ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Progress in Natural Science 19 (2009) 517–522

Progress in Natural Science

www.elsevier.com/locate/pnsc

Short communication

Correlations between recombination rate and intron distributions along chromosomes of *C. elegans*

Hong Li [a,b,*], Guoqing Liu [a], Xuhua Xia [b]

[a] *School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China*
[b] *Department of Biology, University of Ottawa, Ottawa, Ont., Canada K1N 6N5*

Received 13 April 2008; received in revised form 28 May 2008; accepted 4 June 2008

Abstract

Generally speaking, the intron size positively correlates with recombination rate in *Caenorhabditis elegans* genome. Here, we analyze the correlations between recombination rate and some measures of different intron lengths so as to know whether the recombination influences the introns of different lengths in the same way. Results show that the correlation between the recombination rate and the percentage of short introns (<100 bp) is negative, but the correlation between the recombination rate and the percentage of introns that are larger than 500 bp is positive. Average intron length correlates positively with the recombination rate for introns whose length is in the range of 100–1000 bp. We speculate that the recombination mainly exerts impact on introns whose length ranges from 100–1000 bp. We also show that the average intron number per gene correlates negatively with the recombination rate.

Keywords: Recombination rate; Percentage of introns; Intron length; Intron number per gene; *C. elegans*

1. Introduction

Introns are widespread and abundant in eukaryotic genomes. For example, introns contribute about 26% of the *Caenorhabditis elegans* genome sequences. Although introns have some beneficial functions, such as taking part in the process of alternative splicing that allows a single stretch of DNA to code for more than one functional protein [1,2] and encompassing regulatory elements [3,4], their sizes vary within and among genomes indicating the susceptibility of introns to mutations. It has been shown that intron size is influenced by various factors [5,6], such as the insertion of transposable elements [7], the presence of regulatory elements controlling gene expression [8], the presence of RNA genes [9] or RNA involved in gene regulation (e.g. miRNA) [10], the frequency and size of deletion events [11,12], the fixation of multiple species conserved sequences [13], the need to attain higher regulatory capacity [14], and the selection for energy and time economy in gene expression [15,16].

Some studies have shown that the intron size is also correlated with the recombination rate in some eukaryotes. A negative correlation between the intron size and the local recombination rate has been reported for both *Drosophila melanogaster* and human genomes [17–19]. Two models that might explain the results were proposed. First, natural selection favors smaller introns, whereas mutation tends to increase the intron size, and it is the balance between these forces which determines the intron size at equilibrium. In this model, longer introns would accumulate in the regions of reduced recombination rates, since the efficacy of natural selection is decreased in these regions due to Hill–Robertson interference [20,21]. The second model, however,

[*] Corresponding author. Tel.: +86 471 6678889; fax: +86 471 6678236.
E-mail address: lihong499@hotmail.com (H. Li).

1002-0071/$ - see front matter © 2008 National Natural Science Foundation of China and Chinese Academy of Sciences. Published by Elsevier Limited and Science in China Press. All rights reserved.
doi:10.1016/j.pnsc.2008.06.019

predicts that the negative correlation between the intron size and the recombination rate might be caused by selection for favoring longer introns in low recombination regions to reduce Hill–Robertson interference between adjacent exons [18].

Interestingly, Prachumwat et al. observed a positive correlation between the intron size and the local recombination rate in *C. elegans* genome [22], which contrasts with the negative correlation between these variables observed for *D. melanogaster* and human genomes and cannot be predicted by models for the evolution of the intron size [17,18]. The authors explained that Hill–Robertson effect might not be a major determinant of intron size variation in *C. elegans*, but recombination-dependent mutational patterns might be responsible for the variation in the intron size [22], codon bias [23] and transposon density [24].

For example, the tendency for introns to be larger where recombination rates are higher could be caused by the possible preferred insertions of transposons in the recombination active regions in *C. elegans*. An alternative explanation is referred to selection for keeping active chromosomal domains relatively small, which might cluster in autosomal centers where the recombination rate is lower [22].

The global positive correlation between the intron size and the recombination rate in *C. elegans* [22] was obtained considering all the introns together, and the behaviors of different length introns were not distinguished. Generally, intron length varies in a wide range, for example, the longest intron in *C. elegans* genome is up to 99,191 bp, 6.3% introns are longer than 1000 bp and 56.5% introns are shorter than 100 bp. It is hard to imagine that the effect of natural selection and mutation remains to be constant for all the introns with lengths varying from 50 bp to 90,000 bp. Furthermore, total intron length is influenced not only by individual intron length but also by the percentage of introns of different lengths.

