Genomics 95 (2010) 151–159

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Heterogeneous distribution of SNPs in the human genome: Microsatellites as predictors of nucleotide diversity and divergence

Miguel A. Varela [a,b,*], William Amos [a]

[a] Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK
[b] Fundación Pública Galega de Medicina Xenómica (Grupo de Medicina Xenómica), CIBERER, Hospital Clínico Universitario, Santiago de Compostela, A Coruña, Spain

ARTICLE INFO

Article history:
Received 2 October 2009
Accepted 11 December 2009
Available online 21 December 2009

Keywords:
Divergence
Diversity
Microsatellites
SNPs
Substitutions

ABSTRACT

Understanding the forces that govern the distribution of single nucleotide polymorphisms is vital for many of their applications. Here we conducted a systematic search to quantify how both SNP density and human–chimpanzee divergence vary around different repetitive sequences. We uncovered a highly complicated picture in which these quantities often differ significantly from the genome-wide average in regions extending more than 20 kb, the direction of the deviation varying with repeat number and motif. AT microsatellites in particular are potent predictors of SNP density, long $(AT)_n$ repeat tracts tending to be found in regions of significantly reduced SNP density and low GC content. Although the causal relationships remain difficult to determine, our results indicate a strong relationship between microsatellites and the DNA that flanks them. Our results help to explain the mixed picture that emerges from other studies and have important implications for the way in which genetic diversity is distributed in our genomes.

Introduction

Single nucleotide polymorphisms (SNPs) are the most widespread form of sequence variation in the genome, representing about 90% of human DNA polymorphism [1]. In recent years, SNPs have replaced microsatellites as the markers of choice for most large-scale studies of model organisms and particularly for humans [2,3]. Applications include gene mapping, inference of patterns of natural selection and the elucidation of population histories. Literally millions have already been identified in the human genome as part of the International HapMap Project [4].

For many applications, an understanding of the forces that govern the distribution of SNPs is either desirable or even vital. Despite this onus, many aspects of SNP evolution remain poorly understood. In particular, SNPs tend to occur non-randomly [5], variously creating regions that can be viewed as high density clusters [5–7] or low density 'bare patches'. Such patterns are suggestive of the action of natural selection, with balancing selection promoting clusters [8] and purifying selective sweeps acting to denude a region and increase linkage disequilibrium [9,10]. However, clusters may also arise through the presence of mutation hotspots where the local mutation rate is strongly elevated [6,11], raising the possibility that low density regions could also reflect regions where the mutation rate is depressed.

Several factors have already been identified as being associated with SNP clusters. The clearest predictor of SNP density appears to be recombination rate [12–16]. Wherever the recombination rate is unusually high, so also tends to be the density of SNPs. As yet it is unclear whether high recombination rates increase the local mutation rate or *vice versa*. It is even possible that both features correlate with some third, as yet unknown factor. SNP clusters may also arise through ascertainment biases, including the development of high density maps in and around genes of medical interest. One further factor that may be linked is the occurrence of microsatellites, whose presence also correlates with recombination rates [17–19].

An association between the distribution of SNPs and the presence of microsatellites seems entirely plausible and is supported by several lines of evidence. First, microsatellites exhibit high levels of length polymorphism such that heterozygous individuals can be viewed as carrying microdeletions, potentially enhancing the local mutation rate [20]. Second, stresses associated with the unusual base-stacking of purine–pyrimidine repeats or other structural properties [21–23] could also be responsible for the mutational biases observed in regions flanking microsatellites [24–26]. Third, even within microsatellites, mutations appear to occur non-randomly, favouring the 3′ end and showing a decreased transition/transversion ratio relative to non repetitive sequences [27]. Thus, although the exact mechanisms remain largely obscure, it seems clear that microsatellites influence both the nature and, in all probability, the rate of mutations occurring in their vicinity.

To study the possible impact of microsatellites on the local distribution of SNPs we used data from the HapMap Project Database

* Corresponding author. Fax: +44 0 1223 336677.
E-mail address: mav33@cam.ac.uk (M.A. Varela).

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.12.003

[4] to reconstruct patterns of average SNP density around microsatellites of different lengths and motifs and compared these diversity patterns with the divergence of flanking sequences of orthologous human and chimpanzee microsatellites.

Results

The numbers and lengths of all microsatellites found are summarised in Table 1. We began by exploring how the density of SNPs varies around $(AC)_n$, $(AG)_n$ and $(AT)_n$ microsatellites at two resolutions, 1 kb (Fig. 1) and 10 kb either side (Fig. 2). Fig. 1 reveals two main trends. First, SNP density varies significantly around the three motifs, as indicated by the way the average density at one location often lies outside the 95% confidence interval at another. Overall, the consensus pattern appears to be one in which SNP density decreases towards the microsatellite, revealed as a central trough of SNP density. However, the 3D smoothing tends to mask an apparent peak in SNP density at very short lengths where the microsatellite is only two repeats long. This pattern is seen most convincingly for $(AT)_2$ and is, if anything, a dip for $(AG)_2$. The second noticeable trend is for the relative frequencies of SNPs to vary with the length of the microsatellite. Thus, all three motifs occur in low SNP density regions relative to the rest of the genome when short (two repeats), generally lie in higher than average regions at five repeats and then tend to have dropped back down again at the largest repeat number (20 repeats), particularly close to the microsatellite.

Moving to a broader scale, using 1 kb windows totalling 10 kb either side of the microsatellite (Fig. 2), somewhat related patterns are seen. The trough that is apparent for all three motifs at fine resolution is now only apparent for $(AT)_n$ and even then only among the longer repeat classes, though here it does seem very pronounced. Rather as expected, with only two repeats all three motifs generally lie in regions of average SNP density. Despite this, all three motifs lie in regions of above-average SNP density when they carry five repeats. Interestingly, at this length the average SNP density varies significantly between motifs, being lowest for AT (~1.5), intermediate for AC (~1.55) and highest for AG (>1.55). At the largest repeat number examined, 20 repeats, the motifs differ both in mean SNP density and in the shape of the graph: AC is relatively flat with suggestions of a peak near the microsatellite, AG exhibits some level of asymmetry, with SNP density being greater 5′ while AT reveals a dip that lies predominantly below the genome average density. To determine the maximum extent of these patterns we conducted one further, low resolution study based on 50 kb each side of $(AC)_{20}$, $(AG)_{20}$ and $(AT)_{20}$ (Fig. 3). Both the $(AC)_{20}$ and $(AG)_{20}$ plots show little variation, all lying at or near the genome-wide mean. In contrast, the $(AT)_{20}$ plot exhibits a profound dip, extending approximately 10 kb either side of the microsatellite.

