

Abundance, Distribution, and Mutation Rates of Homopolymeric Nucleotide Runs in the Genome of *Caenorhabditis elegans*

Dee R. Denver,¹ Krystalynne Morris,² Avinash Kewalramani,¹ Katherine E. Harris,³ Amy Chow,⁴ Suzanne Estes,⁵ Michael Lynch,¹ W. Kelley Thomas²

¹ Department of Biology, Indiana University, 327 Jordan Hall, 1001 East Third Street, Bloomington, IN 47405, USA

² Hubbard Center for Genome Studies, University of New Hampshire, Durham, NH 03824, USA

³ Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA

⁴ Saint Luke's Hospital of Kansas City, Kansas City, MO 64111, USA

⁵ Department of Zoology, Oregon State University, Corvallis, OR 97331, USA

Received: 1 August 2003 / Accepted: 16 December 2003

Abstract. Homopolymeric nucleotide runs, also called mononucleotide microsatellites, are a ubiquitous, dominant, and mutagenic feature of eukaryotic genomes. A clear understanding of the forces that shape patterns of homopolymer evolution, however, is lacking. We provide a focused investigation of the abundance, chromosomal distribution, and mutation spectra of the four strand-specific homopolymer types (A, T, G, C) ≥ 8 bp in the genome of *Caenorhabditis elegans*. A and T homopolymers vastly outnumber G and C HPs, and the run-length distributions of A and T homopolymers differ significantly from G and C homopolymers. A scanning window analysis of homopolymer chromosomal distribution reveals distinct clusters of homopolymer density in autosome arms that are regions of high recombination in *C. elegans*. Dramatic biases are detected among closely spaced homopolymers; for instance, we observe 994 A homopolymers immediately followed by a T homopolymer (5' to 3') and only 8 instances of T homopolymers directly followed by an A homopolymer. Empirical homopolymer mutation assays in a set of *C. elegans* mutation-accumulation lines reveal an ~ 20 -fold higher mutation rate for G

and C homopolymers compared to A and T homopolymers. Nuclear A and T homopolymers are also found to mutate ~ 100 -fold more slowly than mitochondrial A and T homopolymers. This integrative approach yields a total nuclear genome-wide homopolymer mutation rate estimate of ~ 1.6 mutations per genome per generation.

Key words: *Caenorhabditis elegans* — Genome — Homopolymer — Microsatellite — Mutation

Introduction

Simple sequence runs are a major component of every eukaryotic genome that has been surveyed and homopolymeric nucleotide runs (HPs), also called mononucleotide microsatellites, are consistently among the most abundant types of simple sequence (Dechering et al. 1998; Toth, et al. 2000; Katti et al. 2001). HP loci can be highly mutagenic and HP instability is associated with many types of cancer (Catasus et al. 2000; Richetta et al. 2001; Zhang et al. 2001). However, these simple sequences are also thought to have roles in transcriptional regulation and recombination (Brahmachari et al. 1997; Kashi et al. 1997; Templeton et al. 2000). HPs in mismatch repair gene

Novel sequences are deposited in GenBank under accession numbers AY219759–AY219789.

Correspondence to: Dee R. Denver; email: ddenver@bio.indiana.edu

exon sequences have been postulated to act as mutation rate modulators (Chang et al. 2001). Structural in vitro analyses of HPs composed of A:T base pairs show that these mononucleotide runs have a very rigid structure stabilized by additional, non-Watson–Crick, cross-strand hydrogen bonding (Nelson et al. 1987). Alternatively, HP runs containing G:C base pairs form triplex structures and G quartets in vitro (Sen and Gilbert 1988; Cheung et al. 2002). The hypervariable nature of HPs along with their profusion in eukaryotic genomes suggests that these repetitive sequences play a significant role in genome evolution and may serve as good models for understanding general patterns of simple sequence evolution.

Multiple mutational and selective forces contribute to the evolution of simple sequences such as HPs. For instance, the canonical high incidence of length change mutations associated with HPs and other microsatellites has contributed to their exclusion from exon sequences over evolutionary time (Metzgar et al. 2000). Explaining the overall abundance and biased distribution patterns of simple sequences in eukaryotic genomes, however, has proven more difficult. The cryptic simplicity hypothesis maintains that simple sequences are seeded by an active mechanism, most likely slippage during replication (Tautz et al. 1986). This theory also predicts that the distribution of simple sequences in the genome is essentially random and reflects the underlying base composition. Alternative hypotheses maintain that repetitive sequences in the human and *Drosophila* genomes are seeded from retrotransposon insertions, suggesting that simple sequences and retrotransposons are similarly distributed in genomic sequences (Nadir et al. 1996; Wilder and Hollocher 2001). A third model proposes a balance between slippage events and point mutations inside repeats in shaping simple sequence distribution patterns (Kruglyak et al. 1998, 2000). In this “mutational balance” model, new repeat units are generated by the splitting of existing repeat units by internal point mutations, suggesting that repeats are distributed near other repeats of the same motif type (Kruglyak et al. 2000). Finally, some models suggest that unequal crossovers during recombination (Smith 1976) and/or slippage associated with DNA synthesis following gene conversion (Richard and Paques 2000) contributes to the spreading of simple sequences in eukaryotic genomes. None of these theories alone, however, have proven sufficient to explain the observed patterns of simple sequence evolution in eukaryotic genomes (Deschering et al. 1998; Toth et al. 2000).

HPs are among the most dominant forms of simple sequences observed in eukaryotic genomes, and accurate estimates of the rates and patterns of HP mutation are critical for understanding their evolutionary properties. HPs have been characterized as hotspots

for length change mutations (Tran et al. 1997; Denver et al. 2000) and in *Saccharomyces cerevisiae* HPs composed of G:C base pairs have been shown to mutate at higher rates than HPs composed of A:T base pairs (Gragg et al. 2002). Most studies that target simple sequence mutational properties, however, consider only a few representative repeat loci or involve indirect plasmid-based assays. Although the above approaches have significantly contributed to a basic understanding of the mechanisms and factors involved in HP mutation, a large-scale and direct analysis of the rates and patterns of spontaneous mutation at multiple genomic HP loci is essential for an accurate understanding of HP dynamics and their roles in eukaryotic genome evolution.

This study provides an integrative approach to understanding the evolutionary and mutational properties of HPs in the genome of *C. elegans*. First, we assay the abundance and distribution of HPs with respect to chromosome position and physical association with other HP loci. Second, we perform a large-scale, direct screen for mutations at 38 distinct HP loci of different types and sizes in a set of 72 *C. elegans* mutation-accumulation (MA) lines (Vassilieva and Lynch 1999; Vassilieva et al. 2000) as well as 23 natural geographic isolates (Hodgkin and Doniach 1997; Denver et al. 2003). The *C. elegans* MA lines are propagated across generations as single random worms to minimize the efficiency of natural selection against new mutations. All but the most deleterious mutations accumulate over time in the MA lines, allowing for direct and unbiased estimates of baseline mutation rates and patterns (Denver et al. 2000). Together, knowledge of the abundance and distribution patterns of HP loci in the *C. elegans* genome with a direct and unbiased evaluation of HP mutation spectra in the MA lines provides insights into the forces shaping *C. elegans* HP evolution.

