Note

Intron Size Correlates Positively With Recombination Rate in *Caenorhabditis elegans*

Anuphap Prachumwat, Laura DeVincentis and Michael F. Palopoli[1]

Department of Biology, Bowdoin College, Brunswick, Maine 04011

Manuscript received October 2, 2003
Accepted for publication December 10, 2003

ABSTRACT

A negative correlation between intron size and recombination rate has been reported for the *Drosophila melanogaster* and human genomes. Population-genetic models suggest that this pattern could be caused by an interaction between recombination rate and the efficacy of natural selection. To test this idea, we examined variation in intron size and recombination rate across the genome of the nematode *Caenorhabditis elegans*. Interestingly, we found that intron size correlated *positively* with recombination rate in this species.

SPLICEOSOMAL introns are widespread and abundant in eukaryotic genomes (Hawkins 1988; Deutsch and Long 1999). For example, it appears that introns constitute ~26, 11, and 24% of the *Caenorhabditis elegans*, *Drosophila melanogaster*, and human genome sequences, respectively (*C. elegans* Sequencing Consortium 1998; Adams *et al.* 2000; Venter *et al.* 2001). Introns impose a burden on organisms harboring them in terms of the energy, time, and materials required for both DNA replication and gene transcription. The large amount of genomic DNA that is devoted to introns in eukaryotes, despite these unavoidable costs, raises the question of what forces drive the evolution of intron size.

Several beneficial functions that are associated with introns have been identified. For example, introns are required for alternative splicing, a post-transcriptional mechanism that allows a single stretch of DNA to code for more than one functional protein (Hanke *et al.* 1999; Caceres and Kornblihtt 2002). Introns also contain functional DNA sequences, such as regulatory elements, alternative promoters, and other genes (Dibb 1993; Duret and Bucher 1997). Interspecific comparisons, however, indicate that intron sequences are not usually under strong functional constraint, since they often differ substantially in nucleotide sequence and length between closely related species (Shabalina and Kondrashov 1999; Kent and Zahler 2000; Robertson 2000; Shabalina *et al.* 2001).

A negative correlation between intron size and local recombination rate has been reported for both the *D. melanogaster* and human genomes (Carvalho and Clark 1999; Comeron and Kreitman 2000). To explain this pattern, Carvalho and Clark (1999) proposed a model in which natural selection favors smaller introns, whereas mutation tends to increase intron size, and it is the balance between these forces that determines intron size at equilibrium. Since the efficacy of natural selection is decreased in regions of the genome that experience reduced recombination rates (Hill and Robertson 1966; Felsenstein 1974), this model predicts the evolution of longer introns where recombination rates are lower (and selection is less effective).

Comeron and Kreitman (2000) argued that the assumption of a mutational bias toward increasing intron size is not supported by recent observations, which suggest that there is an overall mutational bias toward *deletions* in diverse animals (Ogata *et al.* 1996; Petrov *et al.* 1996; Ophir and Graur 1997; Petrov and Hartl 1998, 2000). If mutations are biased toward deletions, then intron size is expected to collapse over evolutionary timescales, unless an opposing force has prevented this collapse from occurring. Comeron and Kreitman (2000) argued that, since introns do not collapse in size, longer introns must often be favored by natural selection. In particular, they proposed that longer introns increase the rate of recombination between adjacent exons and that this effect is beneficial because it allows adjacent exons to evolve with less selective interference (for a more detailed treatment of this model, see Comeron and Kreitman 2002). Like the model proposed by Carvalho and Clark (1999), this model predicts that longer introns will accumulate in regions of reduced recombination, since it is here that selective interference presents the greatest hindrance to the evolution of adjacent exons.

[1] *Corresponding author:* Department of Biology, Bowdoin College, 6500 College Station, Brunswick, ME 04011.
E-mail: mpalopol@bowdoin.edu

To test these models further, we analyzed intron size and recombination rate variation within the genome of the nematode *C. elegans*. Recombination rates and intron sizes vary substantially in this species (Barnes *et al.* 1995; *C. elegans* Sequencing Consortium 1998; Deutsch and Long 1999). On the basis of the population-genetic models described above, we predicted that intron size would correlate negatively with local recombination rate in the *C. elegans* genome.