Therefore, to get a more detailed relationship between the intron size and the recombination rate so as to know whether the recombination influences the introns of different lengths in the same way in *C. elegans* genome, here, we re-evaluate the association between local recombination rate and intron length through applying some other measures, such as the percentage of introns of different lengths, the average intron length and the average intron number per protein-coding gene, which are defined on the basis of dividing the introns into different length groups.

2. Materials and methods

2.1. Data collection and analysis

All intron annotation data of *C. elegans* genome were taken from GenBank (http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?). The data (including the introns embedded in CDS and those embedded in 5′ UTR and 3′ UTR) contain 105,050 introns, whose length information and chromosomal locations are given. The rough information on intron length is as follows: the length of the intron ranges from less than 10–99,191 bp; there are 56.5% introns whose length is in the range of 40–100 bp; about one third of introns whose length is in the range of 100–1000 bp; 6.29% introns larger than or equal to 1000 bp; only 0.3% introns shorter than 40 bp; and there are 78 introns shorter than 20 bp. Introns smaller than 20 bp might have resulted from the errors made during database creation or by gene prediction software, so we excluded these from our sample. Thus, the remaining 104,972 introns were in our final sample.

We divided each chromosome into *M* equal-sized statistical units. The numbers of the statistical unit were 33, 40, 32, 37, 54 and 33 for chromosomes I, II, III, IV, V, and X, respectively. The average intron number in each statistical unit was 461. All introns were divided into four groups according to their length, such as very short group (group VS) for introns shorter than 100 bp, short group (group S) for introns within the range of 100–500 bp, long group (group L) for introns within the range of 500–1000 bp and very long group (group VL) for introns larger than or equal to 1000 bp.

In each statistical unit, we calculated the percentage of intron number (*PIN*) and average intron length (*AIL*) for each of the four different length groups as well as the average intron number (*AIN*) per gene.

$$PIN_{ij} = N_{ij}/N_i \tag{1}$$

$$AIL_{ij} = L_{ij}/N_{ij} \tag{2}$$

$$AIN_i = N_i/N_{ig} \tag{3}$$

where PIN_{ij} is the percentage of intron number of group *j* in statistical unit *i*, AIL_{ij} is the average intron length of group *j* in statistical unit *i*, AIN_i is the average intron number per gene in statistical unit *i*, N_{ij} is the intron number of group *j* in statistical unit *i*, N_i is the total intron number in statistical unit *i*, L_{ij} is the total intron length of group *j* in statistical unit *i*, and N_{ig} is the protein-coding gene number in statistical unit *i*. Here, $i = 1,2, \ldots, M$ and $j =$ VS, S, L, VL.

We obtained the distributions of the three measures along the chromosomes, and then made linear regressions with the distribution of recombination rate, respectively.

2.2. Recombination rate

The distribution functions of the recombination rate for each chromosome of *C. elegans* were kindly provided by Mr. Palopoli [22] (mpalopol@bowdoin.edu). The distribution function of the recombination rate (RR) was estimated as a function of nucleotide position along each chromosome by taking the first derivative of the polynomial function that described the best-fit curve for the recombination-map position versus the nucleotide coordinate in the genomic sequence. The distributions of the recombination rate along the six chromosomes of *C. elegans* are simple and regular. Generally, high recombination regions are located in the two arms of chromosomes and low recombination regions are located in the centers of the chromosomes. The middle of each statistical unit was chosen to represent its physical position when calculating the recombination rate for each statistical unit using the distribution functions of the recombination rate.

3. Results

3.1. Percentage of intron number

The results of linear regression between the recombination rate and the percentage of intron number (*PIN*) of four intron groups for the six chromosomes of *C. elegans* are shown in Table 1. Because the distributions of *PIN* along chromosomes are similar for the six chromosomes, as an example, only the distributions for chromosome I are shown in Fig. 1.

We can see that the correlations between recombination rate and *PIN* of different length groups are different. *PIN* was correlated significantly negatively with the recombination rate for very short introns (VS) on euchromosomes. For long (L) and very long (VL) introns, however, *PIN* was significantly positively correlated with the recombination rate on euchromosomes. There was no significant correlation between *PIN* and the recombination rate for the S group introns (100–500 bp) across all the chromosomes.

