6522–6527 *Nucleic Acids Research, 2005, Vol. 33, No. 20*
doi:10.1093/nar/gki970

The excess of 5′ introns in eukaryotic genomes

Kui Lin* and Da-Yong Zhang

MOE Key Laboratory for Biodiversity Science and Ecological Engineering and College of Life Sciences, Beijing Normal University, Beijing 100875, China

Received September 27, 2005; Revised and Accepted November 2, 2005

ABSTRACT

In this work, 21 completely sequenced eukaryotic genomes were analyzed using an intragene comparison approach. We found that all of these genomes show a significant 5′-biased distribution of introns of protein-coding genes. Our findings are different from previous studies based on the intergene method, where introns are biased towards the 5′ end of genes only in intron-poor genomes, but are evenly distributed in intron-rich genomes. In addition, by analyzing the patterns of intron distribution of a set of well-compiled housekeeping genes from human and their respective orthologs identified by a bidirectional best BLAST hit method from the other genomes, we found that the trend of 5′-biased intron positions of the set of housekeeping genes for each genome is much more skewed than that of all genes of the same genome, and rarely if any of the housekeeping genes examined have an extremely 3′-biased position distribution in which all introns of a gene are located only at the 3′ portion of the gene. The most parsimonious explanation for our findings may be the model in which intron loss is caused by homologous recombination between the genomic copy of a gene and a reverse transcriptase product of a spliced mRNA.

INTRODUCTION

With more and more completely sequenced genomes available, a deeper understanding of basic biology can be gained from a comparison of genomes in different evolutionary lineages. In the past decade, comparisons of eukaryotic genome sequences across a broad range of taxa have been unveiling some interesting patterns, such as bigger genomes tending to contain more genes, more and longer introns, and more transposable elements than smaller genomes. In eukaryotes, although the origin of spliceosomal introns, the dynamics of their evolution, and the potential factors that affect it are poorly understood, the pattern of intron positional distribution have been studied recently. Intron positions of protein-coding genes are observed to be unevenly distributed towards the 5′ end of genes in *Saccharomyces cerevisiae* (1). The first genome-wide analysis of intron positional distribution revealed that this bias is both significant in intron-poor genomes and in genes with a single intron from intron-rich genomes (2). A general correlation has been reported between intron density and positional bias at a genome-wide scale, which claims that introns are biased towards the 5′ end of protein-coding genes in intron-sparse genomes, but are evenly distributed within the coding sequence of genes in intron-rich genomes (3). On the other hand, analyzing intron loss in 684 groups of orthologous genes from seven completely sequenced genomes in eukaryotes, including human, has shown that introns closer to the 3′ end of these genes are observed to be preferentially lost during the course of evolution (4,5). All of these studies suggest that both the paucity and positional bias of introns are due to intron loss through a mechanism of homologous recombination of intronless copies of transcripts from 3′ poly-adenylated tails (1–5), although, by comparative genomic analysis in four filamentous fungi, Nielsen *et al*. (6) found no increased frequency of intron loss towards the 3′ end of genes.

To conduct a comprehensive analysis of intron positional distribution within each genome, an intergene comparison approach is usually used at a genome-wide scale (2,3). The position of each intron within its host gene is mapped into an (0,1)-interval relative to its coding sequence length. For each genome, all of the mapped intron positions are pooled into n (e.g. 10) categories, where each category size is one-nth. These n fractions of introns in each category for each genome, then, are used in assessing whether there is a bias of intron positional distribution for the genome or not. However, factors such as differences between genes in terms of gene lengths, expression levels, RT rates, distributions of RT product lengths and rates of gene conversion, may affect the results from such intergene comparisons (4). This is because it may be problematic to compare a part of one gene along its length with another part along a different gene. In this study, we used an intragene analysis method to look at each protein-coding gene within a genome individually. In this case, intron positions of each gene