SNP density provides an indication of current patterns of mutation rate variation around microsatellites. For a longer term view, we compared orthologous sequences in humans and chimpanzees, and calculated nucleotide divergence, using the two different bin sizes used above (Figs. 4 and 5). At the finer scale, totalling 1 kb either side (Fig. 4), all three motifs exhibit a similar 3D pattern that appears actually to be the exact converse of the pattern seen for SNP density. Thus, while SNP density tends to exhibit a low-point near to the microsatellite, human–chimpanzee divergence tends to exhibit a peak. This contradiction is also seen in the 2D splines, particularly for the two shortest length classes where all three motifs tend show a peak in divergence when SNP density shows a trough and *vice versa*. Moving to the broader scale (Fig. 5) tends to reduce the peaks and troughs while at the same time emphasising tendencies to lie in regions with either above or below average human–chimpanzee divergence. The most striking features are the strong decline in divergence between the regions that contain $(AC)_5$ (~12.5) and $(AC)_{20}$ (~11.9), and for $(AT)_n$ to show a well-supported hump at 20 repeats, despite a dip at five repeats (Fig. 5). In fact, in the region adjacent to $(AT)_5$ there is already a small hump in divergence that can only be seen at the finer scale (Fig. 4, $(AT)_5$). Here the peak in divergence seems to expand along with the microsatellite inside a greater region of low divergence.

Table 1
Abundance and length of all microsatellites found in the entire human genome.

Repeats	$(AC)_n$	$(AT)_n$	$(AG)_n$
2	256,985	256,788	259,612
3	160,525	168,735	175,188
4	37,841	47,842	44,614
5	10,413	8492	8211
6	4481	4173	3019
7	2426	2410	1574
8	1490	1505	905
9	1021	1003	577
10	931	758	366
11	846	577	258
12	877	464	166
13	903	404	154
14	910	310	142
15	1000	286	108
16	1037	223	95
17	1072	232	91
18	1027	177	72
19	945	180	80
20	957	195	49

Although Figs. 1, 2, 4 and 5 provide a good summary for how SNP density and divergence vary with distance from a microsatellite carrying either 2, 5 or 20 repeats, they provide substantially less information about how these traits vary with microsatellite length in general. Consequently, we constructed plots for how SNP density and human–chimpanzee divergence vary with repeat number over three different regions: 100 bp either side of the microsatellite, 1 kb either side but excluding the nearest 100 bp and 10 kb either side excluding the nearest 1 kb (Fig. 6). AC and AG exhibit broadly similar patterns, in the sense that, within any given panel, SNP density appears not to vary significantly with repeat number, yet across all panels the fluctuations are probably consistent enough to suggest that SNP density does rise and fall. The same is partly true for divergence, though here the AC panels show a striking peak of divergence at 10 repeats closest to the microsatellite and a tendency for low divergence further out for the largest repeat numbers. AT reveals much stronger patterns, particularly for SNP density, which declines almost monotonically as repeat number increases at all three resolutions. This trend appears somewhat contradicted by the divergence panels, which show an increase near to the microsatellite and little variation at lower resolution.

One aspect we have so far ignored is broader sequence context, specifically local GC content. To learn whether there is a general relationship between microsatellites and the sequences in which they occur, we calculated the average GC content of sequences flanking each of the three motifs at lengths 2, 5 and 20 repeats, based on the 100 bases either side of each microsatellite (Table 2). We find that most motifs occur in sequences with an average GC content of around 42%. The exceptions are $(AC)_5$ and, particularly, longer AT microsatellites.

Discussion

In this study we explore the relationship between microsatellites and the nucleotide diversity and divergence of the sequences in which they occur. We find a complicated picture in which both the density of SNPs and human–chimpanzee divergence vary with motif type, distance from the microsatellite and the length of the microsatellite itself. AT microsatellites in particular are potent predictors of SNP density, loci with higher repeat numbers tending to lie in regions of significantly reduced SNP density.