Table 1
Results of linear regression of the percentage of intron number (*PIN*) and the average intron length (*AIL*) with recombination rate for six *C. elegans* chromosomes.[a]

CHR	Length group	*PIN*		*AIL*	
		R	*P*	*R*	*P*
I	VS	−0.71	<0.0001	0.14	0.44
	S	−0.4	0.02	0.46	0.007
	L	0.77	<0.0001	0.74	<0.0001
	VL	0.64	<0.0001	−0.44	0.01
II	VS	−0.64	<0.0001	0.36	0.03
	S	0.17	0.27	0.51	0.001
	L	0.67	<0.0001	0.47	0.002
	VL	0.57	<0.0001	−0.22	0.18
III	VS	−0.76	<0.0001	0.48	0.005
	S	−0.27	0.14	0.65	<0.0001
	L	0.73	<0.0001	0.48	0.005
	VL	0.74	<0.0001	−0.32	0.08
IV	VS	−0.62	<0.0001	0.22	0.18
	S	−0.10	0.59	0.61	<0.0001
	L	0.71	<0.0001	0.42	0.009
	VL	0.66	<0.0001	−0.17	0.32
V	VS	−0.54	<0.0001	0.14	0.33
	S	0.00	0.94	0.45	0.001
	L	0.61	<0.0001	0.56	<0.0001
	VL	0.46	0.001	−0.20	0.16
X	VS	0.25	0.15	0.00	0.74
	S	0.10	0.64	−0.20	0.24
	L	−0.41	0.018	0.02	0.92
	VL	−0.49	0.004	0.17	0.33

[a] *R* is the Pearson correlation coefficient and *P* is the *P* value obtained from significance test; VS, S, L and VL represent the different length intron groups whose corresponding ranges of intron length are <100 bp, 100–500 bp, 500–1000 bp and ⩾1000 bp, respectively.

The *PIN* values varied dramatically between high and low recombination regions. Take chromosome I as an example (Fig. 1), with the recombination rate decreasing from the two arms to the center of the chromosome, the *PIN* values of very short introns increased from about 30% to 60%, and for long and very long introns, *PIN* values decreased from about 25% to or lower than 5%.

In contrast with the results found on euchromosomes, *PIN* and the recombination rate were significantly negatively correlated for L and VL groups on chromosome X. However, there were no significant correlations between the two variables for introns shorter than 500 bp.

3.2. Average intron length

We got the distributions of the average intron length (*AIL*) in four different intron length groups along the six chromosomes. As an example, the distributions for chromosome I are shown in Fig. 2. The results of linear regression between the recombination rate and the *AIL* values are also shown in Table 1.

Positive and significant correlations occurred between *AIL* and the recombination rate when intron length was in the range of 100–1000 bp (S and L groups). For the introns shorter than 100 bp (VS group), the positive correlations between the two variables were significant only on chromosomes II and III. For very long introns, the correlations were negative but not significant except for chromosome I.

On chromosome X, there was no significant correlation between *AIL* and the recombination rate. Compared with euchromosomes, although not significant, the correlations between *AIL* and recombination rate on chromosome X seem to be in contrast with those on euchromosomes, especially for groups S and VL.

3.3. Average intron number per protein-coding gene

The average intron number per protein-coding gene (*AIN*) was calculated in each statistical unit. Linear regression results between the average intron number per gene and the recombination rate for *C. elegans* are shown in Table 2. According to the distribution of the recombination rate along the chromosomes, we divided each chromosome roughly into high recombination region(s) (HRR) and low recombination region(s) (LRR). The high recombination regions corresponded to the ending parts of the two arms of each chromosome, either of which accounted for 25% of the corresponding arm, and the low recombination region was located in the center of each chromosome, which accounted for 50% of each chromosome. The difference (*t*-test) of *AIN* between the high recombination and the low recombination regions are shown in Table 3.

