Genome Nucleotide Composition Shapes Variation in Simple Sequence Repeats

Xiangjun Tian,1 Joan E. Strassmann,1 and David C. Queller*1

1Department of Ecology and Evolutionary Biology, Rice University

*Corresponding author: E-mail: queller@rice.edu.

Associate editor: Hideki Innan

Abstract

Simple sequence repeats (SSRs) or microsatellites are a common component of genomes but vary greatly across species in their abundance. We tested the hypothesis that this variation is due in part to AT/GC content of genomes, with genomes biased toward either high AT or high CG generating more short random repeats that are long enough to enhance expansion through slippage during replication. To test this hypothesis, we identified repeats with perfect tandem iterations of 1–6 bp from 25 protists with complete or near-complete genome sequences. As expected, the density and the frequency are highly related to genome AT content, with excellent fits to quadratic regressions with minima near a 50% AT content and rising toward both extremes. Within species, the same trends hold, except the limited variation in AT content within each species places each mainly on the descending (GC rich), middle, or ascending (AT rich) part of the curve. The base usages of repeat motifs are also significantly correlated with genome nucleotide compositions: Percentages of AT-rich motifs rise with the increase of genome AT content but vice versa for GC-rich subgroups. Amino acid homopolymer repeats also show the expected quadratic relationship, with higher abundance in species with AT content biased in either direction. Our results show that genome nucleotide composition explains up to half of the variance in the abundance and motif constitution of SSRs.

Key words: comparative genomics, nucleotide composition, simple sequence repeat, microsatellite.

Introduction

Simple sequence repeats (SSRs) are DNA tracts comprising perfect or near-perfect tandem iterations of short DNA motifs among which the repeats with motifs of 1–6 bp are generally called microsatellites. These kinds of repeated DNA sequences are very common in eukaryotes (Toth et al. 2000; Katti et al. 2001). They are less often found in prokaryotes but are not entirely absent (Mrazek et al. 2007). For example, microsatellites are present in a bacterium, Mycoplasma genitalium, that was thought to be the smallest self-replicating organism (Hancock 1996b) as well as in plant mitochondria and chloroplasts (Allender et al. 2007; Rajendrakumar et al. 2007).

SSRs are believed to be one of the most variable types of DNA sequences in genomes, with variation in number of repeats primarily attributed to slipped-strand mispairing and the subsequent excision and repair during DNA replication (Levinson and Gutman 1987). Microsatellites within noncoding regions have extensive length variation and are often assumed to form and evolve neutrally. Therefore, they are often used as molecular markers for genome mapping (Danin-Poleg et al. 2000; Bindler et al. 2007; Solignac et al. 2007) and DNA fingerprinting (Garcia-Martinez et al. 2006). SSRs are also applied in population genetics to determine population divergence and to survey gene flow between populations (Ellegren 2004; Allender et al. 2007; Vali et al. 2008) and genetic relatedness (Queller et al. 1993; Webster and Reichart 2005). However, in coding regions, the generation and expansion of SSRs are usually restricted by selection against frameshift mutations except for triplet (tri- or hexa-) nucleotide repeats (Metzgar et al. 2000; Metzgar and Wills 2000). Mutations in the number of triplet repeats may be tolerated by the gene but may also cause deleterious effects on the host. The most well-known cases are human “triplet repeat diseases,” such as neurodegenerative disorders and cancers caused by unstable triplet repeats (Reddy and Housman 1997; Arzimanoglou et al. 1998; Karlin et al. 2002; Perutz et al. 2002).

The frequency, density, and distribution of microsatellites vary widely in genomes of different species. The human and mouse genomes show quite different patterns (Lander et al. 2001; Waterston et al. 2002), but there is also considerable variation within genera, such as Drosophila (Pascual et al. 2000; Schlötterer and Harr 2000) and Mycoplasma (Mrazek 2006). The frequency and the density of SSRs are sometimes correlated with genome sizes. For example, the SSR density is higher in large genomes than in small genomes among mammals (Ellegren 2004). However, among plants, the SSR frequency is lower in large genomes than in small genomes (Morgante et al. 2002). The distributions of SSRs also vary in different regions along genome sequences. It is well known that noncoding regions generally contain more abundant SSRs than coding regions (Hancock 1995, 1996a). However, there is no apparent difference of SSR contents between intergenic regions and introns (Weber 1990). In addition, the SSR density is higher at the end of chromosome arms than at other regions in...
human and mouse genomes (Lander et al. 2001; Waterston et al. 2002), and it is higher on autosomes than on X chromosomes in mammals (such as humans, mice, and rats) but vice versa in Drosophila (Bachtrog et al. 1999).