*To whom correspondence should be addressed at: College of Life Sciences, Beijing Normal University, Beijing 100875, China. Tel: +86 10 58805045; Fax: +86 10 58807721; Email: linkui@bnu.edu.cn

are only compared along with its length and each gene is counted as an independent test in assessment of bias of the intron positions for a given genome. We re-examined the positional distributions of introns for protein-coding genes within eukaryotic genomes, in particular for intron-rich genomes. Our findings show that all genomes we studied have a significant 5′ intron bias. Combining our findings with the results from the previous studies, we suggest that the 5′ bias of intron positions might be due to reverse transcriptase-mediated intron loss. This suggestion is further supported by independent evidence from an analysis of a set of housekeeping genes in each genome. As expected, if the 5′-bias is due to RT-mRNA-mediated intron loss, there must be a more skewed ratio of the number of 5′-biased genes to that of 3′-biased for highly expressed genes in germ line cells other than those of other genes within the same genome, in particular, it must be more unlikely to observe genes with extremely 3′-biased intron positions, namely where all introns are located only in the 3′ portions of the genes, for the highly expressed genes. To this end, by examining the patterns of intron positions from a set of well-compiled housekeeping genes from human genome (7) and their possible orthologs from the other genomes, we indeed obtained a positive observation as expected. Thus, the most parsimonious explanation to our findings is more likely that intron loss in eukaryotes during evolution is mediated by a reverse transcriptase.

MATERIALS AND METHODS

Genome datasets

Twenty-one completely sequenced eukaryotic genomes were studied in this work, including *Anopheles gambiae*, *Apis mellifera*, *Arabidopsis thaliana*, *Caenorhabditis elegans* (WS97), *Candida glabrata*, *Canis familiaris*, *Debaryomyces hansenii*, *Drosophila melanogaster*, *Encephalitozoon cuniculi*, *Eremothecium gossypii*, *Gallus gallus* (NCBI build 1 version 1), *Guillardia theta*, *Homo sapiens* (NCBI build 34 version 3), *Kluyveromyces lactis*, *Mus musculus* (NCBI build 32), *Pan troglodytes* (NCBI build 1 version 1), *Plasmodium falciparum*, *Rattus norvegicus* (NCBI build 2), *S.cerevisiae*, *Schizosaccharomyces pombe* and *Yarrowia lipolytica*. Their genomic annotations were downloaded from the NCBI GenBank database (ftp://ftp.ncbi.nih.gov) and were parsed locally using our scripts. Only the longest coding region was analyzed if multiple alterative spliced transcripts of a gene existed. The 5′-untranslated regions (5′-UTRs) and 3′-UTRs end were not considered in analysis. This maneuver is a conservative one, for there would lead to more robust results if the 5′-bias patterns of intron positions were detected because it is known that the length of the 3′-UTR is comparatively long and there are few introns in the 3′-UTR (8). For accuracy, more stringent criteria were used. Genes whose products were annotated as hypothetical or putative were excluded from the analysis. In addition, genes with incomplete exon positions (denoted as '<' or '>') were also excluded from our analysis. Thus, there were fewer genes being analyzed from *A.gambiae* and *P.falciparum* genomes than those studied by Mourier and Jeffares (3). About 13 210 genes were excluded from *A.gambiae* due to their incomplete exon positions. For the *P.falciparum* genome, only 624 genes with complete exon positions were analyzed because the products of most predicted genes were annotated to be hypothetic proteins due to the difficulty of gene prediction in the genome (9). A survey of each of these genomes is listed in Table 1.

