, it is unclear whether any thermal stability threshold would be determined more by the mean or by the maximum temperature.

Downloaded from http://rsbl.royalsocietypublishing.org/ on April 20, 2018

However, if body temperature is important, it should tend to impact similarly on all motifs, predicting that the maximum repeat number will correlate more strongly between motif types than with our spot estimates of body temperature. This is what we find: all pairwise correlations are significant (weakest = AT versus AAC, $r=0.48$, $n=25$, $p=0.014$; strongest = AT versus AAT, $r=0.70$, $n=25$, $p=0.000009$). Thus, despite the sparse triplet repeat data, the maximum repeat number of all motifs correlates strongly with the maximum AT repeat number, which in turn correlates with body temperature.

Our analysis also provides evidence suggestive of transitions between different equilibrium states. At equilibrium, one might expect to find a rather monotonic decline in the log microsatellite frequency with an increasing repeat number, as we observe for, for example, the bushbaby (figure 1). However, more than half of all profiles exhibit some form of 'hump', in which microsatellite frequency either remains flat or even increases with increasing repeat number (e.g. the hedgehog; figure 1). In extreme cases, the profile can appear either truncated, with frequency declining from many copies to zero within a few repeat units, or to have a tail in which the transition from rare to absent occurs over a wide range of repeat numbers. Such patterns are evocative of transitional states that might arise variously when the genome-wide mutation rate changes or when the upper length boundary evolves. Discovering the exact meaning of these diverse profiles will require further modelling.

In summary, we have shown that body temperature and maximum microsatellite repeat number are correlated, providing a direct tie between physiology and genome evolution. Although establishing a causal link will require further work, our findings extend the idea that higher body temperature increases the rate of point mutations (Gillooly *et al*. 2005) and suggest that temperature can impact on molecular evolution qualitatively as well as quantitatively. Moreover, given that slippage is known to be temperature sensitive (Tuntiwechapikul & Satazar 2002; Kelkar *et al*. 2008), a correlation between body temperature and maximum length would seem to support the thermal stability threshold over mutational degradation as the key factor preventing extreme expansion. The range of shapes seen among the frequency–repeat number plots, from linear declines to various humped distributions (figure 1), is suggestive of transitional states, perhaps caused during adaptation to new thermal niches, but understanding their significance will require further work.

Amos, W. 1999 A comparative approach to the study of microsatellite evolution. In *Microsatellites: evolution and applications* (eds D. B. Goldstein & C. Schlötterer), pp. 66–79. Oxford, UK: Oxford University Press.

Amos, W., Sawcer, S. J., Feakes, R. & Rubinsztein, D. C. 1996 Microsatellites show mutational bias and heterozygote instability. *Nat. Genet.* **13**, 390–391. (doi:10.1038/ng0896-390)

Aschoff, J. 1982 The cricadian rhythm of body temperature as a function of body size. In *A companion to animal physiology* (eds C. R. Taylor, K. Johansen & L. Bolis), pp. 173–188. Cambridge, UK: Cambridge University Press.

Bininda-Edwards, O. R. P., Gittleman, J. L. & Purvis, A. 1999 Building large trees by combining phylogenetic information: a complete tree of the extant Carnivora (Mammalia). *Biol. Rev.* **74**, 143–175. (doi:10.1017/S0006323199005307)

Cardillo, M., Bininda-Edwards, O. R. P., Boakes, E. & Purvis, A. 2004 A species-level phylogenetic super-tree of marsupials. *J. Zool. Lond.* **264**, 11–31. (doi:10.1017/S0952836904005539)

Clarke, A. & Rothery, P. 2008 Scaling of body temperature in mammals and birds. *Funct. Ecol.* **22**, 58–67. (doi:10.1111/j.1365-2435.2007.01341.x)

Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M. & Slatkin, M. 1994 Mutational processes of simple sequence repeat loci in human populations. *Proc. Natl Acad. Sci. USA* **91**, 3166–3170. (doi:10.1073/pnas.91.8.3166)

Ellegren, H. 2000 Heterogeneous mutation processes in human microsatellites. *Nat. Genet.* **24**, 400–402. (doi:10.1038/74249)

Garza, J. C., Slatkin, M. & Freimer, N. B. 1995 Microsatellite allele frequencies in humans and chimpanzees with implications for constraints on allele size. *Mol. Biol. Evol.* **12**, 594–604.

Gillooly, J. F., Allen, A. P., West, G. B. & Brown, J. H. 2005 The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proc. Natl Acad. Sci. USA* **102**, 140–145. (doi:10.1073/pnas.0407735101)

Huchon, D., Madsen, O., Sibbald, M. J. J. B., Ament, K., Stanhope, M. J., Catzeflis, F., de Jong, W. W. & Douzery, E. J. P. 2002 Rodent phylogeny and a timescale for the evolution of Glires: evidence from an extensive taxon sampling using three nuclear genes. *Mol. Biol. Evol.* **19**, 1053–1065.

Jin, L., Macaubas, C., Hallmayer, J., Kimura, A. & Mignot, E. 1996 Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. *Proc. Natl Acad. Sci. USA* **93**, 15 285–15 288. (doi:10.1073/pnas.93.26.15285)

Kelkar, Y. D., Tyekucheva, S., Chlaromonte, F. & Makova, K. 2008 The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res.* **18**, 30–38. (doi:10.1101/gr.7113408)

Kruglyak, S., Durrett, R. T., Schug, M. D. & Aquadro, C. F. 1998 Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl Acad. Sci. USA* **95**, 10 774–10 778. (doi:10.1073/pnas.95.18.10774)

Primmer, C., Ellegren, H., Saino, N. & Møller, A. P. 1996 Directional evolution in germline microsatellite mutations. *Nat. Genet.* **13**, 391–393. (doi:10.1038/ng0896-391)

Purvis, A. & Rambaut, A. 1995 Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *CABIOS* **11**, 247–251.

Schlötterer, C. & Tautz, D. 1992 Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* **20**, 211–215. (doi:10.1093/nar/20.2.211)

Tuntiwechapikul, W. & Satazar, M. 2002 Mechanism of *in vitro* expansion of long DNA repeats: effect of temperature, repeat length, repeat sequence and DNA polymerases. *Biochemistry* **41**, 854–860. (doi:10.1021/bi0110950)

Xu, X., Peng, M., Fang, Z. & Xu, X. 2000 The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* **24**, 396–399. (doi:10.1038/74238)