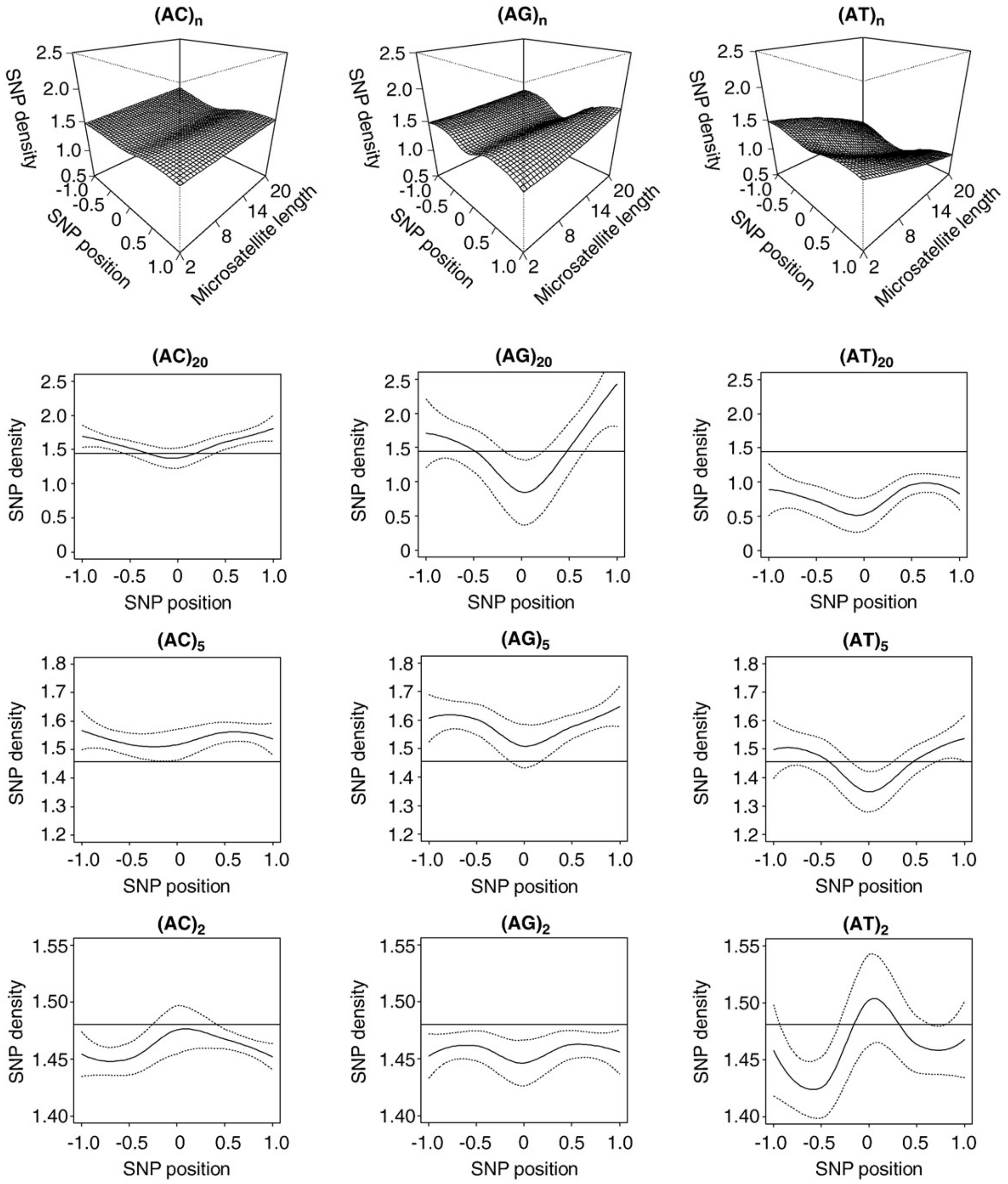

Fig. 1. SNP density in 1 kb sequences flanking $(AC)_n$, $(AG)_n$ and $(AT)_n$ microsatellites measured as the number of SNPs per kb. 3D plots show how SNP density varies in sequences flanking microsatellites of different sizes. Since 3D spline-fitting can over-smooth fine-scale patterns and make it difficult to display confidence intervals we also include 2D slices through the 3D graph, taken at 2, 5 and 20 repeats. Dotted lines represent 95% confidence intervals of the best fit local regression of all points. Horizontal lines represent the autosomal average SNP density.

Correlational studies of microsatellite and SNP characteristics are hampered by two important confounding factors. First there is the question of observation bias. Put generally, microsatellites selected for study are inevitably a subset of all microsatellites, and the selection criteria run the risk of influencing what is found [28,29]. In our study we chose microsatellites without neighbouring runs of the same motif. In doing so, we were forced to assume that these regions are representative of those in which all other equivalent microsatellites are located, but this assumption may not hold. For example, if microsatellites form naturally in regions with a high mutation rate, it might be the case that an 'average' microsatellite usually has a neighbour with the same motif and that the selection of isolated microsatellites biases the data in favour of low mutation rate regions. These and related biases are difficult to control and arguably under-acknowledged, but they must be born in mind when interpreting results in studies of this kind.

A second important problem relates to causality. For the most part we and others are able to demonstrate that two characteristics are correlated [30,31], but it is substantially more difficult to determine in which direction (if any) the causal link operates. Thus, if long microsatellites are found in regions with high SNP density this could result either if long microsatellites generate instability in their flanking sequences, or if high mutability regions allow enhanced rates of both point and slippage mutations. A further explanation would be if long microsatellites are short-lived, and are most often seen in regions of high mutability because there they are likely to

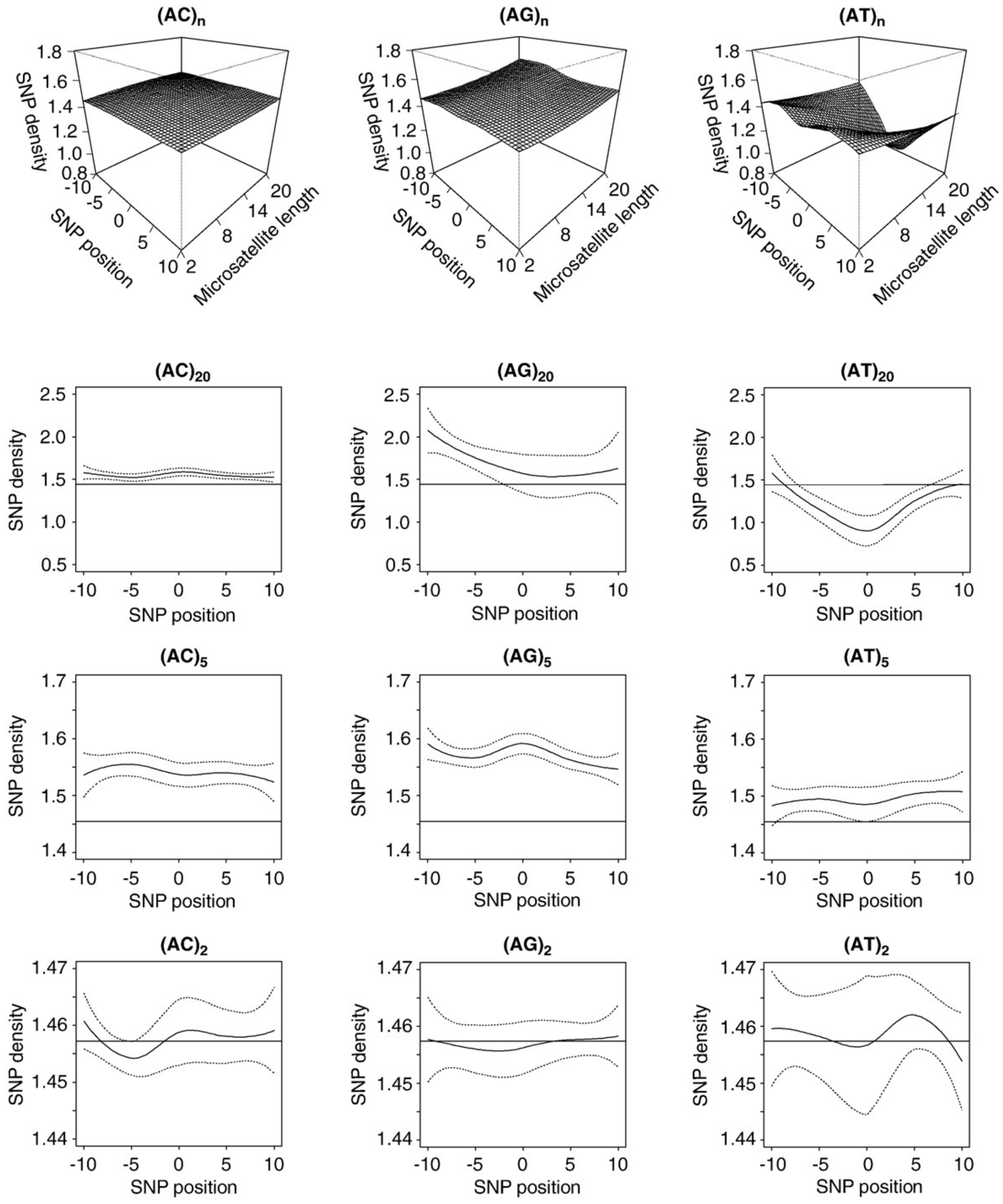

Fig. 2. SNP density in 10 kb sequences flanking $(AC)_n$, $(AG)_n$ and $(AT)_n$ microsatellites measured in number of SNPs per kb. For full plot details see legend to Fig. 1.

suffer internal point mutations which reduce slippage [32,33], prolong the time spent being long and make them more likely to be recorded, i.e. an observation bias. The general difficulty of inferring causal links must therefore be remembered when considering the patterns we present.