Statistical test for the biased distribution of introns

Similar to the measurement method for the relative position of introns used by Sakurai *et al*. (2), we mapped the positions of introns of each protein-coding gene into an (0, 1)-interval

Table 1. The intron characteristics of interest in the 21 completely sequenced genomes

Species	Number of CDSs	Number of introns	Introns per CDS[a]	Number of intronless CDSs	Percentage of intronless CDSs
Canis familiaris	16 827	198 889	13	1518	9
Gallus gallus	14 250	152 758	11.4	799	5.6
Homo sapiens	20 552	154 358	8.8	2977	14.5
Rattus norvegicus	21 053	162 464	8.7	2333	11.1
Pan troglodytes	21 673	165 510	8.4	1988	9.2
Mus musculus	23 913	154 360	7.7	3865	16.2
Apis mellifera	5798	38 069	6.8	167	2.9
Caenorhabditis elegans	11 754	67 455	5.9	313	2.7
Arabidopsis thaliana	26 258	111 649	5.4	5638	21.5
Drosophila melanogaster	13 181	37 216	3.6	2850	21.6
Anopheles gambiae	1953	4021	2.6	399	20.4
Schizosaccharomyces pombe	1688	1662	2.2	921	54.6
Plasmodium falciparum	567	621	2	258	45.5
Yarrowia lipolytica	6058	703	1.1	5424	89.5
Debaryomyces hansenii	6697	354	1.1	6362	95
Encephalitozoon cuniculi	1468	15	1.1	1454	99
Saccharomyces cerevisiae	4491	260	1	4238	94.4
Eremothecium gossypii	4708	220	1	4492	95.4
Kluyveromyces lactis	5217	128	1	5091	97.6
Candida glabrata	5255	84	1	5173	98.4
Guillardia theta	214	15	1	199	93

All genes annotated either as hypothetical or with incomplete exon positions were excluded. See the Materials and Methods section for details.
[a]The number of introns per CDS was calculated as the number of total introns divided by the subtraction of the number of total CDS from the number of intronless CDS within a genome studied. This is different from the definition by Mourier and Jeffares (3).

relative to its coding sequence length. For each protein-coding gene with m (>0) introns parsed, it was assigned to just one of three categories representing three different intron distribution patterns, namely, 5′-biased, 3′-biased and equally distributed. We counted each gene as an independent test as suggested by Roy and Gilbert (4). Such metrics can avoid the many problematic issues during comparison using intergene method mentioned above.

Since short exons tend to be very rare, the presence of an intron in a given position in a gene should greatly decrease the possibility that another intron will be found in a nearby position, thus introns will tend to be more equally spaced along a gene than would otherwise be expected. Therefore, we restricted our comparison to genes only with unequal intron distributions between 5′ and 3′ end, since the simple expectation that numbers of genes with 5′- and 3′-biases should be equal and should hold no matter what other factors may govern more fine scale positioning of introns. Thus, our assumption is that if introns are randomly distributed along a gene, the probability that a gene should have more introns in its 5′ end of the gene should be equal to that in the 3′ end.

To explore the pattern of intron positional distribution for each species, we counted genes with unequal numbers of introns in the two halves of genes and test whether these two numbers are equal in each genome. The numbers of genes falling into 5′-biased and 3′-biased categories were counted and denoted by O_i, $i = 1, 2$, respectively. The expected numbers of genes in both categories are thus simply equal to $(O_1 + O_2)/2$. Finally, we applied the χ^2-test for goodness of fit to determine whether the intron distribution within a genome is biased or not.