MATERIALS AND METHODS

Data collection and analysis: The first and last nucleotide positions of both exons and introns for every predicted and confirmed gene were obtained from the flat text file format of the *C. elegans* genome database (Wormbase, http://www.wormbase.org, release WS46, April 2001; Stein *et al.* 2001). For all analyses, data were first imported into Microsoft Excel (version 2001 for Macintosh; Microsoft, Redmond, WA) for data sorting and manipulations. Data were then imported into StatView (version 5.0.1 for Macintosh; SAS Institute, Cary, NC) to conduct statistical analyses and to generate graphs. For intron size *vs.* recombination rate comparison, a bivariate scattergram was generated and Spearman's rank correlation coefficient (corrected for ties) was calculated to test for a significant association between variables. This analysis was completed for each chromosome separately as well as for the entire genome. Results for the complete data set—which includes both predicted and confirmed genes—were checked using only confirmed genes (Wormbase, release WS81, July 2002). To examine regional variation, averages were calculated for 10 equal divisions of each chromosome. The physical lengths of each 10% division were 1.51, 1.53, 1.38, 1.75, 2.09, and 1.77 Mb for chromosomes I, II, III, IV, V, and X, respectively.

Intron size and location: The initial sample included 21,049 predicted genes and 109,128 predicted introns. On the basis of comparisons of genomic positions, we determined that 7706 introns (7.1% of the initial sample) were duplicated in the database. Manual inspection of several hundred of these duplicates indicated that they were in gene sequences that had been assigned more than one name in the database. Only one copy of each duplicated intron was retained for further analysis. A total of 64 predicted introns were <20 bp in length, whereas the shortest known intron size that allows for a successful splicing reaction is ~20 bp (Russell *et al.* 1994). We assumed that introns <20 bp in length resulted from errors made during database curation or by gene prediction software and excluded these from our sample. This left 100,553 introns in our final sample. This number agrees well with the number of introns used in other studies of the complete *C. elegans* genome; for example, Mourier and Jeffares (2003) report an analysis based on 100,569 introns in the complete genome of this species. The middle of each intron was chosen to represent its physical position. Intron size was calculated as one plus the absolute value of first minus last nucleotide positions.

Recombination rate: Recombination rate was estimated as a function of nucleotide position along a chromosome by taking the first derivative of the polynomial function that described the best-fit curve for recombination-map position *vs.* nucleotide coordinate in the genomic sequence, as described in Kliman and Hey (1993). Best-fit curves and first derivatives were calculated using Mathematica (version 4.0 for Macintosh; Wolfram Research, Champaign, IL) and the equations are available upon request. The numbers of loci present in the recombination maps and identified in the genomic sequence were 111, 118, 121, 117, 125, and 181 for chromosomes I, II, III, IV, V, and X, respectively. Small numbers of introns, positioned at the extreme ends of the chromosomes, were outside of the known recombination map. Altogether, 2474 introns fell in these regions, which represented 2.5% of the final sample; for these introns, the local recombination rate was assumed to be the same as the recombination rate of the nearest locus on the genetic map of that chromosome. To make sure that this assumption did not affect our results, all analyses were repeated using a data set that excluded these introns.

RESULTS

Intron size correlated positively with recombination rate for the entire *C. elegans* genome (Figure 1a; Spearman's rank correlation, $R = 0.174$, $P < 0.0001$). A similar pattern was observed when each autosome was considered separately (Figure 1b), but not when the X chromosome was considered separately (Figure 1c). Spearman's rank correlation coefficients were $R =$ 0.206, 0.179, 0.247, 0.212, 0.145, and −0.018 for chromosomes I, II, III, IV, V, and X, respectively. All five autosomes exhibited a rank correlation significantly different from zero, whereas the X chromosome did not (after Bonferroni correction, $P < 0.001$ for each autosome, $P > 0.05$ for the X chromosome). Consistent with the correlation analysis, the slope of the least-squares line for the X chromosome was much closer to zero than that observed for any of the autosomes separately or for the genome as a whole (Figure 1, a and b *vs.* c).