A recurring feature seen in many of the graphs is a clear perturbation around the point at which the microsatellite is located, seen most consistently and clearly at the 100 bp bin resolution, though also at wider scales for $(AT)_n$, seen either as a peak or a trough even within the same motif. Indeed, at the 100 bp bin resolution there are several instances of motifs with two and five repeats showing a peak at one length and a trough at the other (e.g. SNP density around AT and divergence around AC). Assuming that the populations of microsatellites studied for both repeat numbers are similar, these patterns suggest that as a microsatellite changes in length, the mutation processes around it also change. A pattern in which substitution rates near microsatellites change with repeat number could arise in any of several ways. First, there might be a causal link in the direction of microsatellites influencing their surroundings [25], with higher repeat numbers generating stronger conformational stresses that then impact on which and how many mutations occur. Under this model, we could speculate that SNP density and divergence would rise around a newly formed microsatellite to mitigate conformational stress, but might then fall as more and more sites carry bases that best relieve stress. Second, there could be an indirect link suggested by the work of Tian et al. [20]. Specifically, as a microsatellite gets longer, heterozygosity increases [31,34,35] and the locus will appear more often to carry a microdeletion, something that

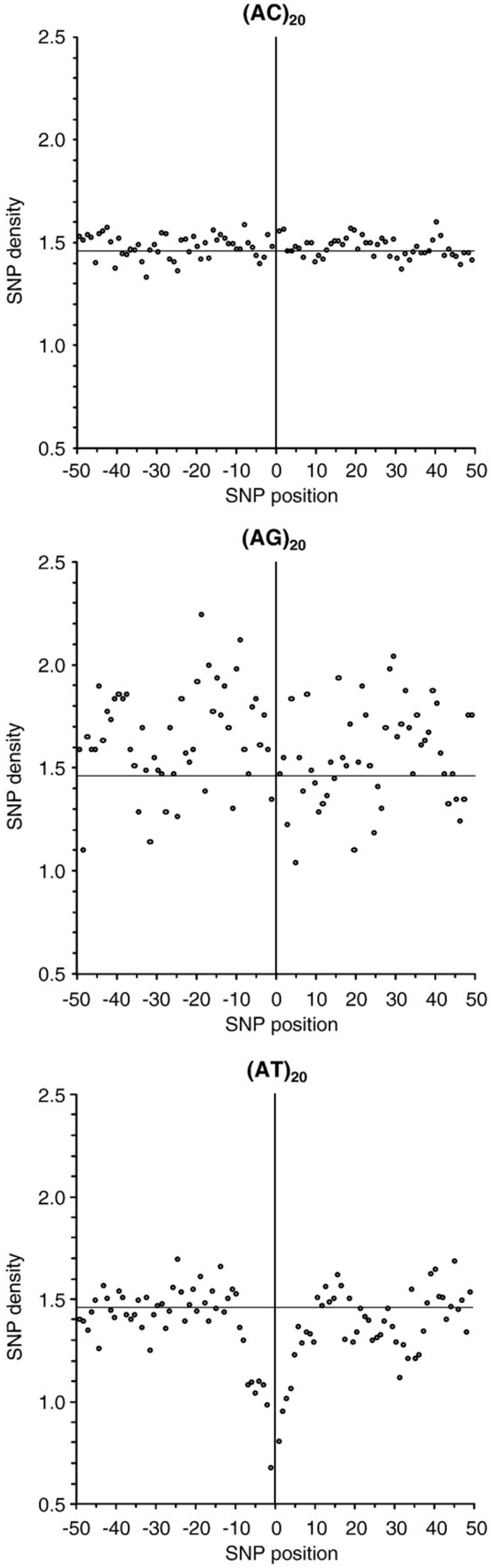

Fig. 3. SNP density in 50 kb sequences flanking $(AC)_n$, $(AG)_n$ and $(AT)_n$ microsatellites. SNPs were counted in a series of 50 equal-sized bins of 1 kb on either side of each microsatellite.

they show increases the local mutation rate. A third possibility is for a causal link in the direction of the sequence context influencing the probability of a microsatellite reaching any given length. For example, locally mutation rates might act both to create new short microsatellites and to slow the rate of slippage of longer microsatellites by causing interruption mutation with the repeat tract. As indicated above, the result could be an observation bias in which selection of microsatellites of a particular length co-selects for genomic regions with particular mutation properties.

Overall, the strongest patterns we find tend to occur closest to the microsatellite, in the smallest, 100 bp bin analyses. These reveal two unexpected, but potentially related features, namely a tendency for peaks to become troughs or vice versa as repeat number changes and, in most cases, an inverse relationship between the patterns seen for SNP density and divergence. In a simple model of evolution SNP density should be a strong predictor of human–chimpanzee divergence, almost the converse of what we observe. A solution to this conundrum may lie with the following speculative model. New microsatellites raise local mutation rates and induce mutation biases that together act to create a sequence context that minimises conformational stress. As this is achieved, the mutation rate falls both because the stresses are reduced and because many of the favoured changes have already happened. Sequence divergence accumulates over a much longer timescale and thus lags behind the SNP pattern such that maximum divergence is not achieved until SNP density has fallen, a pattern that persists as the microsatellite becomes longer. Whether or not this model is correct in detail, it emphasises the need at some level to invoke the occurrence of convergent or parallel mutations in order to explain the way that SNP density falls to below background levels and to reconcile the mismatch between SNP density and sequence divergence. This pattern could be related to the presence of cryptic periodicities in sequences flanking microsatellites [25,36] and is worthy of greater study.

Here it is worth considering the timescale over which changes might occur. Microsatellites may persist in the same location over many millions of years [37,38] and a very high proportion is conserved between humans and chimpanzees. In contrast, SNP density appears highly labile at fine scales, with a tendency towards clustering [7] but where clusters found in chimpanzees are often not found at the same site in humans [39,40]. Such transience might make it difficult for local SNP density to influence the behaviour of the much longer-lived microsatellites, and hence could be seen to favour models in which microsatellites influence the SNP density of their immediate surroundings [25,41] more than the other way round.

At a broader scale, another trend can be discerned for how SNP density varies around microsatellites. Thus, while microsatellites with two repeats tend to form or persist in regions with average SNP density, equivalent microsatellites with five repeats tend to lie in regions of at least 20 kb with significantly above average SNP density. By implication, either the presence of a short microsatellite increases the local mutation rate at this scale, or only a biased subset of two-repeat microsatellites expand up to five repeats, those lying in higher mutability regions. Such a broad sphere of influence seems to us far too large to reflect a direct consequence of the microsatellite. Instead, we suggest this pattern is more likely to reflect regional factors influencing microsatellite formation and expansion.