Materials and Methods

Detection of Homopolymers in C. elegans Chromosome Sequences

C. elegans chromosome sequences were downloaded from the Sanger Centre ftp site (ftp://ftp.sanger.ac.uk/pub/C.elegans_sequences/chromosomes/current_release) on 12 September 2000 for subsequent analysis; 99, 192, 192 bp was analyzed. HP positions were detected by two independent computational approaches. Our first analysis utilized the Mac Vector 7.0 (Oxford Molecular Group) computer program to find HPs ≥ 8 bp. We selected 8 bp as the lower limit for this bioinformatic analysis, as it was the lower limit for empirical mutational analyses of HP mutation described below. Blocks of chromosome sequence (2 MB) were opened in Mac Vector; HP loci were detected using the subsequence search option (searched for “AAAAAAA,” “TTTTTTT,” “GGGGGGG,” and “CCCCCCC” in four independent searches for each sequence block). MacVector subsequence output for the beginning position of each HP of 8 bp was then analyzed in

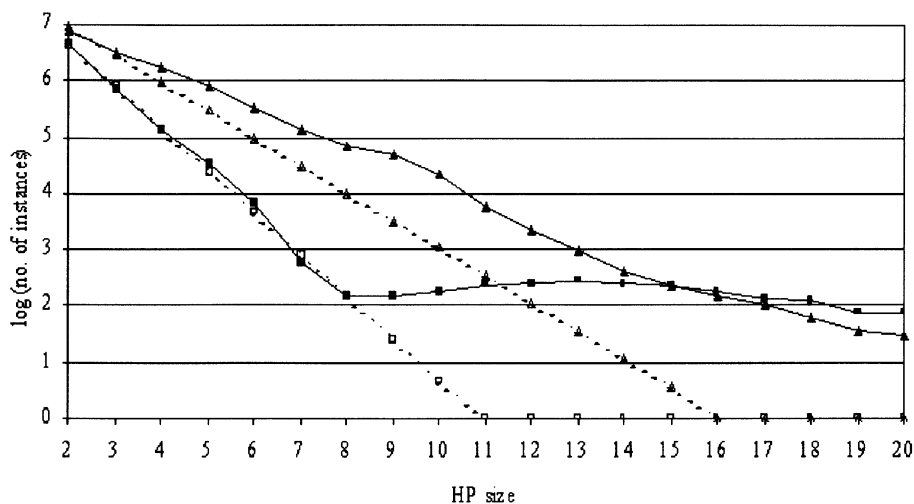


Fig. 1. Size class distributions of HPs in the genome of *C. elegans*. The log-transformed abundances of HPs 2–20 bp are shown. The solid line with filled triangles corresponds to the observed numbers of A and T HPs, whereas the solid line with filled squares represents the observed numbers of G and C HPs. The dashed lines represent the expected numbers for A and T (open triangles) as well as G and C (open squares) HPs, based on individual nucleotide frequencies.

Microsoft Excel to discover the actual size for each homopolymer. For example, a homopolymer of 10 bp was returned in the MacVector query as three 8-bp homopolymers with consecutive beginning positions. After determining the length of the HPs in Microsoft Excel, the data were input into Microsoft Access for archiving and subsequent analyses.

To gain a better understanding of smaller HP size distributions, the positions of all mononucleotide runs ≥ 2 bp were determined in full chromosome sequences (13 March 2001) using a search program in Microsoft Visual C++ 6.0. HP sizes were entered into the program that generates the four possible HPs (i.e., if the size entered is 3, it would generate AAA, GGG, CCC, TTT), which were then stored in an array for use in subsequent text searches. Complete chromosome sequences were stored in a second set of arrays, one for each chromosome sequence. The contents of the chromosome sequence arrays were subjected to text search analyses, searching with the HPs generated in the first array. The program only searched for unique HPs (i.e., the program searched for AAA HPs that have no A preceding or following it). This analysis yielded results highly similar to the subsequent searches performed in MacVector 7.0 for HPs ≥ 8 bp (later sequence release used here contained ≤ 20 additional HPs for each size class compared to previous results).

Determination of Distances Between Adjacent Homopolymers

To evaluate the number of intervening nucleotides between adjacent homopolymers, the beginning position and length of the homopolymers from the Microsoft Excel data set were used to determine the ending position of the homopolymer run. The ending position for each homopolymer was subtracted from the beginning position of the next homopolymer to determine the number of intervening base pairs. In order to obtain a random expectation for this analysis, Microsoft Visual C++ 6.0 was used to generate sets of random numbers for each chromosome. The same number of random numbers was generated for each chromosome, each number between 1 and the chromosome length in base pairs, as the number of HPs ≥ 8 bp that was empirically detected for each respective chromosome. The random numbers for each chromosome were then sequentially arranged in Microsoft Excel—the number of bp between this randomized HP set was then determined. The total number of observations for each class (0 intervening bp, 1 intervening bp, etc.) was then calculated and plotted along with the observed data.

Mutational Analysis of *C. elegans* Homopolymer Loci

Propagation of the *C. elegans* MA lines was described by Vassilieva and Lynch (1999); the *C. elegans* natural isolates were described by Doniach and Hodgkin (1997) and Denver et al. (2003). HP loci (8 to 16 bp in length) were randomly selected for amplification and sequencing from MA lines and natural isolates to detect size changes. Amplifications of DNA were performed in 50- μ L reactions with Taq polymerase (Applied Biosystems). Amplifications were carried out by 35 cycles of denaturation at 95°C for 1 min, annealing at 55 to 65°C for 1 min, and extension at 72°C for 2 min. Primers were designed to *C. elegans* sequences flanking HPs to yield PCR products ~ 300 to 1000 bp that contain targeted HP sequences. PCR reaction aliquots (3.5 μ L) were evaluated on ethidium bromide-stained agarose gels and subsequently purified by solid-phase reversible immobilization (Elkin et al. 2001). Purified PCR products were then cycle sequenced by 25 cycles of denaturation at 96°C for 30 s, annealing at 50°C for 15 s, and extension at 60°C for 4 min, 0.5% DMSO was included in sequencing reactions to improve sequence quality. After ethanol precipitation, sequences were determined with an ABI Prism 377 automated DNA sequencer (Applied Biosystems). HP sequences were analyzed on both strands by manually counting dye peaks in the electropherograms, using the control progenitor (N2) template as a standard. Sequences were submitted to GenBank under accession numbers AY219759–AY219789. HP mutation rates were determined as in Denver et al. (2002). Mutation rates were calculated using the equation $\mu = n/(t \times 1)$, where μ = the mutation rate, n = the number of mutations, t = the number of generations, and 1 = the number of MA lines. For HP loci of run lengths outside of those assayed here (<8 bp, >16 bp), regression analyses were performed in Microsoft Excel to extrapolate mutation rates.

Results

Homopolymer Size Class Distributions

We initially evaluated the abundances of different HP size classes, analyzed for all runs ≥ 2 bp in the *C. elegans* genome (Fig. 1). All four of the strand-specific HP types (A, T, G, and C) were considered distinct. The size class abundances of A and T HPs in *C. elegans* were virtually identical, as were those for G and C HPs. The expected number of A and T HPs

Table 1. Abundance of homopolymers (≥ 8 bp) in the genome of *C. elegans*

HP	Chr. I	Chr. II	Chr. III	Chr. IV	Chr. V	Chr. X	Genome
A	11,452	11,134	11,339	13,679	13,776	11,673	73,053
T	11,776	11,212	11,160	13,861	13,647	11,515	73,171
G	178	144	139	170	164	340	1,135
C	215	191	171	191	152	346	1,266
Total	23,621	22,681	22,809	27,901	27,739	23,874	148,625
HP/MB	1,607	1,492	1,715	1,604	1,308	1,372	1,498

Note. HPs were counted on the (+) strand of *C. elegans* chromosome sequences obtained from the Sanger Centre on 12 September 2000 (ftp://ftp.sanger.ac.uk/pub/C.elegans_sequences/chromosomes/current_release/).

based on individual nucleotide frequencies predicted a steady, linear reduction in A and T HP abundance with increasing run length and that no HPs ≥ 16 bp in length would be present in the *C. elegans* genome. Although a roughly steady reduction in A and T HP abundance with increasing run length was observed, the number of A and T HPs detected for each size class was greater than the random expectation (Fig. 1). Furthermore, many instances of A and T HPs > 16 bp in length were observed; the largest was a 35-bp T HP at position 15,708,621 on the X chromosome.