Identification of housekeeping genes for each genome

Is the 5′-biased intron distribution caused by intron loss through gene conversion with a reverse transcriptase product of a spliced mRNA? To test this hypothesis, housekeeping genes with introns from each genome should be ideal objects of analysis. In order to minimize potential bias in selecting housekeeping genes from different genomes, a well-compiled set of housekeeping genes from human (7) is used to identify its possible orthologs in the 20 other genomes. For the simplicity of computation, the amino acid sequences of 86 housekeeping genes from human genome were searched against the corresponding predicted proteome of each of the 20 other genomes. This comparison was performed by a bidirectional best BLAST hit approach (i.e. top reciprocal BLAST hits), respectively. The underlying premise is that orthologs are more similar to each other than they are to any other proteins from the respective genomes at the sequence level, although the resulting gene pairs may not be always closest relatives phylogenetically (10). In our work, we used NCBI BLAST 2.2.6 [April 09, 2003, for Linux IA-64 systems] (11) to search possible homologous housekeeping genes and apply threshold of E-value $<10^{-10}$, identity >40% and aligned length >0.9 * max(L_q, L_s), where L_q (L_s) is the query (subject) sequence length. Here, the relation of gene x in genome i and gene y in genome j is called a bidirectional best hit, when x is the best hit of query y against all genes in genome j and vice versa. Table 3 (the second column) lists the corresponding numbers of the matched housekeeping genes in each genome. We used each set of these housekeeping genes to conduct further analysis of its intron positional distribution for the respective genome.

RESULTS

5′-biased distribution of introns

Table 2 lists the results of a χ^2-test for a biased distribution of introns within each genome we studied. Surprisingly, all

Table 2. The observed and expected distributions of intron positions and results of χ^2-test for biased intron distribution for each genome

Species	Total CDSs	Observed equally distributed	Expected 5′-biased	Observed 5′-biased	Expected 3′-biased	Observed 3′-biased	χ^2-value[a]	χ^2-value[b]	Ratio of 5′-to 3′-biased
C.familiaris	16 827	2501	6404	7214	6404	5594	204.9	118.9	1.3
G.gallus	14 250	2558	5446	5936	5446	4957	88	48.7	1.2
H.sapiens	20 552	3334	7120	8362	7120	5879	432.9	256.6	1.4
R.norvegicus	21 053	3822	7449	8954	7449	5944	608.1	385	1.5
P.troglodytes	21 673	3711	7987	8787	7987	7187	160.3	84.8	1.2
M.musculus	23 913	4035	8006	9535	8006	6478	583.6	364.7	1.5
A.mellifera	5798	1370	2130	2506	2130	1755	132.4	103.8	1.4
C.elegans	11 754	2769	4336	5676	4336	2996	828.2	712.1	1.9
A.thaliana	26 258	3799	8410	8749	8410	8072	27.2	0.1	1.1
D.melanogaster	13 181	1986	4172	5354	4172	2991	669.1	290	1.8
A.gambiae	1953	271	642	689	642	594	7	0.4	1.2
S.pombe	1688	101	333	528	333	138	228.4	137.9	3.8
P.falciparum	567	21	144	199	144	89	42	2.5	2.2
Y.lipolytica	6058	6	314	582	314	46	457.5	35.9	12.7
D.hansenii	6697	2	166	304	166	29	227.1	7.1	10.5
E.cuniculi	1468	1	6	13	6	0	13	–	–
S.cerevisiae	4491	2	126	239	126	12	205.3	–	19.9
E.gossypii	4708	1	108	200	108	15	159.2	–	13.3
K.lactis	5217	0	63	119	63	7	99.6	–	17
C.glabrata	5255	1	40	75	40	6	58.8	–	12.5
G.theta	214	0	8	15	8	0	15	–	–

[a]χ^2-test was performed with df = 1, χ^2-value is 10.83 (6.63) at an α level of 0.001 (0.01).
[b]This χ^2-value is calculated when excluded from all of CDSs with only single intron.

genomes presented significantly 5′-biased intron distribution patterns at the $\alpha = 0.001$ level compared with the expected intron position distributions, except *A.gambiae*, which is significant at the $\alpha = 0.01$ (Table 2, column 8 and 9). The ratios of the number of the 5′-biased genes to that of the 3′-biased genes for each genome vary greatly, ranging from 1.1 (*A.thaliana*) to 19.9 (*S.cerevisiae*), respectively (Column 10 of Table 2). Our results are different from the findings of previous studies. In those studies, introns only in intron-poor eukaryotic genomes (2,3) or in single intron genes (2) had a significant location bias towards to the 5′ end of genes. There was no tendency for 5′-bias observed for intron-rich genomes, such as those of worm, mouse, rat and human (3). In addition, in our results, the observed bias in intron position for the *C.elegans* genome is the most skewed (ratio of 5′ to 3′ reaches to 1.9, Table 2) among the intron-rich genomes studied (in which the number of introns per gene is larger than 3.0 in this study).