Despite their many similarities, over the entire analysis it is clear that the three motifs differ rather profoundly in their relationships with their flanking sequences. These differences are seen most obviously in Fig. 6, which summarises how SNP density and divergence at difference distances from a microsatellite vary with repeat number. Most strikingly, $(AT)_{20}$ repeats reveal a sphere of influence extending some 10 kb either side. Just why this motif exhibits such a broad pattern compared with $(AC)_{20}$ and $(AG)_{20}$ remains unclear and is worthy of further research. Interestingly,

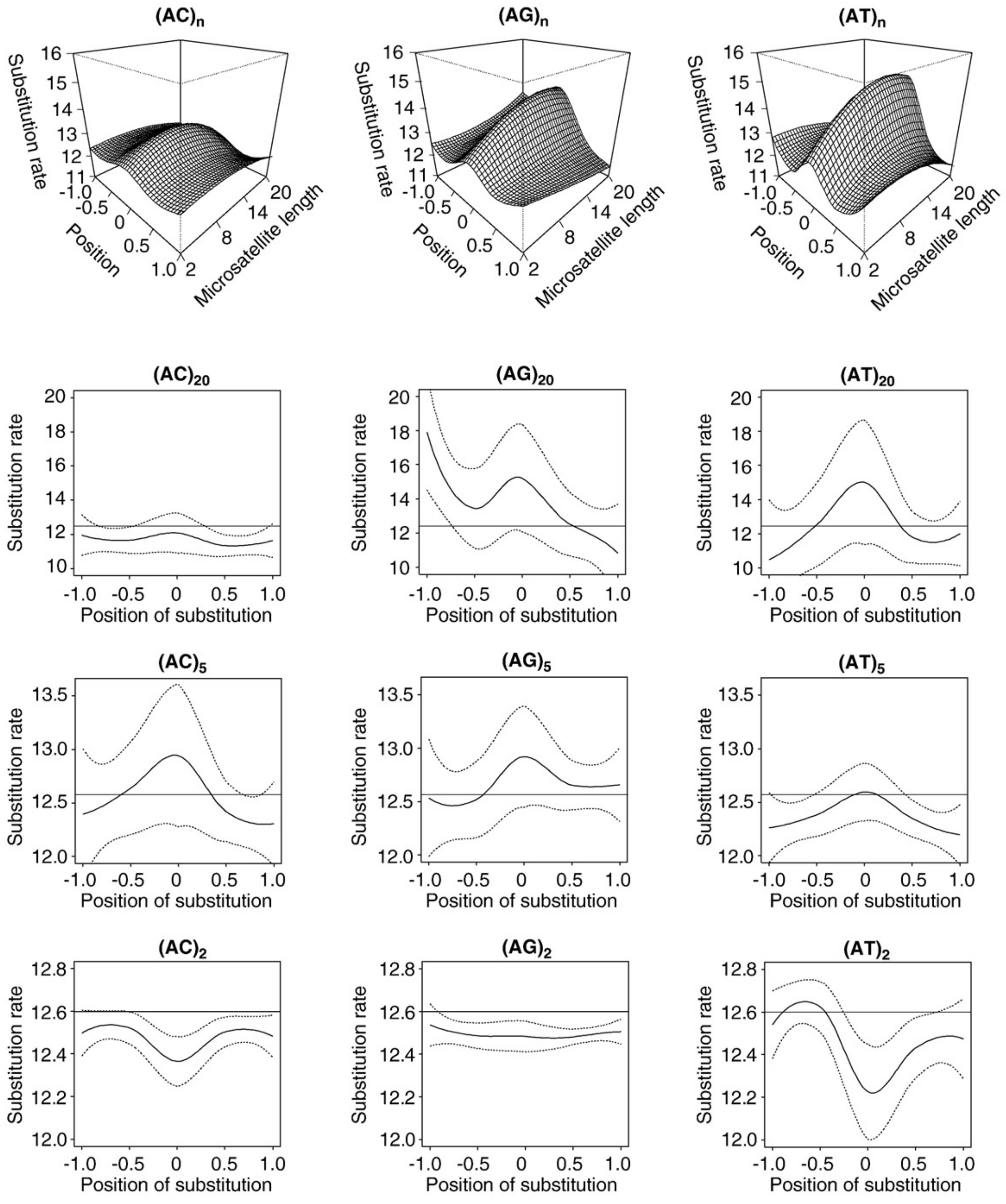

Fig. 4. Substitution rate in 1 kb sequences flanking $(AC)_n$, $(AG)_n$ and $(AT)_n$ microsatellites measured as the number of substitutions per kb. For further plot details see legend to Fig. 1. Horizontal lines represent the autosomal average divergence.

longer AT microsatellites tend to occur in regions with a lower GC content compared with the other microsatellites we studied, implying either that AT microsatellites are formed and expand more readily in AT-rich regions, or that AT microsatellites instil local mutation biases that favour mutations from G and C to A and T. Nonetheless, such inter-motif differences do tend to argue against a role for ascertainment bias in SNP discovery creating the patterns because the biases should be similar across different motifs. Instead, we feel these differences reflect genuine differences in the way the properties of each motif, for example the nature of the based-stacking and the propensity for slippage, together interact with the nature of the sequence in which the microsatellite is formed.

Stepping back, our analysis reveals a highly complicated picture that is difficult to reconcile with simple models of microsatellite and DNA sequence evolution, containing several features that require further study for proper elucidation. However, this very complexity helps to reconcile a number of earlier, apparently contradictory observations. Thus, SNP density in flanking sequences is reported to be positively correlated with microsatellite polymorphism [42], implying also a positive correlation between microsatellite length and flanking sequence divergence, yet the exact opposite appears true [43]. Similarly, microsatellites tend to occur preferentially in genomic regions with low SNP density [30,31], but this appears now to be an over-simplification given the variation we have observed among microsatellites of different lengths, coupled with the tendency for microsatellites to change in length over time.

In conclusion, it is normally assumed that microsatellites expand more where substitution rates are low and, therefore, with a tendency to show low diversity. Our study suggests that this is an oversimplification. We found that microsatellite length and motif are related to flanking sequence, SNP density and divergence in a highly complicated fashion. Regional effects could have an influence at least at some stages of microsatellite formation and expansion. Moreover, around microsatellites, SNP density and divergence fail to show the expected positive relationship. Although the direction of causality remains difficult to establish, we conclude that microsatellites and their flanking sequences are intimately associated to the extent that it becomes difficult to consider the evolution of one without also considering the evolution of the other. Moreover, whether microsatellites change the mutation patterns around them or local variation in mutation rate influences how microsatellites evolve (or both), the sheer number of microsatellite in higher organisms implies a substantial impact on our understanding of how genetic variability is generated.