G and C HP size class abundances were distinct from those of A and T HPs (Fig. 1). As with A and T HPs, a linear reduction in G and C HP abundance with increasing run length was expected based on individual nucleotide frequencies, though far fewer G and C HPs for each size class were predicted due to the AT-rich genome of *C. elegans*. Unlike A and T HPs, the observed numbers of G and C runs ≤ 8 bp roughly followed the expectations based on mononucleotide frequencies. The observed numbers of G and C HPs ≥ 8 bp, however, were much greater than the random expectations. Unlike A and T HPs, G and C HPs did not display a steady decline in abundance with increasing run length. In fact, the number of G and C HPs 13 bp in length was greater than each of the 8- to 12-bp G and C HP size classes (Fig. 1). Although none were expected at random, many G and C HPs ≥ 11 bp in length were detected; the largest was a 32-bp C HP at position 1,980,365 on chromosome III.

Genomic Distribution of Homopolymers

We characterized the genomic distribution patterns of the 148,625 unique HPs ≥ 8 bp long detected on the (+) strands of *C. elegans* chromosomes (Table 1). The number of A and T HPs ≥ 8 bp long (146,224) vastly exceeded the number of G and C HPs ≥ 8 bp long (2401). The ratio of A and T HPs to G and C HPs was $\sim 61:1$, much higher than the ratio of A and T-to-G and C nucleotides ($\sim 2:1$) in the *C. elegans* genome. The observed number of A and T HPs ≥ 8 bp long

was more than five times greater than the expected number based on individual nucleotide frequencies (21,812 A and T HPs expected). The observed number of G and C HPs ≥ 8 bp was ~ 17 -fold greater than expectations based on individual nucleotide frequencies (138 G and C HPs ≥ 8 bp expected). The numbers of A and T HPs were nearly equal on the (+) strands of all chromosomes, as were the numbers of G and C HPs (Table 1). This equality of complementary HP motif types on both strands of all chromosomes indicated that no strong strand-specific biases exist for HP distribution in the *C. elegans* genome.

An analysis of HP abundance in each *C. elegans* chromosome revealed significant variation in HP density ($p < 0.005$), from an average of 1,308/MB for the largest chromosome (V) to 1,714/MB in the smallest chromosome (III) (Table 1). HPs were most dense in chromosome arms and less abundant in the core regions (arm and core regions defined as by Barnes et al. [1995]). Furthermore, the frequency of any HP type (A, T, G, or C on the [+] strand) in each MB was highly correlated with all other types of HPs. This was particularly striking for the correlation of A with T HPs and G with C HPs. These strong correlations ($r = 0.96$ for A with T, 0.76 for G with C, and 0.96 for A and T with G and C) were consistent across all chromosomes. Therefore, regions of chromosomes with high HP density had increased abundances of all four HP motif types.

To examine the distribution of HPs across *C. elegans* chromosomes, we employed a scanning window approach. For each chromosome, the number of HPs in a 100-kb window of sequence was determined, beginning at position 1 of each chromosome. The window was then moved one nucleotide 3' on the (+) strand across the length of the chromosome, counting HPs in each window. The data were averaged over 1000 consecutive windows and displayed graphically by chromosome in Fig. 2. This analysis revealed the presence of many HP "clusters" that routinely exceeded 300 HP/100 kb. These clusters were strongly overrepresented in autosome arms, while the core regions of chromosomes generally displayed a baseline HP density of approximately 90 to 100 HP/

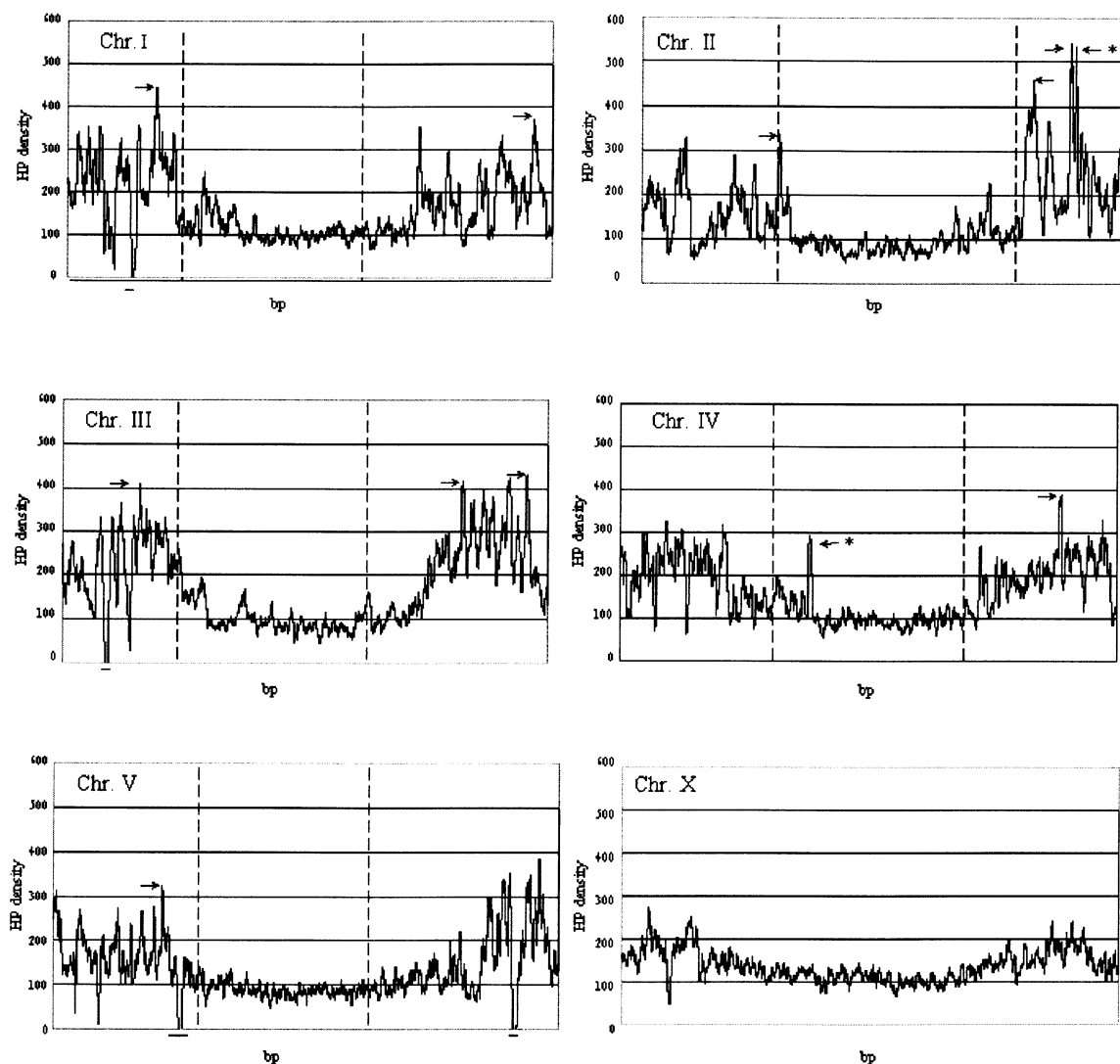


Fig. 2. Distribution of HPs across *C. elegans* chromosomes. HPs are counted across a 100-kb window and then averaged over 1000 windows. HP density (*y* axes) is then plotted against chromosome physical maps (*x* axes). Patterns in HP densities are resolved on a very fine scale with this approach, allowing for the definition of discrete HP clusters. Arrows indicate the densest HP clusters that

are described in the Results. Arrows with asterisks indicate the two clusters where HP motif-type biases are observed. The dashed vertical lines in the top panels indicate the conventional arm and core boundaries as defined by Barnes et al. (1995). Four small horizontal lines below the *x* axis indicate gaps > 5000 bp in the *C. elegans* genome sequence analyzed.