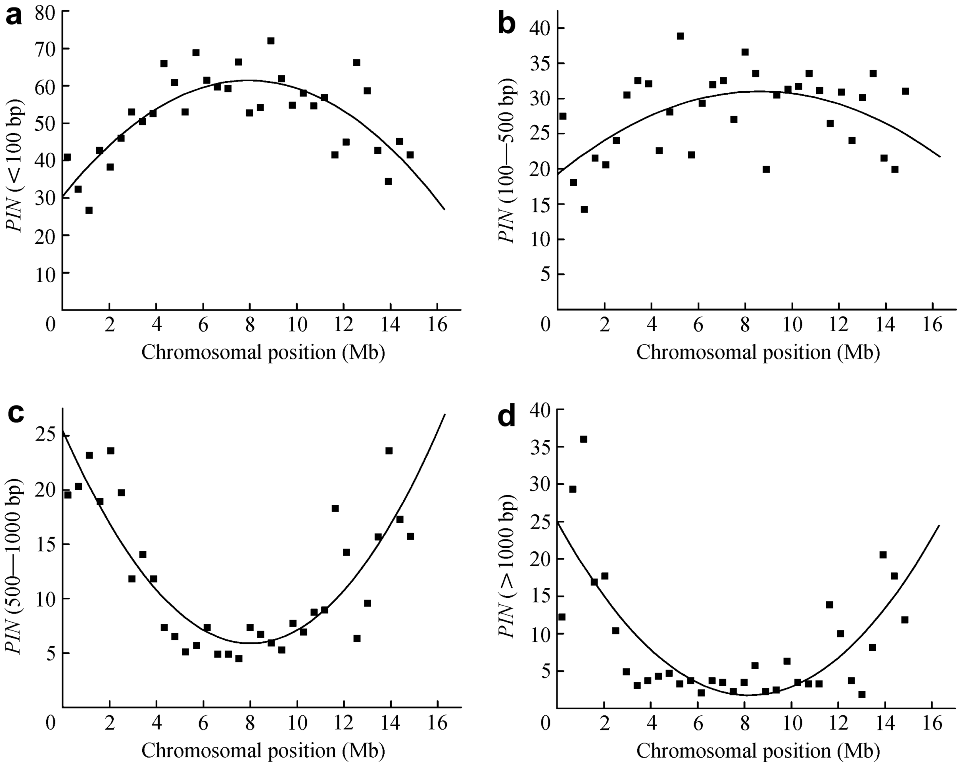

Fig. 1. Distributions of the percentage of intron number (*PIN*) of four intron length groups along chromosome I of *C. elegans*. The abscissa represents the chromosomal positions of the 33 statistical units on chromosome I, and the ordinate represents the *PIN* of intron length groups whose length ranges are, respectively, given in the parentheses of (a), (b), (c), and (d). The fitting curves made by the second-order polynomial function are shown by the solid lines for the illustration of trends.

As shown in Table 2, significant negative correlations were detected between the average intron number per gene and the recombination rate on euchromosomes of *C. elegans* except chromosome IV ($P = 0.79$). That is to say, the average intron number per protein-coding gene generally tends to decrease with the increasing recombination rate on euchromosomes. No significant correlation was detected between the two variables on chromosome X. The average intron number per gene in the high recombination region was less than that in the low recombination region (Table 3).

4. Discussion

The positive correlation [22] between the intron size and the recombination rate can be explained by recombination-dependent mutations rather than by selection. Because if the selection for short introns [15] is responsible for the intron size variation in *C. elegans*, a negative correlation should be expected between the intron size and the recombination rate due to Hill–Robertson effect. Note that the present paper focuses more on the influence of recombination on different length introns, but not on the underlying mechanism on how recombination influences intron length.

Regardless of how the recombination affects intron length, based on the analysis of the three measures *PIN*, *AIL* and *AIN*, we propose that the recombination not only influences intron length *per se*, but also influences the percentage of introns of different lengths and intron numbers in genes. The influence of recombination imposed on introns is different for different intron sizes. With the increasing recombination rate, the percentage of very short introns (<100 bp) decreased from about 60% to about 30%, but the average intron length in this range did not vary significantly; the percentage of short introns (100–500 bp) did not vary obviously, whereas the average intron length increased dramatically. On the contrary, the percentages of long and very long introns increased with the increasing recombination rate. The average intron length also increased in long intron group (500–1000 bp) with the increasing recombination rate, but not in a very long group ($\geqslant$1000 bp).

The intensity of recombination impact exerted on introns varies with the length of introns. Combining the results for *PIN* and *AIL*, we conclude that the impact of recombination is mainly exerted on introns with the length of 100–1000 bp, where higher correlation coefficients were obtained. When intron length exceeds 1000 bp through mutation or decreases to shorter than 100 bp under the impact of natural selection, recombination effect would become weaker or non-significant (Table 4). That is to say, the length ranges of both the very short group and the very long group are intron-cumulated ranges, whereas the intermediate range between them is the recombination-acted range.