Materials and methods

Data

We downloaded the complete genomic sequences of all human autosomes from the National Center for Biotechnology Information (NCBI) database (build 36 version 3), (Genbank accession numbers NC_000001–NC_000022) (www.ncbi.nlm.nih.gov). To represent the distribution of human SNPs we downloaded data for the CEU population (Utah residents with Northern and Western European ancestry from the CEPH collection) from the HapMap Project Database

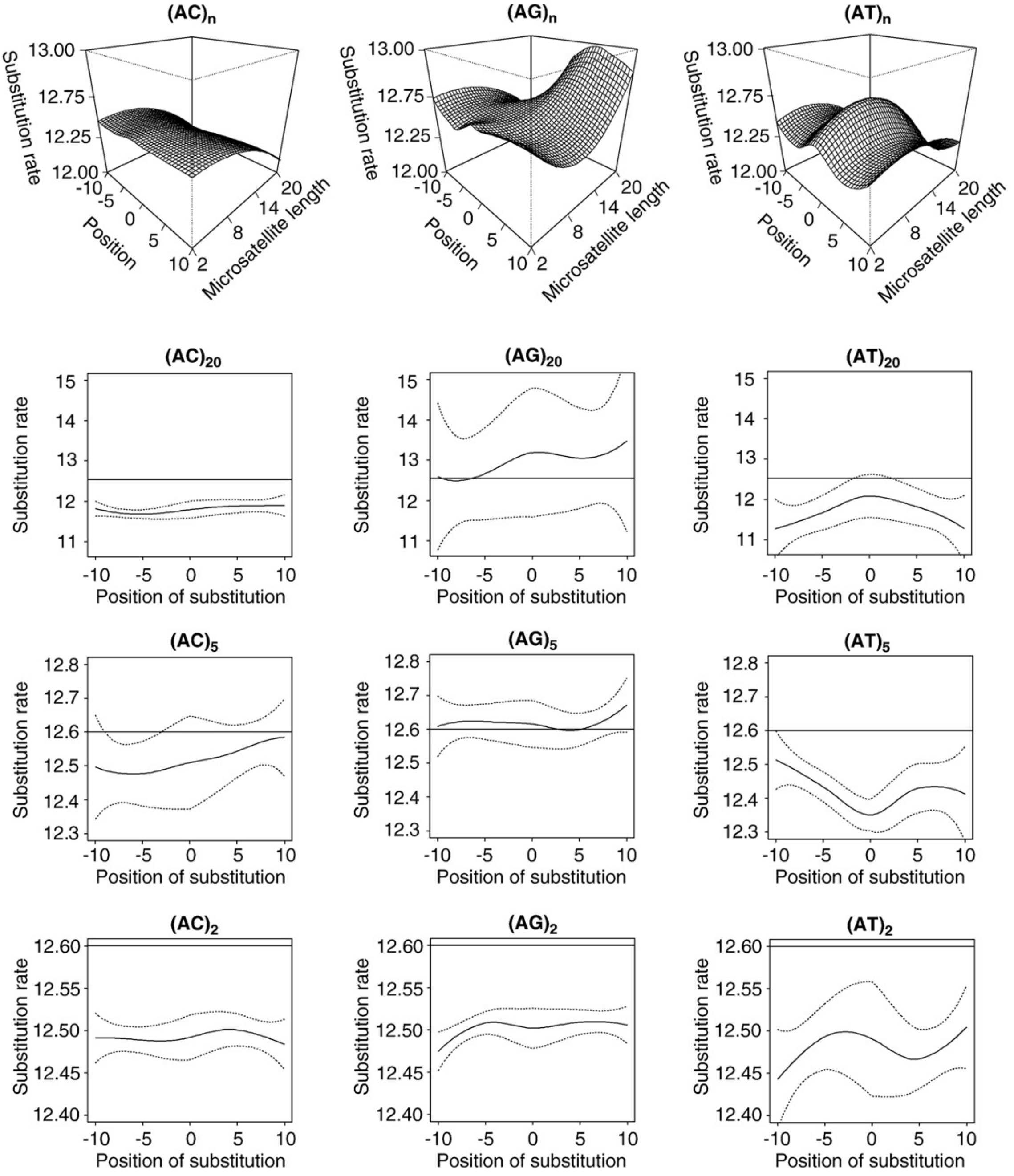

Fig. 5. Substitution rates in 10 kb sequences flanking $(AC)_n$, $(AG)_n$ and $(AT)_n$ microsatellites measured in number of substitutions per kb. 3D plots show how divergence varies in sequences flanking microsatellites of different sizes. For further plot details see legend to Fig. 1. Horizontal lines represent the autosomal average divergence.

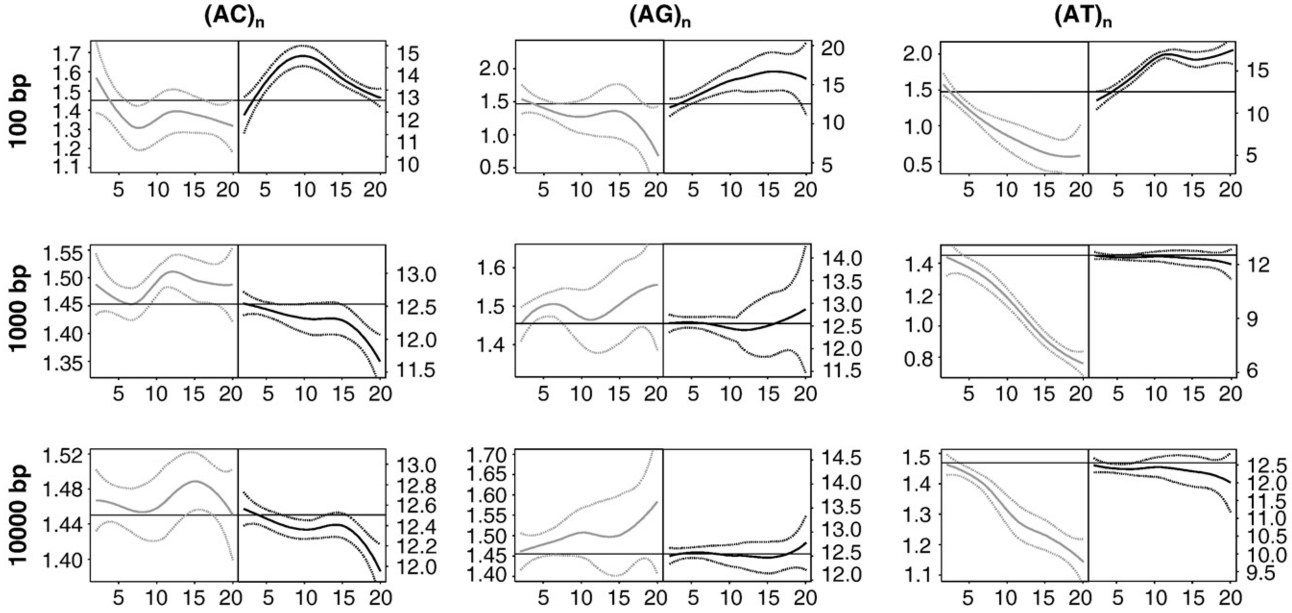

Fig. 6. Summary for how SNP density (grey line) and human–chimpanzee divergence (bold line) vary with microsatellite length over three different regions: 100 bp either side of the microsatellite, 1 kb either side but excluding the nearest 100 bp and 10 kb either side excluding the nearest 1 kb. Dotted lines represent 95% confidence intervals of the best fit local regression of all points. Horizontal lines represent both the autosomal average divergence and SNP density. SNP density and divergence are measured in number of SNPs and substitutions per kb, respectively.