100 kb. In fact, overall we observed a significant ($p < 0.005$) overabundance of HPs in arm regions of autosomes, under the null hypothesis that HPs are randomly distributed across autosome sequence (60,480 and 41,462 HPs expected in autosomal core and arm regions, respectively; 33,521 and 68,421 HPs observed in core and arm regions, respectively). In contrast, the X chromosome lacked a clear demarcation between core and arms and was largely devoid of high-density clusters (all < 300 HP/100 kb).

Twelve of the densest HP clusters were further characterized by analyzing the specific 100-kb windows of DNA sequence that corresponded to these clusters (indicated by arrows in Fig. 2). The gene densities and base compositions of these regions of high HP density were similar to genomic averages. A and T HPs were dominant in these clusters as they

were in the rest of the genome. The distribution of A and T HPs was evaluated across 10-kb segments in each of the 12 clusters—the majority of these clusters (10/12) contained roughly equal numbers of A and T HPs across all segments. However, two clusters (indicated by asterisks in Fig. 2) displayed skewed HP motif type biases. The first cluster (on chromosome II) displayed a 3.6-fold enrichment of A over T HPs in one 10-kb segment followed by a 2.8-fold enrichment over A HPs just 10 kb downstream. Visual examination of the 10 kb of genomic sequence that corresponded to these segments of the cluster showed that the biased distribution of HP types in these regions was not due to the presence highly dense and localized A or T HP “superclusters.” Rather, the dozens of HPs in each 10-kb segment were dispersed and separated by complex sequences. The second

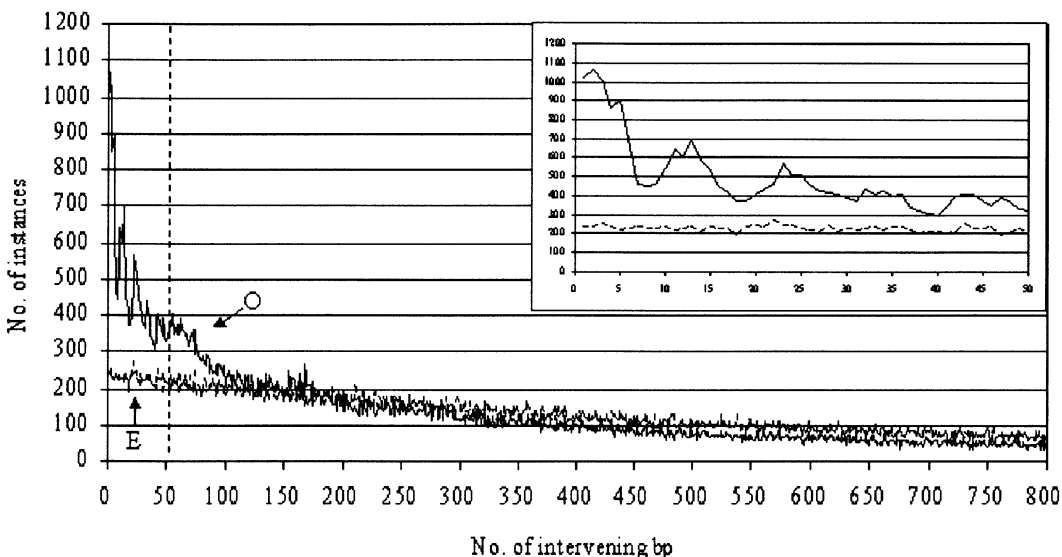


Fig. 3. Frequency distribution of bases between HPs in the *C. elegans* genome. The thick black line indicated by O shows observed values; the thin black line indicated by E shows expected values based on a random distribution of HPs across chromosomes (see Materials and Methods). **Inset:** A more detailed view of the

classes of HP pairs with 0 to 50 intervening bases (expected values are indicated by the dashed line). A clear bias toward small numbers of bases between HPs is evident. Additional biases are detected, such as the abundance of ~10–12 and ~22–24 bp between HPs.

cluster displaying a skewed HP motif-type bias (on chromosome IV) was the only cluster detected in an autosomal core region. An extreme bias toward A over T HPs was observed throughout this cluster. Further examination of this region of chromosome IV revealed the presence of three separate large tandem repeats that dominated this region of the genome. Two of these repeats contained A HPs as a subcomponent of the larger repeating unit, resulting in the observed HP motif-type bias in this region of chromosome IV. Although HP motif-type distribution biases were observed in a few specific instances, most of the clusters examined were heterogeneous with respect to HP motif type.

To further examine the propensity for HP clustering, we determined the number of intervening base pairs between all HP loci ≥ 8 bp in length. A random distribution of HPs across chromosomes (see Materials and Methods) predicted a bias toward closely spaced HPs and ~ 250 instances of directly adjacent HPs (Fig. 3). However, we observed a strongly skewed HP spacing pattern distribution displaying a much stronger bias for closely spaced HPs compared to the random expectation, particularly for HPs separated by < 50 bp (Fig. 3; see inset). To further investigate very closely associated HPs, we identified all HP pairs separated by 0 to 5 intervening base pairs (Table 2). The most striking observations involved $(A)_{\geq 8}(N)_{0-5}(T)_{\geq 8}$ and $(T)_{\geq 8}(N)_{0-5}(A)_{\geq 8}$ HP pairs (where N = G or C). We observed 1019 total cases of immediately adjacent HPs (0 intervening base pairs). The overwhelming majority of these (994) were $(A)_{\geq 8}(T)_{\geq 8}$ pairs, whereas only 8 $(T)_{\geq 8}(A)_{\geq 8}$ pairs were detected (Table 2). A roughly equal number of

$(A)_{\geq 8}(T)_{\geq 8}$ and $(T)_{\geq 8}(A)_{\geq 8}$ HP pairs was expected based on the overall equivalent abundances of A and T HPs throughout the genome. The overabundance of $(A)_{\geq 8}(T)_{\geq 8}$ pairs was not observed when intervening nucleotides were present between the HPs (Table 2). In fact, the most abundant class of HP pairs with one intervening nucleotide was $(T)_{\geq 8}(N)_1(A)_{\geq 8}$ pairs that outnumbered $(A)_{\geq 8}(N)_1(T)_{\geq 8}$ pairs nearly 3:1. In contrast, roughly equal numbers of $(T)_{\geq 8}(N)_{1-5}(T)_{\geq 8}$ and $(A)_{\geq 8}(N)_{1-5}(A)_{\geq 8}$ HP pairs were observed for each class.

Mutation and Natural Variation of C. elegans Homopolymers

In addition to surveying the abundance and distribution of HPs across the *C. elegans* genome, we directly assayed the rates and patterns of mutation at 38 genomic HP loci that ranged in length from 8 to 16 bp (Table 3). The average HP run length of loci assayed was roughly equivalent for the G and C (12.4-bp) and the A and T (11.6-bp) HPs. To obtain an unbiased view of HP mutation spectra, we probed for mutations in a set of 72 *C. elegans* MA lines, propagated for 280 generations through single-progeny descent (Vassilieva and Lynch 1999; Vassilieva et al. 2000). Natural patterns of HP variation were also evaluated by analyzing the same set of HP loci in a set of 23 *C. elegans* natural geographic isolates (Hodgkin and Doniach 1997; Denver et al. 2003).