Whole-genome surveys of frequency, density, and distribution of microsatellites are one of the principal means of seeking clues of the genesis and evolution of repeats (Gur-Arie et al. 2000; Pascual et al. 2000; Schlötterer and Harr 2000; Toth et al. 2000; Katti et al. 2001; Karaoğlu et al. 2005; Mrázek 2006; Mrázek et al. 2007; Brandstrom and Ellegren 2008). However, little is understood yet about the wide variation in SSR abundance between species. We propose that genomic base composition may be a major cause of this variation, with both AT-rich and GC-rich genomes being more prone to evolve SSRs than genomes of balanced composition. The logic is simple. SSRs evolve through slippage, and though there is some evidence that slippage can begin at very low repeat numbers (Zhu, Queller, et al. 2000; Zhu, Strassmann, et al. 2000), it seems likely that many SSRs first arise by chance substitutions that make a short repetitive sequence, which can undergo slippage if they are above some threshold size (Levinson and Gutman 1987; Stephan and Cho 1994; Messier et al. 1996; Rose and Falush 1998; Bachtrog et al. 1999). In this two-part process of initiation and elongation, we hypothesize that SSR frequency and density are determined in large part by how easy it is to get random initial SSRs. Generation of these initial short random SSRs is easier with biased nucleotide compositions. In the extreme, a genome that is completely AT (or completely CG) will contain many more short random repeats than a balanced genome. For example, for a given AT doublet, the probability that it will be followed by three more random AT doublets is 0.5^4 = 0.016 for the two-base code of an AT-only genome and 0.25^6 = 0.00022 for a balanced four-base code. The probability for a genome of intermediate AT composition will of course be between these values. We illustrate this in more detail with simulations below. This factor has been previously noted (Dieringer and Schlötterer 2003), but it was used primarily as a correction factor in examining other influences on SSR abundance. Here, our hypothesis is that nucleotide content is a key variable of interest in its own right, which can be viewed as a major driver of SSR abundance. Similar explanations have been proposed for low-complexity regions within proteins, which are sometimes caused by SSRs (Xue and Forsdyke 2003; DePristo et al. 2006) and for codon bias (Knight et al. 2001). Here, we test this hypothesis for SSRs, both in protein-coding regions and in noncoding regions.

We chose to focus on single-celled eukaryotes or protists for several reasons. Prokaryotes have relatively few SSRs. Multicellular eukaryotes have large amounts of noncoding DNA whose composition may be dominated by other processes. For example, transposition may generate much sequence, and SSRs may be associated with some of these elements (Arcot et al. 1995; Ramsay et al. 1999; Temnykh et al. 2001). Moreover, many of their genomes have local isochores with quite variable AT content, so that genomic AT content may not be a good indicator of AT content in particular areas. Protists tend to have less DNA that may be junk DNA and tend to have genomes that have more uniform AT content, so that the average AT content of the genome is a good indicator of local AT conditions throughout the genome. Protists also give us a good range of genome sizes, AT contents, and SSR abundance for testing our hypotheses. With 25 complete or near-complete genome sequences of protist organisms, we explored and analyzed 1–6 bp SSRs with tract length of over 12 bp, focusing on the effects of genome nucleotide composition on the frequency, density, and base usage of SSRs.