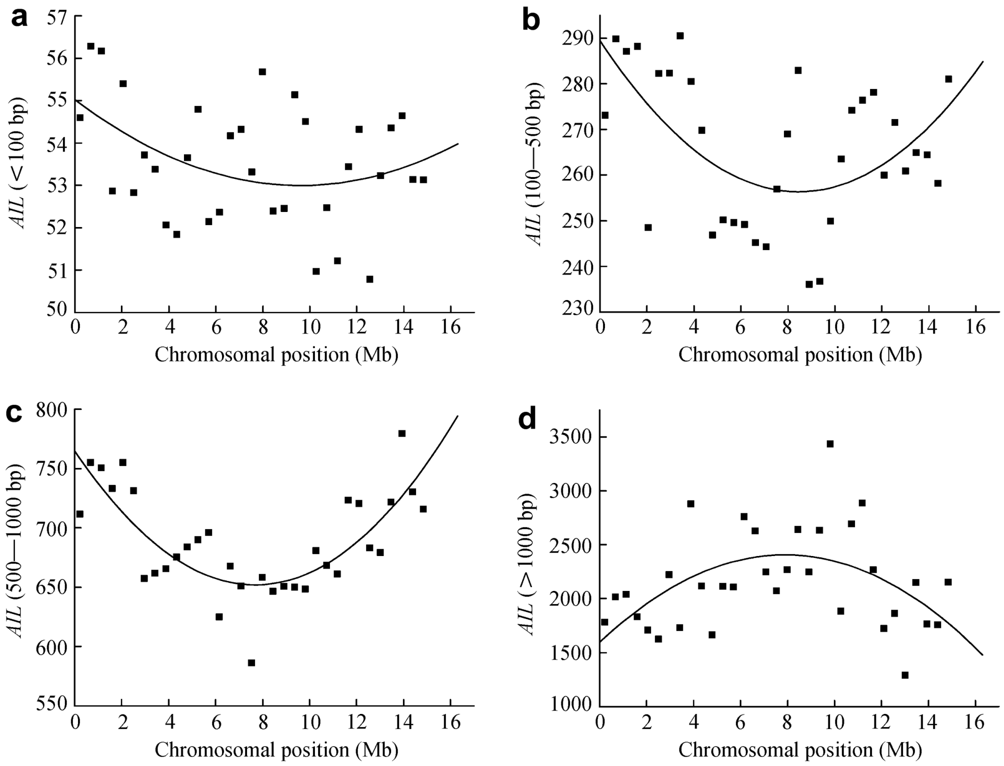

Fig. 2. Distributions of average intron length (*AIL*) in four different intron length groups along chromosome I of *C. elegans*. The abscissa represents the chromosomal positions of the 33 statistical units on chromosome I, and the ordinate represents the *AIL* of intron length groups whose length ranges are, respectively, given in the parentheses of (a), (b), (c), and (d). The fitting curves made by the second-order polynomial function are shown by the solid lines for the illustration of trends.

Table 2
Results of linear regression between average intron number per gene (*AIN*) and recombination rate for *C. elegans*.

	R	*P*
CHR I	−0.61	<0.0001
CHR II	−0.39	0.01
CHR III	−0.53	0.002
CHR IV	−0.00	0.79
CHR V	−0.33	0.01
CHR X	0.10	0.52

Table 3
The difference of average intron number per gene (*AIN*) between low recombination regions (LRRs) and high recombination regions (HRRs) for *C. elegans*.

	HRR	LRR	*P* value
CHR I	5.1	5.8	0.007
CHR II	4.4	5.1	0.001
CHR III	5.1	5.5	0.022
CHR IV	4.7	5.2	0.016
CHR V	4.2	4.8	0.0002
CHR X	5.4	6.1	0.035

The average intron number per protein-coding gene shows significant negative correlations with the recombination rate on euchromosomes, except on chromosome IV.

On chromosome X, the average intron length and the average intron number per gene do not show significant correlations with the recombination rate. While intron length is over 500 bp, the percentage of intron number shows significant negative correlation with the recombination rate on chromosome X, which is in contrast with that on the euchromosomes. These results are consistent with the conclusion that the patterns and processes of molecular evolution may differ between chromosome X and the euchromosomes in *D. melanogaster* [23]. However, note that there is no significant correlation between the percentage of intron number and the recombination rate for introns shorter than 500 bp.