We detected 31 total mutations at HP loci in 72 *C. elegans* MA lines; 30 were length change mutations and 1 was a base substitution (Table 3). No significant differences were observed between the frequency

Table 2. Spacing between homopolymers

HP pair ^a	No. of intervening nucleotides					
	0	1	2	3	4	5
A–A	N/A	283	294	239	274	212
A–T	994	143	180	144	116	99
A–G	5	2	2	0	1	0
A–C	0	0	0	1	1	1
T–A	8	369	273	195	258	119
T–T	N/A	257	246	274	242	222
T–G	6	1	0	0	1	0
T–C	0	0	1	0	0	1
G–A	0	0	0	1	1	0
G–T	2	0	0	0	0	0
G–G	N/A	5	1	0	0	0
G–C	0	0	0	0	0	0
C–A	3	1	1	0	4	1
C–T	1	1	2	0	2	1
C–G	0	0	0	0	0	0
C–C	N/A	2	1	1	1	0
Total	1019	1064	1001	855	901	656

^a HP pair indicates the 5'-to-3' order of strand-specific HP types.

of insertions (14) and deletions (16), and virtually all (29/31) of the detected mutations were in G and C HPs. No significant difference in mutation rate was detected between HPs with G on the (+) strand (12 mutations in 8 G HPs) and those with C on the (+) strand (17 mutations in 8 C HPs). G and C HPs also displayed a trend of increasing mutation rate with increasing run length, a property commonly observed in other simple sequences (Kroutil et al. 1996; Wierdl et al. 1997). In contrast to the G and C HPs, we detected only two mutations in the A and T HP loci assayed in the MA lines. Furthermore, these two mutations were of different types: a T → C transition was detected at the first (5') base of the (T)₈ HP assayed in cosmid Y75B8A and a single-base deletion was detected in the (T)₁₃ HP in cosmid H02F09 (Table 3).

We observed 29 total G and C HP mutations in the MA lines, leading to an average mutation rate of $9.0 \times 10^{-5}/\text{gen}$ ($\pm 1.7 \times 10^{-5}$) for G and C HPs (see Materials and Methods for mutation rate calculation). However, this rate was dominated by mutations in larger HPs. Smaller G and C HPs (8–12 bp) mutated at a relatively slower rate of $2.2 \times 10^{-5}/\text{genome (gen)}$ ($\pm 1.2 \times 10^{-5}$), whereas larger G and C HPs (13–16 bp) mutated at a high rate of $1.5 \times 10^{-4}/\text{gen}$ ($\pm 2.8 \times 10^{-5}$). For A and T HPs, a mutation rate of $4.5 \times 10^{-6}/\text{gen}$ ($\pm 3.2 \times 10^{-6}$) was calculated using the two mutations observed in the MA lines, however, there was a large variance for this estimate. This rate was also derived from two mutations that presumably arose from disparate mechanisms. Despite the uncertainty of a precise A and T HP mutation rate, the drastic difference in the total number of mutations observed between the G and C and the A and T HPs clearly indicated a

higher baseline rate of mutation for G and C HPs compared to A and T HPs.

We also assayed HP variation patterns in 23 *C. elegans* natural geographic isolates (Hodgkin and Doniach 1997; Denver et al. 2003). Analysis of allelic variation at HP loci in natural isolates revealed a significantly ($p < 0.005$) greater number of alleles in G and C HPs (4.1 alleles/locus) than in A and T HPs (1.5 alleles/locus). This reduction in the allelic variation of A and T HPs compared to G and C HPs is consistent with the overall lower mutation rate observed in the MA lines. In addition to variation in length, two base substitution polymorphisms were detected in natural isolate HPs. The first was observed in seven natural isolates (AB4, CB4852, CB4855, CB4857, CB4858, PB303, and KR314) and was in the (C)₁₁ HP in cosmid C56E6 (Table 3). The second occurred in a different set of seven natural isolates (CB4853, CB4854, CB4855, CB4856, CB4857, KR314, RW7000) and was detected in the (C)₁₆ HP in cosmid Y75B8A. Both of these polymorphisms were detected at the first base of the HP on the (+) strand and were T → C transitions.

Discussion

Forces Shaping Homopolymer Genomic Distribution Patterns

The evolutionary forces that contribute to shaping genomic distribution patterns of repetitive sequences such as HPs have received much attention, as simple sequence loci are frequently used as genetic markers, associated with human genetic diseases, and often constitute a large fraction of the total sequence in eukaryotic genomes (Dechering et al. 1998; Catusus et al. 2000; Toth et al. 2000; Katti et al. 2001; Richetta et al. 2001; Zhang et al. 2001). Some patterns of HP distribution are clearly associated with specific selective forces (such as the exclusion of HPs from exon sequences [Metzgar et al. 2000]), yet the causal agents of the overall abundance and genome-wide distribution patterns of HPs in eukaryotic genomes remain mysterious. Our observations provide insights into the complex and interacting forces that shape HP evolution in the *C. elegans* genome. HP loci (148,625 detected, Table 1) are far more abundant than other classes of simple sequence in the *C. elegans* genome; for instance, only 953 di-, tri-, tetra-, and pentanucleotide microsatellites are present in the genome (Frisse 1999). However, *C. elegans* HPs display the same distributional bias toward autosomal arms that is observed with other classes of simple sequence. Hence, studying the distribution patterns of the abundant HPs may provide general insights into the forces shaping simple sequence evolution in the *C. elegans* genome.

Table 3. Homopolymer mutations (mut) in the MA lines and variation in natural isolates

Chr.	Chr. position	Cosmid/YAC	HP	MA lines			Natural isolates, no. alleles
				+1 mut	-1 mut	Others	
I	742,019	Y18H1A	(G) ₁₄	0	0	0	3
I	6,667,528	C48B6	(A) ₁₂	0	0	0	1
I	8,150,793	F02E9	(T) ₁₄	0	0	0	1
I	8,469,728	C36B1	(A) ₁₄	0	0	0	2
I	8,675,640	F53B6	(C) ₁₀	1	0	0	3
I	13,206,279	Y87G2A	(G) ₁₅	2	2	0	4
I	13,363,712	Y71A12B	(A) ₁₃	0	0	0	1
I	14,591,757	F11C3	(C) ₁₅	2	5	0	3
II	5,413	C01B12	(G) ₁₃	2	1	0	5
II	3,144,663	Y25C1A	(A) ₁₅	0	0	0	3
II	5,219,911	C34F11	(G) ₈	0	0	0	4
II	6,554,125	C56E6	(C) ₁₁	0	1	0	7
II	6,623,525	C56C10	(T) ₁₃	0	0	0	1
II	7,281,543	F32A5	(A) ₁₃	0	0	0	2
II	7,623,410	F35D2	(C) ₁₀	0	0	0	3
II	11,210,974	F37H8	(A) ₈	0	0	0	1
II	11,211,306	F37H8	(A) ₈	0	0	0	1
II	12,018,574	Y17G7B	(A) ₁₅	0	0	0	2
II	2,189,591	Y48G9A	(C) ₁₆	3	2	-2C	5
III	4,373,088	B0284	(T) ₁₀	0	0	0	1
III	12,064,922	Y75B8A	(T) ₈	0	0	T → C	1
III	12,065,026	Y75B8A	(C) ₁₆	1	1	0	5
IV	6,298,927	Y73B6BL	(C) ₉	0	0	0	3
IV	7,813,086	C33H5	(C) ₈	0	0	0	2
IV	8,058,581	C26B2	(A) ₁₆	0	0	0	3
IV	13,267,578	JC8	(T) ₈	0	0	0	1
V	381,383	ZK6	(G) ₁₂	1	0	0	3
V	12,101,060	R13H4	(G) ₁₃	1	1	0	4
V	18,026,664	C47A10	(T) ₈	0	0	0	1
V	18,026,694	C47A10	(T) ₉	0	0	0	1
V	20,252,790	Y113G7A	(G) ₁₄	1	1	0	8
X	296,663	Y47C4A	(A) ₁₄	0	0	0	2
X	419,582	ZK1193	(G) ₁₄	0	0	0	3
X	1,590,065	H02F09	(A) ₁₀	0	0	0	1
X	1,590,212	H02F09	(T) ₁₃	0	1	0	2
X	9,997,061	F19C6	(A) ₉	0	0	0	2
X	16,370,053	C11G6	(A) ₉	0	0	0	1
X	16,370,065	C11G6	(A) ₉	0	0	0	2
Total			28	14	15	2	98