When each chromosome was divided into 10 regions of equal length, average intron size and average recombination rate exhibited parallel distributions throughout the genome (Figure 2), and this positive correlation between regional averages was statistically significant (Spearman's rank correlation, $R = 0.750$, $P < 0.0001$). Both intron sizes and recombination rates tended to be much greater on the autosomal arms than in the autosomal centers. On the X chromosome, however, average intron size did not exhibit much regional variation, and regional averages in intron size on the X chromosome were generally intermediate between those observed for autosomal centers and arms.

Results were similar when the sample was limited to introns from confirmed genes only or to introns from within the known recombination map. For example, on the basis of introns from confirmed genes only, intron size correlated positively with recombination rate across the entire genome (Spearman's rank correlation, $R =$ 0.204, $P < 0.0001$). Similar trends were observed when each autosome was examined separately: Spearman's rank correlation coefficients for introns from confirmed

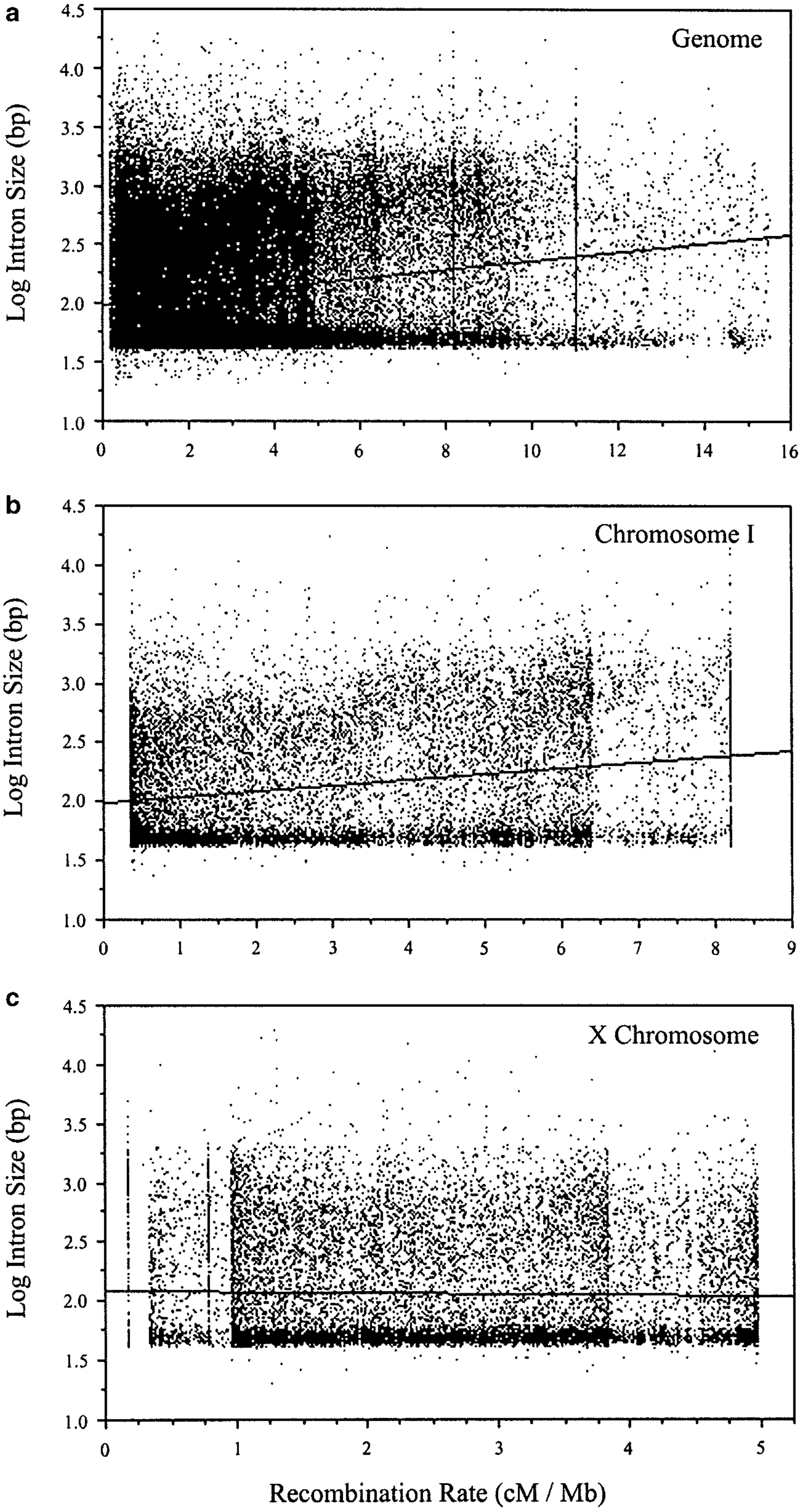

FIGURE 1.—Relationship between intron size (log scale, bp) and local recombination rate (cM/Mb) in the genome of *C. elegans.* (a) Across the entire genome, intron size correlated positively with recombination rate (Spearman's rank correlation, $R = 0.174$, $P < 0.0001$). (b) Similar trends were observed when each autosome was considered separately, such as chromosome I ($R = 0.206$, $P < 0.0001$). (c) When the X chromosome was considered separately, however, there was no significant correlation between intron size and recombination rate ($R = -0.018$, $P > 0.05$). Least-squares lines are provided for illustration of trends.

genes only were $R = 0.268$, 0.243, 0.258, 0.132, and 0.281 for chromosomes I, II, III, IV, and V, respectively. In contrast, no significant correlation was observed for the X chromosome (Spearman's rank correlation, $R = -0.0001$, $P = 0.993$).

DISCUSSION

In the *C. elegans* genome, intron size correlated positively with recombination rate (Figure 1a). This result contrasts with the negative correlation between these variables observed for the *D. melanogaster* and human