(www.hapmap.org). These SNPs correspond to the HapMap Public Release #26 in NCBI build 36 (dbSNP b126) coordinates.

Microsatellite identification

We focused on the most common dinucleotide motifs $(AC)_n$, $(AG)_n$ and $(AT)_n$. Exhaustive searches were then conducted on the entire human genome, excluding the two sex chromosomes, using a custom program written in AWK. To avoid biases due to line breaks, groups of five lines at a time are concatenated and searches conducted only in the central 100 bases. When a repeat is found its coordinates are stored. Only pure repeat tracks of the form $(N)_{50}(XY)_n(N)_{50}$ were retained, where $(N)_{50}$ is a tract of 50 flanking bases in which $(XY)_2$ does not appear and $(XY)_n$ represents n repeats of a microsatellite with motif XY. Complementary and alternative frame motifs were ignored, i.e. when searching for $(AC)_n$ we ignored $(TG)_n$ and $(CA)_n$, unless the latter qualified in its own right. Using the AWK program we generated output files, one each for $(AC)_n$, $(AG)_n$ and $(AT)_n$, where n varied between 2 and 20 repeats. These contain the overwhelming majority of all such repeats in the human genome.

Flanking sequence analysis

To explore the ways in which SNP density varies around any given class of microsatellite, SNPs were counted in a series of 10 equal-sized bins on either side. Since the scale over which patterns may be found is unclear, we examined two different bin sizes, 100 bp and 1 kb, providing information on SNP density 1 and 10 kb either side of the microsatellite respectively. For a broader context still, determined *a posteriori*, we also analysed 50 kb either side using a bin size of 1 kb. To avoid counting the same SNPs twice in any given analysis with the same motif and length, we moved methodically through the genome, accepting only those loci that lie outside the search area of the previously analysed locus. Thus, at the broadest resolution of 50 kb either side of a microsatellite, a new locus is only accepted if it is at least 100 kb distant from the previous one.

Given any particular pattern of SNP densities, it is interesting to place this in a broader temporal context. For this we compared the flanking sequences of orthologous human and chimpanzee microsatellites using tools at the Galaxy website (http://galaxy.psu.edu/) [44]. First we extracted pairwise alignments with the chimpanzee genome (panTro2 March 2006) for each chromosome using the tool Fetch alignments. Then, the genomic position of each base substitution was determined using the Regional Variation tool. Subsequently, the number of substitutions was determined at both the 100 bp bin and 1 kb bin sizes, using our list of non-adjacent microsatellites derived above. Finally, we calculated the GC content in the 100 bp flanking sequence either side of microsatellites in human chromosome 1 using the tool “Geecee” at the Galaxy website (http://galaxy.psu.edu/).

Statistical analysis

All statistical analyses were conducted in R (http://www.r-project.org/). For each given microsatellite motif and bin size, 3D plots of SNP density relative to the microsatellite were constructed using local spline fitting to generate a smoothed surface, as implemented in the ‘locfit’ function. 3D spline-fitting can over-smooth fine-scale patterns and make it difficult both to determine and to display confidence intervals. Consequently, we constructed 2D slices through the 3D graph, taken at 2, 5 and 20 repeats, using the command ‘crit’ to determine the best-fit local regression along with

Table 2
GC content (%) of sequences flanking all microsatellites AC, AG and AT of human chromosome 1 based on the 100 bases either side of each microsatellite. Sequences flanking AT microsatellites have a lower GC content than sequences flanking AG and AC. Furthermore, the longer AT microsatellites are the more likely they have a low GC content in their flanking sequences.

Repeats	GC content	SEM
$(AT)_2$	39.77	0.06
$(AT)_5$	36.81	0.27
$(AT)_{20}$	34.13	1.21
$(AG)_2$	42.36	0.06
$(AG)_5$	42.4	0.28
$(AG)_{20}$	42	5.2
$(AC)_2$	41.97	0.06
$(AC)_5$	39.42	0.3
$(AC)_{20}$	42.07	0.69

95% confidence intervals. For comparisons both with random expectations and between different motifs and different repeat numbers, the autosomal average SNP density was included on all 2D plots, calculated as the total number of SNPs (3907239, HapMap Public Release #26 in www.hapmap.org) divided by the number of nucleotides sequenced (2681518154, see Build 36.3 statistics in www.ncbi.nlm.nih.gov). To compare substitution rates among sequences flanking microsatellites we used similar methods to those used for SNPs, and we included the autosomal average substitution rate on all 2D plots calculated as the total number of substitutions (33829759) in autosomal chromosomes of the human genome (build36/hg18 March 2006), and the chimpanzee genome (panTro2 March 2006) divided by the number of nucleotides sequenced in the human genome (2681518154, see Build 36.3 statistics in www.ncbi.nlm.nih.gov).

Acknowledgments

M.V. was awarded an *Angeles Alvariño* fellowship by *Xunta de Galicia* (IN840D and IN809A).

References

[1] A.J. Brookes, The essence of SNPs, Gene 234 (1999) 177–186.

[2] R.D. Miller, M.S. Phillips, I. Jo, M.A. Donaldson, J.F. Studebaker, N. Addleman, S.V. Alfisi, W.M. Akener, H.A. Bhatti, C.E. Callahan, B.J. Carey, C.L. Conley, J.M. Cyr, V. Derohannessian, R.A. Donaldson, C. Elosua, S.E. Ford, A.M. Forman, C.A. Gelfand, N.M. Grecco, et al., High-density single-nucleotide polymorphism maps of the human genome, Genomics 86 (2005) 117–126.

[3] P.C. Sabeti, P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E.H. Byrne, S.A. McCarroll, R. Gaudet, S.F. Schnaffer, E.S. Lander, T.I.H. Consortium, Genome-wide detection and characterization of positive selection in human populations, Nature 449 (2007) 913–919.

[4] The International HapMap Consortium, A second generation human haplotype map of over 3.1 million SNPs, Nature 449 (2007) 851–853.

[5] R. Sainudiin, A.G. Clark, R.T. Durrett, Simple models of genomic variation in human SNP density, BMC Genomics 8 (2007) 146.

[6] J.W. Drake, A. Bebenek, G.E. Kissling, S. Peddada, Clusters of mutations from transient hypermutability, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 12849–12854.