It has been reported that genes with a single intron of some intron-rich genomes has a significant tendency of 5′-biased intron distribution (2). In order to test whether our findings may only be caused by genes with a single intron, we conducted the similar analysis of genes with at least two introns for each genome of interest. Interestingly, for intron-rich genomes, they still show significant 5′-biased intron distributions except *A.thaliana* and *A.gambiae*. Both two exceptional genomes contain very similar numbers of genes with either 5′-biased or 3′-biased introns (6453/6480 for *A.thaliana* and 47/33 for *A.gambiae*) (Table 2, column 9). Thus, our findings agree that the intragene analysis, which is suggested by Roy and Gilbert (4) who studied intron loss in 684 groups of orthologous genes from seven eukaryotic genomes and found that introns closer to the 3′ end of genes tend to be lost preferentially except for the *C.elegans* genome, should be much useful for detecting signals for biased intron distributions within intron-rich eukaryotic genomes.

Intron positional distribution within housekeeping genes

If the uneven distribution of introns within each genome results from the biased-loss of introns at the 3′ end of each gene during the course of evolution, it would expect that genes highly expressed in the germ line cells within a genome should tend to have less introns at their 3′ portions than that of the 5′ portions. In particular, highly expressed genes should be much more unlikely to have extremely 3′-biased distribution of introns, namely, all introns of a gene should be located only at the 3′ portion of the gene. To test this hypothesis, a set of well-compiled housekeeping genes from human and their possible orthologs identified from the 20 other genomes through a bidirectional best BLAST hit approach were examined. Given that housekeeping genes perform basic biological functions in cells, they must be expressed ubiquitously, including in germ line cells. If housekeeping genes show a ratio of the number of 5′-biased genes to the number of 3′-biased genes comparable with that of the other genes within a genome, then this would cast doubt on the interpretation that the 5′-biased intron locations observed within the genomes we studied resulted from an excess of intron loss at 3′ end of genes. Since our knowledge about the role of genes is still accumulating and evolving, it might be difficult to identify a true set of housekeeping genes for each specific genome. Therefore, for a conservative analysis, we compared the identified housekeeping genes with the total genes rather than the other ones for the biased ratio for each genome. The list of housekeeping genes from human genome examined here was compiled carefully and stringently by Lahn and his colleagues (7) in their study of the correlation of positive selection of nervous system genes with the evolution of human brains. Table 3 shows the distributions of intron locations for the identified housekeeping genes in each genome we studied. We noted that, compared with 95 individual genes in the original paper, there are fewer housekeeping genes matched in the genome annotation files we analyzed, only 86 for *H.sapiens* matched through exact gene name comparison. Although these housekeeping genes are found to be randomly scattered across their respective genomes in human, rat and mouse genomes (7), we found that the ratios of the number of 5′-biased genes to that of 3′-biased genes in each set of housekeeping genes are indeed much higher than those of the total genes in the same genome (Tables 2 and 3), except for *P.troglodytes* and *A.thaliana* in which the ratios seem to be the same as those of the total genes. These two exceptions may be due to fewer housekeeping genes being analyzed. The higher ratio values indicate that the 5′-biased pattern of intron positions of these housekeeping genes are more skewed than that of the total genes from each genome. Most interestingly, rare if any of housekeeping genes are found to be extremely 3′-biased distribution of introns for each of the 21 genomes (Last column of Table 3), whereas, we found that, among genes with 3′-biased intron distribution, many of them (e.g. ~21% in human, ~24% mouse and ~20% rat) are extremely 3′-biased in their distributions of intron positions.