Materials and Methods

Sequence Data

Genome sequences were retrieved from the public databank and the release homepage of the genome sequencing projected of some individual organisms (as of 24 March 2009). Sequences of Cryptosporidium hominis, C. parvum, Entamoeba histolytica, Giardia lamblia, Plasmodium falciparum, P. yoelii, Theileria parva, and Trypanosoma brucei were retrieved from the version 2.7 release of microbial genomes (http://img.jgi.doe.gov) system of Department of Energy Joint Genome Institute (Markowitz et al. 2008). Sequences of Chlamydomonas reinhardtii, Monosiga brevicollis, Naegleria gruberi, Ostreococcus lucimarinus, O. tauri, Phaeodactylum tricornutum, Pythophthora ramorum, P. sojae, and Thalassiosira pseudonana were retrieved from the genome sequencing projects of JGI eukaryotic genomes (http://genome.jgi-psf.org/). Sequences of Leishmania major, L. braziliensis, and L. infantum were retrieved from the release of genome projects of the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/). Sequences of Trichomonas vaginalis were retrieved from the genome sequencing projects of The Institute for Genomic Research (http://www.tigr.org/tdb/e2k1/tvg/). Sequences of Cyanidioschyzon merolae were retrieved from the C. merolae Genome Project of National Institute of Genetics, Japan (http://merolae.biols.u-tokyo.ac.jp/). Sequences of Dictyostelium discoideum were retrieved from DictyBase (http://dictybase.org/). Sequences of Paramecium tetraurelia were retrieved from the release by Genoscope (http://www.cns.fr/). Sequences of P. infestans were retrieved from the P. infestans database released by the Broad Institute (http://www.broadinstitute.org/annotation/genome/pythophthora_infestans/MultiHome.html).

Random DNA sequences were simulated at each of nine AT contents (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%). One hundred megabases were generated for each. The Perl script for simulating random DNA sequences is available on request.

SSR Searches

In both the protist genomes and in the random sequences, we searched for 1–6 bp perfect SSRs that are over 12 bp in length. We chose a length threshold rather than a repeat number threshold because microsatellite variability, and
presumably the underlying dynamics of replication slippage, depend more on the former (i.e., a 12-repeat dinucleotide is expected to be more similar in variability to a tetranucleotide with 6 repeats than one with 12 repeats). The thresholds for each type of SSR are therefore the following: mononucleotide SSRs with at least 13 repeats, dinucleotide SSRs with at least 7 repeats, trinucleotide SSRs with at least 5 repeats, tetranucleotide SSRs with at least 4 repeats, penta- and hexanucleotide SSRs with at least 3 repeats. We grouped motifs with a common repeat structure, for example, AT with TA or CAA with ACA and AAC and chose the repeat motif that constructs the longest repeat. The SSR identification was implemented by using an in-house Perl script, which is available on request.

Calculation of SSR Frequency and Density
We defined the number of any type of SSR by counting the total number of uninterrupted SSR sequences of that type that exceeded 12 bp in length. We defined the length of a type of SSR by summing the total length of all counted SSR sequences of that type. The frequency of a type of SSR was measured as the average number of SSRs per megabase:

$$F_i = \frac{N_i}{G} \times 1,000,000,$$

where $F_i$ was the frequency of the SSR with unit size of $i$ (1 ≤ $i$ ≤ 6), $N_i$ was the total number of SSR motifs with unit size $i$, and $G$ is the size of the whole-genome sequence (or of coding or noncoding sequence). The SSR density was measured by the total SSR length per megabase:

$$D_i = \frac{L_i}{G} \times 1,000,000,$$

where $D_i$ was the density of the SSR with unit size of $i$ (1 ≤ $i$ ≤ 6) and $L_i$ was the total length of SSR motifs with unit size $i$.

Grouping of Motifs by AT content
For each type of microsatellite, motifs were grouped by AT content. If the unit size of a type of SSR is $n$, its motifs were categorized into $n + 1$ subgroups that are AT0, AT1, . . . , and ATn, where the number indicates the number of A’s or T’s in a motif. The percentage of each subgroup of a type of SSR (unit size $n$) was calculated by

$$p_i = \frac{c_i}{\sum_{j=0}^{n} c_j} \times 100\%,$$

where $p_i$ was the percentage of a subgroup SSR with AT = $i$ out of all subgroups of the SSRs of unit size $n$, $c_i$ was the number of SSRs with AT = $i$, and $\sum_{j=0}^{n} c_j$ (where 0 ≤ $j$ ≤ $n$) was the total number of SSRs with unit size $n$.

Within-Species Local AT Content and SSR Frequency and Density
To test whether variation within genomes in AT content correlates with SSR frequency and density, we calculated these measures as above but for sliding windows within each genome. We used nonoverlapping 100 kbp windows. To examine this correlation separately for coding and noncoding sequences, we separately concatenated all the coding sequence and all the noncoding sequence and used new 100 kbp windows for each.