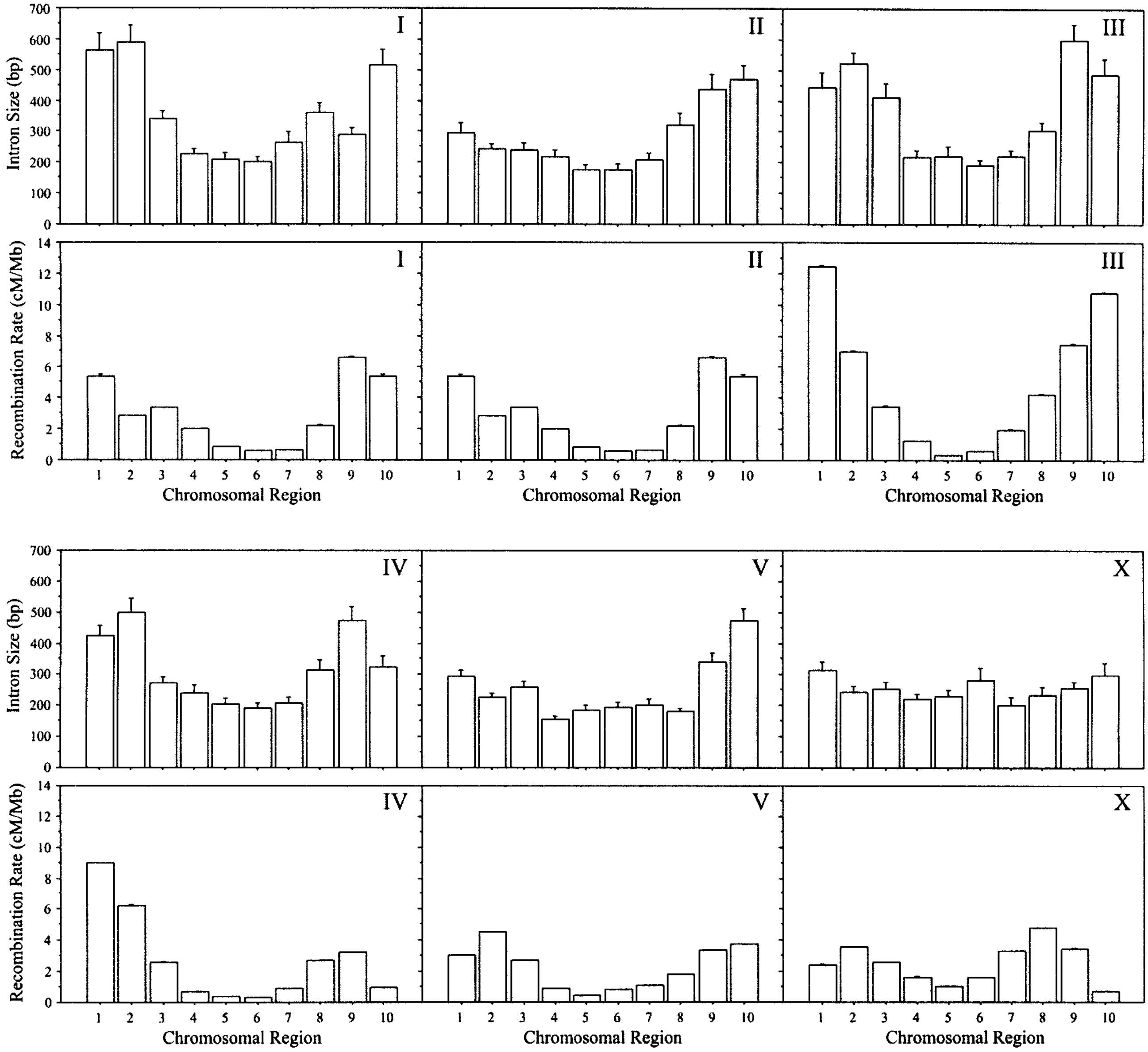

Figure 2.—Comparison of regional averages of intron size (base pairs) and local recombination rate (centimorgans per megabase) across each chromosome in *C. elegans*. These variables exhibited parallel distributions throughout the genome, and the positive correlation between regional averages was statistically significant (Spearman's rank correlation, $R = 0.750$, $P < 0.0001$). Both intron sizes and recombination rates tended to be much greater on the autosomal arms than in the autosomal centers. On the X chromosome, however, average intron size did not exhibit much regional variation. Each chromosome is divided into 10 regions of equal size from left to right. Error bars represent 95% confidence intervals.

genomes and is not predicted by current models for the evolution of intron size (Carvalho and Clark 1999; Comeron and Kreitman 2000).

There were consistent, regional trends in average intron size and average recombination rate across the *C. elegans* genome (Figure 2): (1) autosomal arms tended to have large introns and high recombination rates; (2) autosomal centers tended to have small introns and low recombination rates; and (3) the X chromosome exhibited much less regional variation in average intron size than did any of the autosomes, with average intron sizes intermediate between those observed for autosomal arms and centers. These consistent patterns ruled out the possibility that the genome-wide positive correlation was due to a few regions with widely divergent intron sizes and/or recombination rates.