[7] D.C. Koboldt, R.D. Miller, P.-Y. Kwok, Distribution of human SNPs and its effect on high throughput genotyping, Hum. Mut. 27 (2006) 249–254.

[8] B. Charlesworth, M. Nordberg, D. Charlesworth, The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations, Genet. Res. Camb. 70 (1997) 155–174.

[9] B.F. Voight, S. Kudaravalli, X. Wen, J.K. Pritchard, A map of recent positive selection in the human genome, PLoS Biol. 4 (2006) e72.

[10] E. Wang, G. Kodama, P. Balbi, R.K. Moyzis, Global landscape of recent inferred Darwinian selection for *Homo sapiens*, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 135–140.

[11] M.J. Lercher, L.D. Hurst, Human SNP variability and mutation rate are higher in regions of high recombination, Trends Genet. 18 (2002) 337–340.

[12] I. Hellmann, K. Prüfer, H. Ji, M.C. Zody, S. Pääbo, S. Ptak, Why do human diversity levels vary at a megabase scale? Genome Res. 15 (2005) 1222–1231.

[13] A. Kong, D.F. Gudbjartsson, J. Sainz, G.M. Jonsdottir, S.A. Gudjonsson, B. Richardsson, S. Sigurdottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S.T. Palsson, M.L. Frigge, T.E. Thorgeirsson, J.R. Gulcher, K. Stefansson, A high-resolution recombination map of the human genome, Nat. Genet. 31 (2002) 241–247.

[14] B.A. Payseur, M.W. Nachman, Microsatellite variation and recombination rate in the human genome, Genetics 156 (2000) 1285–1298.

[15] M.W. Nachman, Single nucleotide polymorphisms and recombination rate in humans, Trends Genet. 17 (2001) 481–485.

[16] L. Duret, P.F. Arndt, The impact of recombination on nucleotide substitutions in the human genome, PLoS Genet. 4 (2008) e10000071.

[17] E. Biet, J.-S. Sun, M. Dutriex, Conserved sequence preference in DNA binding among recombination proteins: an effect of ssDNA secondary structure, Nucleic Acids Res. 27 (1999) 596–600.

[18] Y.-C. Li, A.B. Korol, T. Fahima, A. Beiles, E. Nevo, Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review, Mol. Ecol. 11 (2002) 2453–2465.

[19] W.-J. Guo, J. Ling, P. Li, Consensus features of microsatellite distribution: microsatellite contents are universally correlated with recombination rates and are preferentially depressed by centromeres in multicellular eukaryotic genomes, Genomics 93 (2009) 323–331.

[20] D. Tian, Q. Wang, P. Zhang, H. Araki, S. Yang, M. Kreitman, T. KNagylaki, R. Hudson, J. Bergelson, J.-Q. Chen, Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes, Nature 455 (2008) 105–109.

[21] M. Legendre, N. Pochet, T. Pak, K.J. Verstrepen, Sequence-based estimation of minisatellite and microsatellite repeat variability, Genome Res. 17 (2007) 1787–1796.

[22] A. Bacolla, J.E. Larson, J.R. Collins, J. Li, A. Milosavljevic, P.D. Stenson, D.N. Cooper, R.D. Wells, Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties, Genome Res. 18 (2008) 1545–1553.

[23] Y.D. Kelkar, S. YTyekucheva, F. Chlaromonte, K. Makova, The genome-wide determinants of human and chimpanzee microsatellite evolution, Genome Res. 18 (2008) 30–38.

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

[25] E.J. Vowles, W. Amos, Evidence for widespread convergent evolution around human microsatellites, PLoS Biol. 2 (2004) e199.

[26] M.A. Varela, W. Amos, Evidence for non-independent evolution of adjacent microsatellites in the human genome, J. Mol. Evol. 68 (2009) 160–170.

[27] M.A. Varela, R. Sanmiguel, A. Gonzalez-Tizon, A. Martinez-Lage, Heterogeneous nature and distribution of interruptions in dinucleotides may indicate the existence of biased substitutions underlying microsatellite evolution, J. Mol. Evol. 66 (2008) 575–580.

[28] A.R. Rogers, L.B. Jorde, Ascertainment bias in estimates of average heterozygosity, Am. J. Hum. Genet. 58 (1996) 1033–1043.

[29] H. Ellegren, C.R. Primmer, B.C. Sheldon, Microsatellite evolution: directionality or bias in locus selection, Nature Genet. 11 (1995) 360–362.

[30] J.-Q. Chen, Y. Wu, H. Yang, J. Bergelson, M. Kreitman, D. Tian, Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria, Mol. Biol. Evol. 26 (2009) 1523–1531.

[31] M. Brandström, H. Ellegren, Genome-wide analysis of microsatellite polymorphism in chicken circumventing ascertainment bias, Genome Res. 18 (2008) 881–887.

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

[33] S. Kruglyak, R.T. Durrett, M.D. Schug, C.F. Aquadro, Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 10774–10778.

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

[35] X. Xu, M. Peng, Z. Fang, X. Xu, The direction of microsatellite mutations is dependent upon allele length, Nat. Genet. 24 (2000) 396–399.

[36] M.T. Webster, J. Hagberg, Is there evidence of convergent evolution around human microsatellites? Mol. Biol. Evol. 24 (2007) 1097–1100.

[37] N.N. FitzSimmons, C. Moritz, S.S. Moore, Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution, Mol. Biol. Evol. 12 (1995) 432–440.

[38] C. Rico, I. Rico, G. Hewitt, 470 million years of conservation of microsatellite loci among fish species, Proc. R. Soc. Lond. B 263 (1996) 549–557.

[39] S. Ptak, D.A. Hinds, K. Koehler, B. Nickel, N. Patil, D.G. Ballinger, M. Przeworski, K.A. Frazer, S. Pääbo, Fine-scale recombination patterns differ between chimpanzees and humans, Nat. Genet. 37 (2005) 429–434.

[40] W. Winckler, S.R. Myers, D.J. Richter, R.C. Onofrio, G.J. McDonald, R.E. Botrop, G.A.T. McVean, S.B. Gabriel, D. Reich, P. Donnelly, et al., Comparison of fine-scale recombination rates in humans and chimpanzees, Science 308 (2005) 107–111.

[41] A. Hodgkinson, E. Ladoukakis, A. Eyre-Walker, Cryptic variation in the human mutation rate, PLoS Biol. 7 (2009) e1000027.

[42] M. Brandström, A.T. Bagshaw, N.J. Gemmell, H. Ellegren, The relationship between microsatellite polymorphism and recombination hot spots in the human genome, Mol. Biol. Evol. 25 (2008) 2579–2587.

[43] M.F. Santibáñez-Koref, R. Gangeswaran, J.M. Hancock, A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes, Mol. Biol. Evol. 18 (2001) 2119–2123.

[44] P. Schattner, Genomics made easier: an introductory tutorial to genome datamining, Genomics 93 (2009) 187–195.