Our observations are inconsistent with many current models of simple sequence evolution. The slip-page-mediated spontaneous generation of simple sequences (cryptic simplicity) cannot be ruled out as a source of de novo HP units by our observations. However, the nonrandom HP distribution patterns, such as the observed bias toward closely spaced HPs (Fig. 3), and the apparent lack of relationship between HP type abundances and the underlying base composition in *C. elegans* (see Results) are inconsistent with an exclusive role for cryptic simplicity in shaping *C. elegans* HP evolution. Seeding of HPs by retrotransposons seems highly unlikely, as *C. elegans* retrotransposons and HPs display different chromosomal distribution patterns: HP clusters are observed almost exclusively in autosomal arms (Fig. 2), whereas retrotransposons display no bias toward arm or core re-

gions (Duret et al. 2000). Furthermore, present levels of active retrotransposons are not sufficiently high in the *C. elegans* genome to account for the profusion of HPs. Unless retrotransposons were extraordinarily abundant and active earlier in the evolutionary history of *C. elegans*, the seeding model fails to explain the distinctive patterns of *C. elegans* HP distribution. The mutational balance model (Kruglyak et al. 1998, 2000) is also inconsistent with our observations. The three base substitution changes observed in the *C. elegans* MA lines and natural isolates are all transitions at the first base in the HP—no internal base substitutions are observed in natural or MA-line HPs (Table 3). This substitution mutational polarity within the repeat unit has also been detected in other simple sequences such as microsatellites (Brohede and Ellegren 1999). The observed heterogeneous motif types of closely spaced

HPs are also inconsistent with the mutational balance model (Table 2).

The striking distributional bias of HP clusters toward autosomal arm regions (Fig. 2), a property shared by most *C. elegans* simple sequences classes (The *C. elegans* Sequencing Consortium 1998), suggests that recombination is associated with the generation of HP clusters, as a hallmark of *C. elegans* autosomal arms is their high incidence of recombination (Barnes et al. 1995). Unequal crossing-over between HP loci is one mechanism for the generation of new HP units that is consistent with some of the distribution patterns of *C. elegans* HPs. In this model, each of the four recombinant molecules resulting from unequal meiotic crossovers harbors either a deletion or a tandem duplication of the points of crossover (HPs in this case) and the genomic region lying between the crossover points in the parent molecule (Smith 1976). Similarly, slippage during gene conversion-associated DNA synthesis may lead to the addition or subtraction of entire HP repeat units (Richard and Paques 2000). Other forces such as slip-strand mispairing of HPs during replication and intramolecular recombination between HPs also likely contribute to the spreading of HP loci in eukaryotic genomes (Gordenin and Resnick 1998). Although the distinctive clustering patterns of HP loci are consistent with a significant role for recombinational forces in shaping HP evolution in *C. elegans*, some of the more striking biases in HP genomic distribution patterns, such as the extreme biases observed among directly adjacent HPs (Table 2), are not accounted for by recombination-related models.

Distributional Biases Among Adjacent Homopolymers

One of the most remarkable observations from this analysis is the dramatic dominance of directly adjacent $(A)_{\geq 8}(T)_{\geq 8}$ HP pairs (994 observed) over $(T)_{\geq 8}(A)_{\geq 8}$ pairs (8 observed) (Table 2). A converse bias is observed among HP pairs separated by one nucleotide: 143 $(A)_{\geq 8}N_1(T)_{\geq 8}$ HP pairs and 369 $(T)_{\geq 8}N_1(A)_{\geq 8}$ pairs observed. In both cases roughly equal numbers are expected based on the highly similar overall abundances and genomic distributions of A and T HP loci. AT dinucleotides have higher relative abundances ($\rho^* = 0.86$) compared to TA dinucleotides ($\rho^* = 0.62$) (Gentles and Karlin 2001), however, this dinucleotide bias is not nearly strong enough to account for the extreme dominance of $(A)_{\geq 8}(T)_{\geq 8}$ HP pairs over $(T)_{\geq 8}(A)_{\geq 8}$ pairs. All of the previously considered models of simple sequence evolution fail to account for these extreme biases.

Natural selection may be hypothesized to play a role in shaping these biases—this would require, for example, extreme negative selection against directly adjacent $(T)_{\geq 8}(A)_{\geq 8}$ pairs and positive selection for

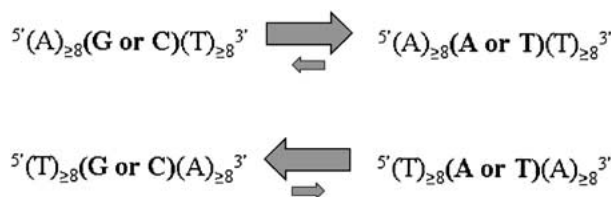


Fig. 4. Proposed substitution mutational bias for closely spaced homopolymers. The directional transition mutation biases presented are proposed to explain the observed distributional biases among HP pairs separated by zero or one intervening base pair (Table 2).

directly adjacent $(A)_{\geq 8}(T)_{\geq 8}$ pairs, as ~ 125 instances of each pair are expected based on base composition (see Results). The specific selective forces required to explain these biases, however, are not clear. The unique structure of directly adjacent A and T mononucleotide runs has been implicated in transcriptional regulation in yeast (Iyer and Struhl 1995), however, the distribution of directly adjacent $(A)_{\geq 8}(T)_{\geq 8}$ pairs in different functional coding sequences is indistinguishable from the overall distribution of all HPs in *C. elegans* (data not shown). The structural transition from a run of A nucleotides to a run of T nucleotides ($5'$ to $3'$) has been shown to result in a distinctive kink that is not observed in transitions from runs of T to A nucleotides (Coll et al. 1987). It is possible that there is selection for the specific secondary structure associated with directly adjacent $(A)_{\geq 8}(T)_{\geq 8}$ pairs and/or selection against structures associated with directly adjacent $(A)_{\geq 8}(T)_{\geq 8}$ pairs.

Rather than selective forces, mutational biases may account for the observed disparity in the distribution patterns of closely spaced HPs. One possibility involves length change mutation biases. The N_1 nucleotide in $(A)_{\geq 8}N_1(T)_{\geq 8}$ pairs may be especially subject to slippage-mediated deletion, and conversely, directly adjacent $(T)_{\geq 8}(A)_{\geq 8}$ HP pairs may be prone to a unique slippage-mediated insertion bias toward G:C base pairs between the HPs. Although this model cannot be ruled out with the present data, virtually all slippage-mediated length change mutations of HPs and other forms of repetitive DNA alter only the number of units in the repeat (Tran et al. 1997; Weirld et al. 1997; Levy and Cebula 2001). The insertion/deletion of base pairs that are of differing types than the surrounding repeat unit is fundamentally inconsistent with the basic slip-strand mispairing mechanism associated with slippage-mediated mutation of repetitive DNA.