Correlation Between Species of Amino Acid Usage in Homopolymers
To test whether species with similar AT contents have similar amino acid usage in their homopolymer amino acid repeats, we calculated a correlation between each pair of species. For each species, we found all the amino acid homopolymers (of length at least five amino acids) and summed up their lengths. The data points for a species were the fractional contributions of each amino acid to that total (e.g., the total might contain 2.6% alanine, 1.4% arginine, and so on, summing to 100%). These data points were then correlated for each species pair. Species pairs with identical amino acid percentages in their homopolymers would have a correlation of one.

Statistics
Quadratic regressions of SSR frequency and SSR density on sequence AT content were performed using R as was the regression of percentages of SSR subgroups on genome AT contents. The goodness of fit of a quadratic regression was assessed by $R^2$, f-test for checking the statistical significance of regression, and t-test for checking the statistical significance of coefficients. The Student’s t-test and Pearson’s correlation coefficient were also performed using R.

The correlations of AT content and SSR frequency and density were phylogenetically corrected by phylogenetic generalized least squares (PGLS) regressions (Martins and Hansen 1997). The phylogeny of the 25 single cell eukaryotes was constructed by maximum parsimony method based on the combined protein sequences of alpha- and beta-tubulin genes as used by Baldauf et al. (Baldauf 1999; Baldauf et al. 2000). The parsimony tree was derived from Mega 4 with a bootstrap test of 1,000 replicates (Kumar et al. 2008). The PGLS quadratic regressions were performed by the R package “ape” (Paradis et al. 2004) based on a standard distance matrix built from the phylogeny by fitting the Brownian Motion model.

Results
SSR Contents in Random DNA Sequences
We randomly generated nine simulated DNA sequences that are 100 Mbp in sequence length, each with a fixed AT content, ranging from 10% to 90%. All 1–6 bp SSRs longer than 12 bp were identified using the same method and criteria as we used for real protist genomes (supplementary table S1, Supplementary Material online). As expected, the number of occurrences (per megabase) and the density (per megabase) of all types of SSRs are significantly related to sequence AT content. Quadratic regressions were all significant and explained most of the variance ($R^2 > 0.80$,
SSR Contents in Protist Genomes

We identified 1–6 bp SSRs (microsatellites) longer than 12 bp in genome sequences of 25 protists with AT content ranging from 35% to 80% (supplementary table S3, Supplementary Material online) and calculated counts and lengths for each type of SSR for the whole-genome sequence (supplementary table S4, Supplementary Material online), coding sequence (supplementary table S5, Supplementary Material online), and noncoding sequence (supplementary table S6, Supplementary Material online). All counts and total length of SSRs varied widely across species, ranging across about three orders of magnitude. As shown in supplementary table S4, Supplementary Material online, the social amoeba *D. discoideum* has the largest number (133,201) and length (3.07 Mbp, occupying 9.1% of the whole-genome sequence) of overall SSRs among our sample collections, followed by *C. reinhardtii* and *P. falciparum*.

Interestingly, though *C. reinhardtii* has the third largest genome in our data set, the other two species with highest SSR abundance are below average in genome size. The data show that in general, genome size is not significantly associated with either the total number or the total length of SSRs (supplementary fig. S1, Supplementary Material online). The same holds for both coding sequences alone and noncoding sequences alone.

Of the individual types of SSRs in coding sequences, trinucleotide SSRs predominate for most of the species, followed by hexanucleotide SSRs, except for a few species with predominant di- or mononucleotide SSRs, such as *M. brevicollis*, *P. falciparum*, and *P. yoelii* (supplementary table S5, Supplementary Material online). In noncoding regions, mono- or dinucleotide SSRs are more often the most abundant type, though trinucleotide repeats are most common in a number of species (supplementary table S6, Supplementary Material online). Among all the genomes, *D. discoideum* led in the number and length of mono- and trinucleotide SSRs, *P. falciparum* for dinucleotide SSRs, and *C. reinhardtii* for tetra-, penta-, and hexanucleotide SSRs (supplementary table S4, Supplementary Material online).