Population-genetic models have assumed that regional variation in intron size across the genome is determined largely by an interaction between recombination rate and the efficacy of natural selection, termed the Hill-Robertson effect (Carvalho and Clark 1999; Comeron and Kreitman 2000). These models were developed to explain the *D. melanogaster* results and predict a negative correlation between intron size and recombination rate. Our results suggest that these models have omitted the main factor(s) that determines regional variation in intron size across the *C. elegans* genome. Since the correlation between these variables was in the opposite direction in *C. elegans vs.* flies or humans, our results can be explained in at least two ways.

First, one of the current models invoking the Hill-Robertson effect might be valid for flies and humans, but not for nematodes. This explanation for our results implies that the balance of factors determining intron size varies from one evolutionary lineage to another and is supported by the apparent lack of a Hill-Robertson effect on regional variation in both transposon density and codon bias in the *C. elegans* genome (Duret *et al.* 2000; Marais *et al.* 2001; Marais and Piganeau 2002). Instead, recombination-dependent mutational patterns were hypothesized to drive variation in transposon density and codon bias in this species. The same could be true for intron size. For example, the tendency for introns to be larger where recombination rates are higher could result if recombination tended to cause insertions of transposons locally in *C. elegans.*

Second, it is possible that the Hill-Robertson effect is not a major determinant of intron size variation in any of these organisms. This explanation for our results fails to explain the observed correlations (either positive or negative) between intron size and recombination rate. Nevertheless, the fact that recombination rate can be positively correlated with intron size in some species, but negatively correlated in others, raises the question of what other factors might be driving the evolution of intron size.