A base substitution mutational bias may better explain the skewed distribution patterns of closely spaced HPs in *C. elegans* (Fig. 4). A mutational bias strongly favoring G:C \rightarrow A:T over A:T \rightarrow G:C substitutions for the N_1 nucleotide in $(A)_{\geq 8}N_1(T)_{\geq 8}$ HP pairs would contribute to both the profusion of directly adjacent $(A)_{\geq 8}(T)_{\geq 8}$ HP pairs and the relative

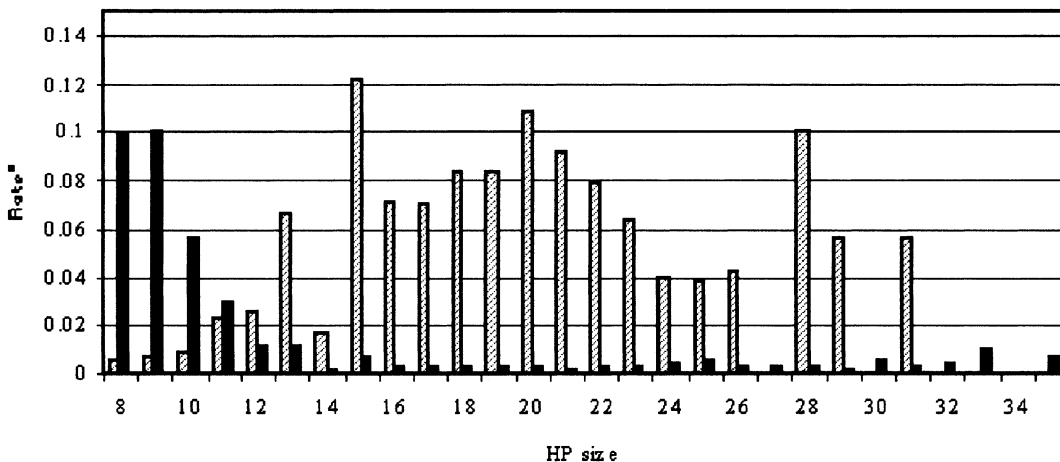


Fig. 5. Genome-wide HP mutation rates with respect to motif type and run length. Black bars indicate A and T HPs and hatched bars indicate G and C HPs. Rate indicates diploid genomic mutation rate (mutations per genome per generation) on the y axis.

Genomic mutation rate is plotted against HP run length on the x axis. Genome-wide HP mutations are dominated by small A and T HPs and large G and C HPs.

paucity of $(A)_{\geq 8} N_1(T)_{\geq 8}$ HP pairs. A converse mutational bias ($A:T \rightarrow G:C$ favored over $G:C \rightarrow A:T$ substitutions at the N_1 nucleotide) among pairs where the $(T)_{\geq 8}$ HP is 5' to the $(A)_{\geq 8}$ HP would explain the nearly complete absence of directly adjacent $(T)_{\geq 8}(A)_{\geq 8}$ HP pairs and the relative dominance of $(T)_{\geq 8}N_1(A)_{\geq 8}$ pairs over $(A)_{\geq 8}N_1(T)_{\geq 8}$ pairs. The differing structural properties of directly adjacent $(A)_{\geq 8}(T)_{\geq 8}$ and $(T)_{\geq 8}(A)_{\geq 8}$ HP pairs suggest that nucleotides at the point of transition may be exposed to distinctive molecular environments and, consequently, differing mutagenic pressures (Coll et al. 1987). Further support is provided by a general substitution mutation bias toward transitions at the first nucleotide position in HP repeats, observed both here among *C. elegans* HPs and in other microsatellites (Brohede and Ellegren 1999). Empirical analyses of mutation patterns among different classes of closely spaced HP loci, however, are required to test for the presence of these specific transition mutational biases (Fig. 4).

Rates and Patterns of Homopolymer Mutation in the *C. elegans* Genome

In addition to comprehensively examining the abundance and distribution of HPs in the *C. elegans* genome, we obtain direct and unbiased estimates of the rates and spectra of HP mutation in the *C. elegans* MA lines, propagated for 280 generations in the relative absence of natural selection (Vassilieva et al. 2000). G and C HPs mutate at much higher rates than A and T HPs, consistent with previous observations in yeast (Gragg et al. 2002). G and C HP alleles in the natural isolates are more polymorphic than A and T HPs, also consistent with an overall higher mutation rate for G and C HPs compared to A and T HPs. The pattern of mutation at G and C HP loci in the MA lines is dominated by single-base length changes

(Table 3). Alternatively, the two mutations observed at A and T HPs in the MA lines consist of one single base deletion and one base substitution (Table 3). The low nuclear A and T HP mutation rate estimate ($4.5 \times 10^{-6}/\text{gen.}$) observed here is surprising, as an $(A)_{11}$ HP in the *C. elegans* mitochondrial genome mutated at a drastically higher rate ($3.2 \times 10^{-4}/\text{gen}$ [Denver et al. 2000]) in the same set of MA lines. Differences in the structural properties and/or replication of HPs between the two subcellular locations may explain the differences in observed mutation rates. Alternatively, differential DNA repair activities (such as mismatch repair) between the nuclear and mitochondrial genomes may contribute to the mutation rate disparity.

The genome-wide rate of HP mutation for *C. elegans* can be estimated with the observations presented in this study. A total of 29 mutations is detected among the 16 G and C HP loci (8 to 16 bp in length) assayed in the MA lines (Table 3). The mutation rate increases with increasing run length for G and C HPs, as is observed in other simple sequence classes such as microsatellites (Wierdl et al. 1997). If we extrapolate the direct mutation rate estimates for G and C HPs 10 to 16 bp in length (no mutations detected in G and C HPs 8–9 bp in length) to all G and C HPs ≥ 8 bp and assume an exponential relationship between HP run length and mutation rate (Tran et al. 1997), a total diploid rate of 1.2 mutations per genome per generation is predicted for G and C HPs ≥ 8 bp (Fig. 5). If we assume that A and T HPs mutate at a rate 19.9-fold reduced relative to G and C HPs (see Results), then a total mutation rate of 0.4 mutation per diploid genome per generation is predicted for A and T HPs ≥ 8 bp (Fig. 5). Together, these two genome-wide mutation rate estimates yield an overall diploid mutation rate of 1.6 mutations per genome per generation for all HPs ≥ 8 bp in the genome of *C. elegans*.

Genomic A and T HP mutations are dominated by the highly abundant but minimally mutagenic runs 8–10 bp in length. Alternatively, genomic G and C HP mutations largely involve the relatively rare but highly mutagenic runs ≥ 15 bp (Fig. 5).

The coupled abundance and high mutation rates of HP loci in the *C. elegans* genome may have significant functional consequences. An analysis of HP distribution patterns in different classes of coding sequence (data not shown) revealed the presence of 3162 HP loci in exon sequence, the majority of which were short (8–10 bp) A and T HPs. By comparison, fewer than 1000 total di-, tri-, tetra-, and pentanucleotide microsatellites are present in the *C. elegans* genome, and only 22 of these are found in exon sequence (Frisse 1999). Given that $\sim 2\%$ of the total HPs in *C. elegans* are in protein-coding sequence, ~ 0.03 HP mutation is predicted to occur in exon sequence in the *C. elegans* genome per generation (or approximately 1 HP mutation in exon sequence per genome every 33 generations). The HP mutation spectrum is dominated by single-base length changes; therefore the majority of HP mutations in exon sequence are likely to result in frameshift mutations.