Effects of Nucleotide Composition on SSR Frequency and Density

To evaluate the effect of genome nucleotide compositions on the frequency of SSRs, we examined the relationship between AT content of genome sequences and the number of SSRs per megabase. The quadratic regressions of frequencies of overall SSRs on AT content were significant and explained about half the variance at the whole-genome level, and for coding and noncoding sequences separately (for all three, $R^2 \geq 0.49$ and $P < 0.001$; see fig. 2A and supplementary table S2, Supplementary Material online). As with the simulated sequences, the quadratic coefficients are all significantly positive ($P \leq 0.001$), indicating greater SSR frequency at more extreme base compositions. We also explored the regressions of the frequency of each type of SSR on AT content, again at the levels of the whole-genome sequence, coding sequence, and noncoding sequence (fig. 2B–D). These separate correlations also fit well with quadratic regressions ($P < 0.05$ for regression functions, except for one $P = 0.054$; and $P < 0.05$ for all quadratic coefficients. For detail, see supplementary table S2, Supplementary Material online).
The graph of a quadratic function \( y = ax^2 + bx + c \) is a parabola, where the quadratic coefficient \( a \) controls the increase/decrease speed of the function from its vertex, and the coefficients \( a \) and \( b \) together control the axis of symmetry of the parabola. Therefore, we used the quadratic coefficient \( a \) to measure the effect of extreme base compositions on increasing SSR frequencies. This generally paralleled frequency; where SSRs were more frequent, their \( a \) coefficients were larger. For example, SSRs are more frequent in noncoding regions than in coding regions \((P = 0.016, \text{paired Student's} t\text{-test})\), and SSR frequency was more driven by sequence AT content in noncoding \((a = 7.56)\) regions than in coding \((a = 2.11)\) regions (supplementary table S2, Supplementary Material online, fig. 2A). Of the individual types of SSRs, mono- and dinucleotides have the highest \( a \) coefficients and are more frequent than the other four types at the whole-genome level, with the former being dominant in high-AT genomes and the latter in high-GC genomes (fig. 2B). In coding sequences, trinucleotides have the highest \( a \) coefficient and are generally the most frequent type (fig. 2C). In noncoding sequences, mono- and dinucleotide SSRs have the highest \( a \) coefficients and are generally the most common (fig. 2D).

In a modest difference from our simulation results, the parabolas for the various SSR types have their vertex or the minimum value (calculated as \(-b/2a\) in supplementary table S2, Supplementary Material online) slightly above 50% AT content. For coding sequences, they fell in the range 50–55% and for noncoding sequences 55–60%.

Very similar results apply for SSR densities or the average lengths of SSRs per megabase. Supplementary figure S2A, Supplementary Material online, shows the SSR densities of these protist genomes and their regressions on genome AT content for the whole-genome sequence, coding sequence, and noncoding sequence. The regressions also explain about half the variance and are significant \((R^2 \geq 0.46\) and \(P \leq 0.01\)). The quadratic coefficients are also significant and positive \((P \leq 0.01; \text{see supplementary table S2, Supplementary Material online})\), indicating that genomes with balanced AT content have the lowest SSR densities, whereas genomes with biased AT or GC content have higher SSR densities. Consistent with prior reports (Hancock 1995), SSR density is higher in noncoding regions than in coding regions \((P = 0.014, \text{paired Student's} t\text{-test})\), especially for genomes with greater GC or AT biases (supplementary fig. S2A, Supplementary Material online).

For individual types of microsatellites varying in motif length (mono- through hexanucleotide), the density was also highly correlated with AT content at the whole-genome level and in coding and noncoding sequences (supplementary fig. S2B–D, Supplementary Material online). The densities of individual types of SSRs are well explained by a quadratic regression on AT content \((P < 0.05\) for regression functions except for two of 0.057 and 0.054; and \(P \leq 0.05\) for all quadratic coefficients. For detail, see supplementary table S2, Supplementary Material online). However, the effect of AT content on SSR densities varied
for different individual types. In coding regions, tri- and hexanucleotide SSRs are generally the densest types with the highest quadratic coefficients except for a few species that have denser dinucleotide SSRs (*M. brevicollis* and *P. falciparum*). In noncoding regions, the density of mono- and dinucleotide SSRs is highest in the most AT-rich and GC-rich protists, respectively, except for two species (*C. hominis* and *N. gruberi*) that have hexanucleotides as the densest kind.

To remove the effect of phylogenetic similarity on the relationships between SSR frequency/density and AT content, we applied PGLS regression analysis on coding sequence SSR. We built phylogenetic distance matrix (supplementary table S8, Supplementary Material online) from a maximum parsimony phylogeny by fitting the Brownian Motion model, which was used in the PGLS regression. The resulting quadratic correlations between AT content and frequency/density of most SSRs are significant after taking the phylogenetic distance into the analysis (*P* < 0.01 for coefficients of AT² and AT except the dinucleotide SSR. See supplementary table S9, Supplementary Material online for details).