One interesting possibility is that intron size varies systematically across the genome because the insertion of nonfunctional ("junk") DNA imposes a greater fitness cost in some chromosomal regions than in others. In eukaryotic cells, chromosomal regions harboring dense clusters of active genes are often located toward the center of the nucleus, adjacent to an interchromatin compartment (Lamond and Earnshaw 1998). This spatial organization is thought to provide the most active genes with the best access to the materials needed for transcription and RNA processing. In contrast, genes that are silenced or exhibit low expression levels tend to be located toward the periphery of the nucleus, deep within the interior of chromatin domains, and these regions are thought to remain relatively inaccessible to the transcriptional machinery (Cremer and Cremer 2001; see Mahy *et al.* 2002 for a recent test of this model). Consistent with this model, the spatial organization of genes in the nucleus can be conserved between divergent species (Tanabe *et al.* 2002). If interchromatin compartments are a limiting resource for the cell, then clusters of genes at high density may have evolved to make the best use of limited access to these compartments. Expansion of noncoding DNA in these clusters would tend to impose a fitness cost, and deletions of noncoding DNA in these regions would often be favored by natural selection. In contrast, in those regions of the genome that are *not* often exposed to interchromatin compartments, expansion of noncoding sequences would not tend to impose much of a burden on the cell; hence, insertions of additional DNA in these regions might be neutral or even favorable, if they contributed positively to overall chromosomal architecture.

This chromosome territory model could explain the observed regional variation in intron size in the *C. elegans* genome. In general, this model predicts that regions of the genome that are often exposed to interchromatin compartments should tend to have less noncoding DNA than regions that are usually distant from interchromatin compartments. If the autosomal centers are the regions of the *C. elegans* genome that are most often exposed to interchromatin compartments, then this model could explain the consistent tendency for introns to be smaller in the autosomal centers. According to this interpretation, introns would have evolved to be smaller in the autosomal centers so that more coding DNA could fit into a limited chromosomal region.

Recently, it was reported that genes expressed at higher levels tend to have shorter introns in both humans and *C. elegans* (Castillo-Davis *et al.* 2002). This correlation was interpreted as evidence that natural selection has driven introns to smaller sizes in highly expressed genes to reduce the cost of transcription. The chromosome territory model provides an alternative, although not mutually exclusive, explanation for this pattern: highly expressed genes may be clustered in both species to make effective use of interchromatin compartments, and small intron sizes may have evolved to fit more genes into a smaller region rather than to reduce transcription costs directly.

We are grateful to two anonymous reviewers, Patsy Dickinson, Bruce Kohorn, and Anne McBride for making suggestions to improve the manuscript. We thank the members of the Biology Department for many stimulating discussions on these and related topics. This work was supported by National Science Foundation grant 0110994 to M.F.P.

LITERATURE CITED

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of *Drosophila melanogaster.* Science **287:** 2185–2195.

Barnes, T. M., Y. Kohara, A. Coulson and S. Hekimi, 1995 Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans.* Genetics **141:** 159–179.

C. elegans Sequencing Consortium, 1998 Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282:** 2012–2018.

Caceres, J. F., and A. R. Kornblihtt, 2002 Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet. **18:** 186–193.

Carvalho, A. B., and A. G. Clark, 1999 Intron size and natural selection. Nature **401:** 344.

Castillo-Davis, C. I., S. L. Mekhedov, D. L. Hartl, E. V. Koonin and F. A. Kondrashov, 2002 Selection for short introns in highly expressed genes. Nat. Genet. **31:** 415–418.

Comeron, J. M., and M. Kreitman, 2000 The correlation between intron length and recombination in Drosophila: dynamic equilibrium between mutational and selective forces. Genetics **156:** 1175–1190.

Comeron, J. M., and M. Kreitman, 2002 Population, evolutionary and genomic consequences of interference selection. Genetics **161:** 389–410.

Cremer, T., and C. Cremer, 2001 Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat. Rev. Genet. **2:** 292–301.