This study provides an integrative approach to understanding the mutational and evolutionary properties of HPs, a dominant class of simple sequence in all eukaryotic genomes surveyed. We observe distinct HP clusters in autosomal arms where other classes of simple sequence are also overrepresented in the *C. elegans* genome. Our observations suggest that recombination-associated forces likely play a major role in shaping the distinctive distributional patterns of HPs and perhaps other classes of simple sequence in the *C. elegans* genome. The dramatic overrepresentation of (A) $_{\geq 8}$ (T) $_{\geq 8}$ and underrepresentation of (T) $_{\geq 8}$ (A) $_{\geq 8}$ HP pairs in the *C. elegans* genome may reflect a special base substitution mutational bias associated with closely spaced HPs. *C. elegans* G and C HPs mutate at a ~ 20 -fold faster rate than A and T HPs; A and T HPs mutate ~ 100 -fold more slowly in the nucleus compared to mitochondria. The high genome-wide rate of HP mutation detected here (1.6 HP mutations per diploid genome per generation) suggests that this group of dominant simple sequences plays a significant role in *C. elegans* genome evolution.

Acknowledgments. Funding for this work was provided by NIH R01 GM-36827 and the University of Missouri Research Board. We thank two anonymous reviewers for helpful comments.

References

- Brahmachari SK, Sarkar PS, Raghavan S, Narayan M, Maiti AK (1997) Polypurine/polypyrimidine sequences as cis-acting transcriptional regulators. *Gene* 190:17–26
- Brohede J, Ellegren H (1999) Microsatellite evolution: Polarity of substitutions within repeats and neutrality of flanking sequences. *Proc R Soc Lond B Biol Sci* 266:825–833
- Catusas L, Matias-Guiu X, Machin P, Zannoni GF, Scambia G, Benedetti-Panici P, Prat J (2000) Frameshift mutations at coding mononucleotide repeat microsatellites in endometrial carcinoma with microsatellite instability. *Cancer* 88:2290–2297
- C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* 282:2012–2018
- Chang DK, Metzgar D, Wills C, Boland CR (2001) Microsatellites in the eukaryotic DNA mismatch repair genes as modulators of evolutionary mutation rate. *Genome Res* 11:1145–1146
- Coll M, Frederick CA, Wang AH, Rich A (1987) A bifurcated hydrogen-bonded conformation in the d(A.T) base pairs of the DNA dodecamer d(CGCAAATTTGCG) and its complex with distamycin. *Proc Natl Acad Sci USA* 84:8385–8389
- Costanzo MC, Crawford ME, Hirschman JE, Kranz JE, Olsen P, Robertson LS, Skrzypek MS, Braun BR, Hopkins KL, Kondu P, Lengieza C, Lew-Smith JE, Tillberg M, Garrels JI (2001) YPD, PombE PD and Worm PD: Model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res* 29:75–79
- Culigan KM, Meyer-Gauen G, Lyons-Weiler J, Hays JB (2000) Evolutionary origin, diversification and specialization of eukaryotic MutS homolog mismatch repair proteins. *Nucleic Acids Res* 28:463–471
- Dechering KJ, Cuelenaere K, Konings RNH, Leunissen JAM (1998) Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res* 26:4056–4062
- Denver DR, Morris K, Lynch M, Vassilieva LL, Thomas WK (2000) High direct estimate of the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*. *Science* 289:2342–2344
- Denver DR, Morris K, Thomas WK (2003) Phylogenetics in *Caenorhabditis elegans*: An analysis of divergence and outcrossing. *Mol Biol Evol* 20:393–400
- Duret L, Marais G, Biemont C (2000) Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. *Genetics* 156:1661–1669
- Elkin CJ, Richardson PM, Fourcade HM, Hammon NM, Pollard MJ, Predki PF, Glavina T, Hawkins TL (2001) High-throughput plasmid purification for capillary sequencing. *Genome Res* 11:1269–1274
- Frisse LM (1999) Understanding the mechanisms of microsatellite formation and mutation using the model organism *Caenorhabditis elegans*. PhD dissertation. University of Missouri, Kansas City
- Gentles AJ, Karlin S (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res* 11:540–546
- Gordenin DA, Resnick MA (1998) Yeast ARMs (DNA at-risk motifs) can reveal sources of genome instability. *Mutat Res* 400:45–58
- Gragg H, Harfe BD, Jinks-Robertson S (2002) Base composition of mononucleotide runs affects DNA polymerase slippage and removal of frameshift intermediates by mismatch repair in *Saccharomyces cerevisiae*. *Mol Cell Biol* 22:8756–8762
- Hancock JM (1995) The contribution of slippage-like processes to genome evolution. *J Mol Evol* 41:1038–1047
- Harfe BD, Jinks-Robertson S (2000) Sequence composition and context effects on the generation and repair of frameshift intermediates in mononucleotide runs in *Saccharomyces cerevisiae*. *Genetics* 156:571–578
- Hodgkin J, Doniach T (1997) Natural variation and copulatory plug formation in *Caenorhabditis elegans*. *Genetics* 146:149–164
- Barnes T, Kohara M, Coulson Y, Hekimi S (1995) Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. *Genetics* 141:159–179

- Iyer V, Struhl K (1995) Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J* 14:2570–2579
- Kashi Y, King D, Soller M (1997) Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet* 13:74–78
- Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol Biol Evol* 18:1161–1167
- Kroutil LC, Register K, Bebenek K, Kunkel TA (1996) Exonucleolytic proofreading during replication of repetitive DNA. *Biochemistry* 35:1046–1053
- Kruglyak S, Durrett RT, Schug MD, Aquadro CF (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci USA* 95:10774–10778
- Kruglyak S, Durrett R, Schug M, Aquadro CF (2000) Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. *Mol Biol Evol* 17:1210–1219
- Lee RC, Ambros V (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* 294:862–864
- Levy DD, Cebula TA (2001) Fidelity of replication of repetitive DNA in mutS and repair proficient *Escherichia coli*. *Mutat Res* 474:1–14
- Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res* 10:72–80
- Nadir E, Margalit H, Gallily T, Ben-Sasson SA (1996) Microsatellite spreading in the human genome: Evolutionary mechanisms and structural implications. *Proc Natl Acad Sci USA* 93:6470–6475
- Nelson HCM, Finch JT, Luisi BF, Klug A (1987) The structure of an oligo(dA) oligo(dT) tract and its biological implications. *Nature* 330:221–226
- Richard GF, Paques F (2000) Mini- and microsatellite expansions: The recombination connection. *EMBO Rep* 1:122–126
- Richetta A, Ottini L, Falchetti M, Innocenzi D, Bottoni U, Faiola R, Mariani-Costantini R, Calvieri S (2001) Instability at sequence repeats in melanocytic tumors. *Melanoma Res* 11:283–289
- Sen D, Gilbert W (1988) Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* 334:364–366
- Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. *Science* 191:528–535
- Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a major source of genetic variation. *Nature* 322:652–656
- Templeton AR, Clark AG, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF (2000) Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am J Hum Genet* 66:69–83
- Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res* 10:967–981
- Tran HT, Keen D, Kricker M, Resnick MA, Gordenin DA (1997) Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. *Mol Cell Biol* 17:2859–2865
- Vassilieva LL, Lynch M (1999) The rate of spontaneous mutation for life-history traits in *Caenorhabditis elegans*. *Genetics* 151:119–129
- Vassilieva LL, Hook AM, Lynch M (2000) The fitness effects of spontaneous mutations in *Caenorhabditis elegans*. *Evolution* 54:1234–1246
- Wierdl M, Dominska M, Petes TD (1997) Microsatellite instability in yeast: Dependence on the length of the microsatellite. *Genetics* 146:769–779
- Wilder J, Hollocher H (2001) Mobile elements and the genesis of microsatellites in dipterans. *Mol Biol Evol* 18:384–392
- Zhang L, Yu J, Willson JK, Markowitz SD, Kinzler KW, Vogelstein B (2001) Short mononucleotide repeat sequence variability in mismatch repair-deficient cancers. *Cancer Res* 61:3801–3805