Table 3. The distributions of intron locations for housekeeping genes (HKGs) in each genome

Species	Number of HKGs				Ratio of	Extremely
	Total[a]	5′-bias	3′-bias	Equal	5′/3′-bias	3′-biased[b]
C.familiaris	61	29	22	10	1.3	0
G.gallus	55	23	17	15	1.4	0
H.sapiens	86	45	16	25	2.8	0
R.norvegicus	66	35	12	19	2.9	0
P.troglodytes	59	22	20	17	1.1	0
M.musculus	70	34	15	21	2.3	1
A.mellifera	34	15	10	9	1.5	3
C.elegans	48	19	10	19	1.9	5
A.thaliana	38	13	12	13	1.1	2
D.melanogaster	46	24	9	13	2.7	5
A.gambiae	4	2	0	2	–	0
S.pombe	36	16	2	18	8	1
P.falciparum	11	7	0	4	–	0
Y.lipolytica	36	20	0	16	–	0
D.hansenii	34	8	0	26	–	0
E.cuniculi	16	1	0	15	–	0
S.cerevisiae	34	7	0	27	–	0
E.gossypii	33	7	0	26	–	0
K.lactis	35	6	0	29	–	0
C.glabrata	33	6	0	27	–	0
G.theta	8	1	0	7	–	0

[a]The total number of housekeeping genes in each genome except *Homo sapiens* is chosen using bidirectional best BLASTP hit approach that is explained in detail in the Materials and Methods section.

[b]Since our knowledge about the role of genes is still accumulating and evolving, it might be difficult to identify a true set of housekeeping genes for each specific genome.

Intron loss in *C.elegans*

There is a large degree of intron turnover within *Caenorhabditis* (12–16), and intron losses are much more frequent than intron gains (14,17). In the previous studies, there was no pattern of 5′-biased introns uncovered in *C.elegans* (3,4). Very interestingly, however, we have found that the *C.elegans* genome not only emerges with a significant 5′-biased intron pattern, but has the most significant 5′-biased pattern among the intron-rich genomes as revealed by statistical tests (Table 2). Its χ^2-critical value reaches 828 and the ratio of the observed number of genes with 5′-biased introns to that of 3′-biased introns reaches 1.9 (Table 2, columns 8 and 10). We also found that there are more genes observed with an equal-distribution of introns and fewer genes with 3′-biased introns relative to the total number of genes in *C.elegans* than in other genomes. About 24% (2769 out of 11754) of genes in *C.elegans* contain equal numbers of introns at both of their 5′ and 3′ portions and only 26% (2996 out of 11 754) of genes contain more introns at their 3′ portions (Table 2). In addition, *C.elegans* has the least number of intronless genes. Only 2.7% of the total numbers of genes in its genome are intronless, compared to *A.mellifera* (2.9%), but 2- to 7-fold less than *G.gallus* (5.6%), *C.familiaris* (9%), *P.troglodytes* (9.2%), *R.norvegicus* (11.1%), *H.sapiens* (14.5%), *M.musculus* (16.1%), *A.gambiae* (20.4%), *A.thaliana* (21.5%) and *D.melanogaster* (21.6%) (Last column of Table 1).