It is interesting to note that although intron length is not affected by the recombination on chromosome X of *C. elegans*, the number of longer introns (>500 bp) is arranged according to the different recombination rates on chromosome X, especially for the introns longer than 1000 bp (Table 1, $R = -0.49$; $P = 0.004$). The cause of it is unclear. We agree that the phenomena concerned with the recombination are caused by the influence of natural selection and mutation on euchromosomes. For example, the variations of the three variables defined in this paper and the GC3 content in CDS [25,26] are the results of the two forces. On chromosome X, however, we conjecture that neither the natural selection nor the mutation concerned with the recombination is strong enough to cause significant impact on intron size evolution. Recombination events on chromosome X, however, may cause some ordering actions,

Table 4
Intron length ranges in which recombination significantly acts on intron length on *C. elegans* euchromosomes.[a]

Intron length range	<100 bp	100–500 bp	500–1000 bp	⩾1000 bp
AIL	−−	++	++	−−
PIN	++	−−	++	++

[a] '++' represents that the measure shows significant correlation with the recombination rate. '−−' represents that the measure does not show significant correlation with the recombination rate.

such as the rearrangement of genes that differ in intron length along chromosome X.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No.30660044) and the Ph.D. Programs Foundation of Ministry of Education of China (Grant No.20050126003).

References

[1] Hanke J, Brett D, Zastrow I, et al. Alternative splicing of human genes: more the rule than the exception? Trends Genet 1999;15(10):389–90.

[2] Caceres JF, Kornblihtt AR. Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet 2002;18(4):186–93.

[3] Dibb NJ. Why do genes have introns?. FEBS Lett 1993;325(1–2): 135–9.

[4] Duret L, Bucher P. Searching for regulatory elements in human noncoding sequences. Curr Opin Struct Biol 1997;7(3):399–406.

[5] Comeron JM. What controls the length of noncoding DNA? Curr Opin Genet Dev 2001;11(6):652–9.

[6] Duret L. Why do genes have introns? Recombination might add a new piece to the puzzle. Trends Genet 2001;17(4):172–5.

[7] Bartolome C, Maside X, Charlesworth B. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. Mol Biol Evol 2002;19(6):926–37.

[8] Bergman CM, Kreitman M. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. Genome Res 2001;11(8):1335–45.

[9] Maxwell ES, Fournier MJ. The small nucleolar RNAs. Annu Rev Biochem 1995;64:897–934.

[10] Mattick JS. Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep 2001;2(11):986–91.

[11] Petrov DA, Sangster TA, Johnston JS, et al. Evidence for DNA loss as a determinant of genome size. Science 2000;287(5455):1060–2.

[12] Petrov DA. DNA loss and evolution of genome size in *Drosophila*. Genetica 2002;115(1):81–91.

[13] Sironi M, Menozzi G, Comi GP, et al. Fixation of conserved sequences shapes human intron size and influences transposon-insertion dynamics. Trends Genet 2005;21(9):484–8.

[14] Pozzoli U, Menozzi G, Comi GP, et al. Intron size in mammals: complexity comes to terms with economy. Trends Genet 2007;23(1):20–4.

[15] Castillo-Davis CI, Mekhedov SL, Hartl DL, et al. Selection for short introns in highly expressed genes. Nat Genet 2002;31(4):415–8.

[16] Chen J, Sun M, Hurst LD, et al. Human antisense genes have unusually short introns: evidence for selection for rapid transcription. Trends Genet 2005;21(4):203–7.

[17] Carvalho AB, Clark AG. Intron size and natural selection. Nature 1999;401(6751):344.

[18] Comeron JM, Kreitman M. The correlation between intron length and recombination in *Drosophila*: dynamic equilibrium between mutational and selective forces. Genetics 2000;156(3):1175–90.

[19] Comeron JM, Kreitman M. Population, evolutionary and genomic consequences of interference selection. Genetics 2002;161(1):389–410.

[20] Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet Res 1966;8(3):269–94.

[21] Kliman RM, Hey J. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol Biol Evol 1993;10(6):1239–58.

[22] Prachumwat A, Devincentis L, Palopoli MF. Intron size correlates positively with recombination rate in *Caenorhabditis elegans*. Genetics 2004;166(3):1585–90.

[23] Singh ND, Davis JC, Dmitri A, et al. Codon bias and non-coding GC content correlate negatively with recombination rate on the *Drosophila* X chromosome. J Mol Evol 2005;61(3):315–24.

[24] Duret L, Marais G, Biemont C. Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. Genetics 2000;156(4):1661–9.

[25] Marais G, Mouchiroud D, Duret L. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. PNAS 2001;98(10):5688–92.

[26] Comeron JM. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics 2004;167(3):1293–304.