Effects of Nucleotide Composition on SSR Motif Constitution

If our hypothesis is correct, the relationship between AT content and SSR abundance is mediated by the generation of more random AT-rich repeats in high-AT genomes and more random GC-rich repeats in high-GC genomes. We should therefore see more usage of AT-rich repeat motifs in high-AT genomes and more usage of GC-rich motifs in high-GC genomes. To assess the effect of genome nucleotide composition on the constitution of SSR motifs, we

![Fig. 3. AT content of protist SSR motifs as a function of different genome AT content. Panel A, B, C, D, E, and F are for mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats, respectively. Each shows best-fit quadratic regressions for low-AT motifs (Black), intermediate-AT motifs (light gray—the curves with central maxima), and high-AT motifs (darker gray). For each type, AT₀ is a repeat unit with no A’s or T’s, AT₁ is a repeat with 1 A or T, and so on. The solid lines are for the genome data of 25 protist species and the dashed lines are for simulated random sequences (for details of regression functions, see supplementary table S7, Supplementary Material online).](http://mbe.oxfordjournals.org/)

categorized each type of SSR with repeat unit size \( n \) into \( n + 1 \) subgroups by the number of A and T nucleotides, that is, with subgroups being \( AT = 0 \), \( AT = 1 \), \ldots, and \( AT = n \) (\( 1 \leq n \leq 6 \)). For each SSR type, we grouped the motifs into low-AT, intermediate-AT, and high-AT subgroups calculated the percentages of each subgroup and then regressed these percentages on genome AT contents. Figure 3 shows these regressions for the 25 protists (solid lines) and for random sequences (dashed lines). The subgroup percentages show significant correlations with genome AT contents, fitting well with quadratic regressions (\( P, 0.05 \) for all but two, see supplementary table S7, Supplementary Material online for detail). Clearly, the curves for protist genomes are very similar to those for random sequences. Motifs that are low AT and high AT are most common for low-AT and high-AT genomes, respectively. Intermediate-AT motifs generally have peak abundance in AT-balanced genomes.

**Fig. 4.** Correlation between local AT content and local SSR frequency and density within species. Panel A, schematic diagram shows the expected correlation between SSR abundance and AT content in genomes that cover only part of the range of AT content. Dashed lines show the expected correlations, from left to right for GC-biased genomes, balanced genomes, and AT-biased genomes. Panel B, within each species, AT content and SSR frequencies and densities were calculated in sliding windows of 100 kbp and a step size of 100 kbp. The Pearson’s correlation coefficients for each species are plotted against genome-wide AT content, showing the correlations are as expected from part A: positive in AT-biased genomes, negative in GC-biased genomes, and near zero in AT-balanced genomes. The solid and dashed lines indicate the best-fit linear regressions of the Pearson’s correlation coefficients on genome AT content for SSR density and SSR frequency, respectively.

**Effects of Nucleotide Composition on SSR Variation Within Genomes**

The large differences in AT content between genomes give us a good range and power for our tests. We expect the same processes to operate within genomes, but the variation in AT content will be much lower, resulting in loss of power. Moreover, a single genome will generally not contain both very AT-rich and very GC-rich areas, so that we cannot test for the full U-shaped curve. Instead, we make different predictions according to which portions of the U-shaped curve a genome covers. As shown in figure 4A, in a genome that is AT rich (right or ascending portion of the curve), there should be a positive correlation between local AT content and both SSR frequency and density. For a genome that is GC rich (left or descending portion of the curve), the correlation should be negative. For a genome with intermediate AT content (central or flattest part of the curve), the correlation should be weak or nonexistent.

Figure 4B shows the estimated within-species correlations for both SSR frequency (open circles) and density (solid circles) as a function of average genome AT content. The results support our predictions, negative correlations within CG-rich genomes, no correlation within balanced genomes, and positive correlations within AT-rich genomes. The same relationships hold if tested separately for coding and noncoding sequences, though the \( R^2 \) values are somewhat lower (but still above 0.55, supplementary fig. S3, Supplementary Material online).