DISCUSSION AND CONCLUSIONS

In this work, we took an intragene comparison strategy to explore the patterns of intron positional distributions on a genome-wide scale for complete eukaryotic genome sequences. We found that introns both from the fully sequenced intron-poor genomes and the fully sequenced intron-rich genomes are significantly biased towards the 5′ end of genes within each of these genomes. These findings are different from those of the previous study (3), which used intergene analysis methods in analyzing the intron-rich genomes. As mentioned above, differences between genes in gene lengths, expression levels, RT rates, distributions of RT product lengths and rates of gene conversion, may affect the results from such intergene comparisons. Accordingly, our results should provide more comprehensive distribution patterns of intron positions for intron-rich genomes than those of previous studies (2,3). However, we must note that two caveats may exist and weaken the results of our analysis. Many of the species studied in this work are vertebrates, which all may have virtually identical intron–exon structures because there is negligible intron gain/loss since the divergence of murine rodents from primates (18). This may also be true for the multiple ascomycote fungi studied here. Secondly, the accuracy of gene predictions may not be good enough at the 3′ end of genes for some completely sequenced genomes, e.g. for the *P.falciparum* genome (9), although we have excluded those hypothetical and putative genes from analysis in order to guarantee the quality of the datasets (see Materials and Methods).

In addition to its 5′-biased intron pattern, we also found that the *C.elegans* genome has the most significant 5′-biased pattern among the intron-rich genomes as revealed by statistical tests. Interestingly, the ratio of the observed number of genes with 5′-biased introns to that of 3′-biased introns for *C.elegans* is the largest among the intron-rich genomes studied (Table 2). This is different from some other recent studies. Cho *et al*. (14) found no positional bias in five genes from six different *Caenorhabditis* species although there are frequent loss of introns during nematode evolution. In the analysis of 684 groups of orthologous genes from seven completely sequenced eukaryotic genomes, Roy and Gilbert (4) found that introns closer to the 3′ end of genes are preferentially lost in *D.melanogaster*, *A.gambiae*, *H.sapiens*, *S.pombe*, *A.thaliana* and *P.falciparum* genomes, but not in the lineage leading to *C.elegans*. This discrepancy over intron position bias in worm might be due to that either intron loss may occur through a qualitatively different mechanism in nematodes (4) or the dataset of *C.elegans* they used might not have enough signatures for identifying potential bias pattern of intron positions, because there are more genes (∼24%) with equally distributed introns of genes in the genome than those in other genomes (Table 1). More careful comparative analysis should help us in better understanding of these inconsistent observations.

Housekeeping genes perform basic biological functions in cells, and are therefore expressed ubiquitously. In theory, mutational events occurring only in the germ line cells are likely to be transferred to next generation so that various mutations could be detected in contemporary genetic sequences. The results of our analysis of the intron positional distributions of the identified housekeeping genes for each genome are very consistent with the prediction of the intron loss model mediated by reverse transcriptases (1–4), which would cause preferential loss of introns at the 3′ end of genes by homologous recombination. This line of evidence is in tension with other recent studies. By analyzing intron presence-absence polymorphism in *Drosophila*, Llopart *et al*. (19) found that the intron loss does not result from a mRNA-mediated mechanism but from a partial deletion at the DNA level. In analysis of introns unique to one species between *C.elegans* and *C.briggsae*, Kent and Zahler (20) proposed that intron loss may be mediated by the mechanism for repair of double-stranded breaks in DNA sequences. Banyai and Patthy (21) founded interesting examples of intron loss even in the 5′ end of very long multidomain genes in *D.melanogaster* and *C.elegans*. However, we believe that more careful analysis of the complete gene structures of housekeeping genes, including 5′-UTRs and 3′-UTRs, may further help us understand mechanisms that cause this bias of intron positional distributions. Besides the housekeeping genes, highly co-regulated genes expressed in the germ line cells should be another good indicator of intron loss biased towards the 3′ end of these genes, unless there are functional sites within their 3′ introns.