Deutsch, M., and M. Long, 1999 Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res. **27:** 3219–3228.

Dibb, N. J., 1993 Why do genes have introns? FEBS Lett. **325:** 135–139.

Duret, L., and P. Bucher, 1997 Searching for regulatory elements in human noncoding sequences. Curr. Opin. Struct. Biol. **7:** 399–406.

Duret, L., G. Marais and C. Biemont, 2000 Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans.* Genetics **156:** 1661–1669.

Felsenstein, J., 1974 The evolutionary advantage of recombination. Genetics **78:** 737–756.

Hanke, J., D. Brett, I. Zastrow, A. Aydin, S. Delbruck *et al.*, 1999 Alternative splicing of human genes: More the rule than the exception? Trends Genet. **15:** 389–390.

Hawkins, J. D., 1988 A survey on intron and exon lengths. Nucleic Acids Res. **16:** 9893–9908.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on the limits to artificial selection. Genet. Res. **8:** 269–294.

Kent, W. J., and A. M. Zahler, 2000 Conservation, regulation, synteny, and introns in a large-scale *C. briggsae–C. elegans* genomic alignment. Genome Res. **10:** 1115–1125.

Kliman, R. M., and J. Hey, 1993 Reduced natural selection associated with low recombination in *Drosophila melanogaster.* Mol. Biol. Evol. **10:** 1239–1258.

Lamond, A. I., and W. C. Earnshaw, 1998 Structure and function in the nucleus. Science **280:** 547–553.

Mahy, N. L., P. E. Perry, S. Gilchrist, R. A. Baldock and W. A. Bickmore, 2002 Spatial organization of active and inactive genes and noncoding DNA within chromosome territories. J. Cell Biol. **157:** 579–589.

Marais, G., and G. Piganeau, 2002 Hill-Robertson interference is a minor determinant of variations in codon bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genomes. Mol. Biol. Evol. **19:** 1399–1406.

Marais, G., D. Mouchinoud and L. Duret, 2001 Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc. Natl. Acad. Sci. USA **98:** 5688–5692.

Mourier, T., and D. C. Jeffares, 2003 Eukaryotic intron loss. Science **300:** 1393.

Ogata, H., W. Fujibuchi and M. Kanehisa, 1996 The size differences among mammalian introns are due to the accumulation of small deletions. FEBS Lett. **390:** 99–103.

Ophir, R., and D. Graur, 1997 Patterns and rates of indel evolution in processed pseudogenes from human and murids. Gene **205:** 191–202.

Petrov, D. A., and D. L. Hartl, 1998 High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. Mol. Biol. Evol. **15:** 293–302.

Petrov, D. A., and D. L. Hartl, 2000 Pseudogene evolution and natural selection for a compact genome. J. Hered. **91:** 221–227.

Petrov, D. A., E. R. Lozovskaya and D. L. Hartl, 1996 High intrinsic rate of DNA loss in Drosophila. Nature **384:** 346–349.

Robertson, H. M., 2000 The large *srh* family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Res. **10:** 192–203.

Russell, C. B., D. Fraga and R. D. Hinrichsen, 1994 Extremely short 20–33 nucleotide introns are the standard length in *Paramecium tetraurelia.* Nucleic Acids Res. **22:** 1221–1225.

Shabalina, S. A., and A. Kondrashov, 1999 Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. Genet. Res. **74:** 23–30.

Shabalina, S. A., A. Y. Ogurtsov, V. A. Kondrashov and A. S. Kondrashov, 2001 Selective constraint in intergenic regions of human and mouse genomes. Trends Genet. **17:** 373–376.

Stein, L. D., P. Sternberg, R. Durbin, J. Thierry-Mieg and J. Spieth, 2001 WormBase: network access to the genome and biology of *Caenorhabditis elegans.* Nucleic Acids Res. **29:** 82–86.

Tanabe, H., S. Müller, M. Neusser, J. von Hase, E. Calcagno *et al.*, 2002 Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. Proc. Natl. Acad. Sci. USA **99:** 4424–4429.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural *et al.*, 2001 The sequence of the human genome. Science **291:** 1304–1351.

Communicating editor: S. Henikoff