**Effects of Nucleotide Composition on Homopolymer Amino Acids**

SSRs in coding regions encode repeats of amino acid residues. Trinucleotide repeats in particular will encode homopolymer runs of single amino acids. Because genome AT content affects SSRs, it may also affect amino acid repeats. To detect the effect of genome nucleotide composition on the generation and expansion of amino acid repeats, we assayed the correlation between AT content of the protist coding sequences and the density and frequency of homopolymer amino acids. The total length of homopolymer amino acids per kilo-residue (density) in the whole proteome is dependent on the AT content of coding sequences, following a quadratic regression with positive quadratic coefficient (fig. 5A). The fit is significant (\( P < 0.05 \), \( R^2 = 0.30 \)). The frequency of homopolymer runs per kilo-residue in the whole proteome shows a similar quadratic relationship as the density.

The types of amino acids in homopolymer runs are also affected. Genomic AT content should be correlated positively with homopolymers of amino acids with generally AT-rich codons and negatively with those that have CG-rich codons. Figure 5B shows that this is true. As a result,
species with similar genomic AT content tend to have similar relative abundances of amino acids in homopolymers, and species with very different genomic AT content have very dissimilar homopolymer profiles (supplementary fig. S4, Supplementary Material online).

Discussion

Genome nucleotide content is a key factor controlling the abundance of SSRs or microsatellites. Simulations confirm the simple expectation that sequences biased toward either AT or GC will generate more small random SSRs by chance. If chance generation of small SSRs is the first step in their evolution, then we might expect to see many more SSRs in biased genomes (fig. 1). Our survey of 25 protist genomes strongly confirms this. The relationship between AT content and SSR abundance (either frequency or density) is always U-shaped, with a minimum for relatively balanced genomes and rising toward both extremes (fig. 2, supplementary fig. S2, Supplementary Material online). The fits to quadratic functions are strong, with the quadratic coefficients being significantly positive (supplementary table S2, Supplementary Material online). Overall, AT content explains about half the variance in SSR frequency and density, both for whole genomes, and for coding and noncoding sequences considered separately (supplementary table S2, Supplementary Material online). It is quite surprising that such a simple explanation applies, given the number of different forces that could influence SSR abundance in different genomes. Similar quadratic fits apply when the data are broken down by type of SSRs with different repeat lengths (1–6 bp), though the fraction of the variance explained is sometimes lower (supplementary table S2, Supplementary Material online).

Our hypothesis assumes a two-phase process of SSR evolution, with random generation of short repeats followed by expansion during replication slippage, a commonly assumed model (Levinson and Gutman 1987; Stephan and Cho 1994; Messier et al. 1996; Rose and Falush 1998; Dieringer and Schlötterer 2003). An alternative model, for which there is some evidence, is that slippage-like processes are involved in the generation of even the first few repeats (Zhu, Queller, et al. 2000; Zhu, Strassmann, et al. 2000). The success of our prediction supports the view that random repeat generation is very important in the early stages, though it does not exclude the possibility that early slippage also contributes. Our results also are of course consistent with the second stage of expansion by slippage because the abundance of our random SSRs is far lower than that of SSRs in real genomes (compare Y axes of fig. 1 with fig. 2 and supplementary fig. S2, Supplementary Material online) as has been previously noted (Dieringer and Schlötterer 2003). So although random generation of short repeats accounts for the relative frequencies of SSRs, the absolute frequencies and lengths require some additional process such as slippage.

It is tempting to consider that the two stages of the process map to our two measures, frequency and density. Random generation of short repeats would seem to largely determine the frequency of repeats, whereas subsequent slippage would have large effects on density. If this were so, we would expect the relationship with AT content to be stronger for frequency than for density. The \( R^2 \) values are indeed slightly higher for frequency but not by much (supplementary table S2, Supplementary Material online). However, the expectation that the frequencies we observe are due solely to random repeat generation, and not slippage, is probably not correct. We counted numbers of
perfect repeats, and these numbers are expected to be influenced just not only by generation of new SSRs but also by breakup by point mutation of older SSRs.

There is one minor lack of fit between our simulations and data. The simulations, as expected, always show a minimum for SSR frequency or density at 50% AT content. Our data, in contrast, have minima at slightly higher AT contents, 50–55% for coding sequences and 55–60% for noncoding sequences. This would seem to suggest that the generation (random) and elongation (slippage) tend to be slightly biased in favor of GC-rich SSRs, and the bias is stronger in noncoding sequences than in coding sequences. This would also seem to suggest that the degeneration trends to be slightly biased in AT-rich SSRs. But we know of no direct evidence for this, and it is unexpected given the stronger hydrogen bonding of cytosine and guanine in GC-rich SSRs. The different bias between coding and noncoding sequences was probably caused by the slightly higher GC content in coding sequences than in noncoding sequences (supplementary table S3, Supplementary Material online). As expected from our hypothesis, the SSRs in AT-rich genomes tend to use AT-rich motifs, whereas those in GC-rich genomes tend to use GC-rich motifs (fig. 3).

Within genomes, AT content appears to influence SSR abundance in the same manner, but with the complication that the range of variation in AT content is smaller, so we do not observe the full U-shaped curve in any species. Instead, AT-rich species have most of their genome on the ascending part of the curve and are expected to show a positive correlation between AT content and SSR abundance. GC-rich species fall on the descending part of the curve and are expected to show a negative correlation between AT content and SSR abundance. Species in the middle of the range should show little or no correlation. These predictions are well supported by the data (fig. 4, supplementary fig. S3, Supplementary Material online).

The influence of AT content extends to amino acid content. Triplet repeats in coding regions generate homopolymer amino acid repeats. As expected, such amino acid repeats are more frequent and denser in AT-biased or GC-biased genomes (fig. 5A). This relationship is somewhat more surprising because the connection is more indirect; amino acid repeats can arise not only from trinucleotide repeats but also from mixtures of synonymous codons.

Moreover, many amino acids can be coded for either by GC-rich or AT-rich codons. But amino acids do differ in average AT content of the triplets that code for them, and this has an effect. Amino acids with high average AT content in their codons tend to show increased repeats with genome AT content, whereas those with low AT content show declines (fig. 5B). As a consequence, species with similar genome AT contents tend to have similar amino acid compositions in their repeats (supplementary fig. S4, Supplementary Material online).

Although AT content explains a large fraction of the variation in SSR frequency and density, other factors must also play a role (Amos and Rubinstein 1996; Amos et al. 1996, 2008; Schlötterer et al. 2006; Varela and Amos 2009). Different slippage rate mutations might also explain some of the variation between species (Schlötterer et al. 2006), though D. discoideum, despite its very high SSR abundance, has a low slippage rate (McConnell et al. 2007). Mutation rates may even vary with factors such as heterozygosity and proximity of other microsatellites (Amos et al. 2008; Varela and Amos 2009). Within genomes, our data conform to the well-known observation that SSRs in eukaryotes are more common in noncoding regions either because of greater selective purging in coding regions (Hancock 1995, 1996a) or more specifically because slippage of (nontriplet) coding SSRs generally leads to deleterious frameshift mutations (Metzgar et al. 2000; Metzgar and Wills 2000). Certain genomic regions are more prone to accumulate SSRs. For example, the SSR densities at the end of chromosome arms are 2-fold higher than other regions in human and mouse genomes (Lander et al. 2001; Waterston et al. 2002). In addition, high percentages of SSRs sometimes occur within larger repeats such as transposons (Arcot et al. 1995; Ramsay et al. 1999; Temnykh et al. 2001). Another interesting phenomenon is that different genomes each have their own characteristic and nonrandom abundances of nonrepeat dinucleotides (Karlin and Burge 1995; Karlin and Mrazek 1997; Campbell et al. 1999). This means our assumption of random formation of early repeats is not strictly correct, but it seems correct enough to generate the SSR abundance patterns we predict. Study of all of these additional factors will be facilitated if one first removes the effect of simple base composition.

We chose to focus on protists because of the simplicity of their genomes—their relatively constant AT content and relative freedom from junk DNA. They provide clear evidence that when nucleotide content is biased either in favor of AT or GC, and more short random repeats are generated by chance, the resulting larger pool of sequences subject to slippage results in more SSRs. Though the larger genomes of multicellular eukaryotes are more complex, it seems likely that the same principle applies, and this is worth further investigation. For species with isochores of strongly differing AT content, the within-species correlations may be enhanced, whereas the between-species correlations may be weakened because each species average AT content would be less representative of local conditions. Greater abundance of transposons may shift abundance of SSRs that are associated with them. However, we predict that much of the variation in SSR abundance within and between these genomes will still be driven by AT content.

Supplementary Material

Supplementary figures S1–S4 and tables S1–S9 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank the two anonymous referees for the constructive comments and suggestions. This work is based in part upon...
work supported by the National Science Foundation under Grant Numbers EF-0626963 and DEB-0918931.  

References  