Except for the preferential loss of 3′ introns, the observed pattern of a greater number of genes with a 5′-biased intron distribution in a eukaryotic genome could be also due to other reasons, such as biased fixation of gained introns in the 5′ end of genes. The mechanisms underlying intron gain are not understood well. Recently, large-scale analyses have found evidence of intron gains (22), but rarely if at all in mammalian genes (18). Although old introns are found to have a significant bias in the 5′ portions of genes, new introns show a nearly even distribution along the gene, especially, a 3′ bias tendency in

intron-rich genomes (5). By identifying the pattern of intron conservation of orthologous genes from four filamentous fungal genomes, Nielsen *et al*. (6) predicted a set of intron gains with certain signatures. However, no statistically significant bias was detected in the position of gained introns along the coding sequence. All of these studies show that there is no apparent evidence for biased intron gains in evolution. Taken together, the most parsimonious explanation for the excess of 5′ end introns within the 21 eukaryotic genomes is to suggest that genes lose their 3′ end introns preferentially through gene conversion by homologous recombination between a copy of a spliced transcript with its corresponding genomic sequence. In this case, the potency of reverse transcriptases might be one of main forces driving the evolution of eukaryotic gene structures by providing a higher rate of loss for 3′ introns, in particular for genes highly expressed in the germ line cells, such as housekeeping genes.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their invaluable comments. We would like to thank Lei Zhu and Yingqin Luo for parsing annotation information from the downloaded GenBank files into the local database. This study has been supported by the MOE EYTP 2003 and by Beijing Normal University. Funding to pay the Open Access publication charges for this article was provided by Beijing Normal University.

Conflict of interest statement. None declared.

REFERENCES

1. Fink,G.R. (1987) Pseudogenes in yeast? *Cell*, **49**, 5–6.
2. Sakurai,A., Fujimori,S., Kochiwa,H., Kitamura-Abe,S., Washio,T., Saito,R., Carninci,P., Hayashizaki,Y. and Tomita,M. (2002) On biased distribution of introns in various eukaryotes. *Gene*, **300**, 89–95.
3. Mourier,T. and Jeffares,D.C. (2003) Eukaryotic intron loss. *Science*, **300**, 1393.
4. Roy,S.W. and Gilbert,W. (2005) The pattern of intron loss. *Proc. Natl Acad. Sci. USA*, **102**, 713–718.
5. Sverdlov,A.V., Babenko,V.N., Rogozin,I.B. and Koonin,E.V. (2004) Preferential loss and gain of introns in 3′ portions of genes suggests a reverse-transcription mechanism of intron insertion. *Gene*, **338**, 85–91.
6. Nielsen,C.B., Friedman,B., Birren,B., Burge,C.B. and Galagan,J.E. (2004) Patterns of intron gain and loss in fungi. *PLoS Biol*., **2**, e422.
7. Dorus,S., Vallender,E.J., Evans,P.D., Anderson,J.R., Gilbert,S.L., Mahowald,M., Wyckoff,G.J., Malcom,C.M. and Lahn,B.T. (2004) Accelerated evolution of nervous system genes in the origin of Homo sapiens. *Cell*, **119**, 1027–1040.
8. Hawkins,J.D. (1988) A survey on intron and exon lengths. *Nucleic Acids Res*., **16**, 9893–9908.
9. Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S. *et al*. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498–511.
10. Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol*., **52**, 540–542.
11. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
12. Coghlan,A. and Wolfe,K.H. (2004) Origins of recently gained introns in *Caenorhabditis*. *Proc. Natl Acad. Sci. USA*, **101**, 11362–11367.
13. Logsdon,J.M.,Jr (2004) Worm genomes hold the smoking guns of intron gain. *Proc. Natl Acad. Sci. USA*, **101**, 11195–11196.
14. Cho,S., Jin,S.W., Cohen,A. and Ellis,R.E. (2004) A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. *Genome Res*., **14**, 1207–1220.
15. 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.
16. Rogozin,I.B., Wolf,Y.I., Sorokin,A.V., Mirkin,B.G. and Koonin,E.V. (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol*., **13**, 1512–1517.
17. Kiontke,K., Gavin,N.P., Raynes,Y., Roehrig,C., Piano,F. and Fitch,D.H. (2004) *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc. Natl Acad. Sci. USA*, **101**, 9003–9008.
18. Roy,S.W., Fedorov,A. and Gilbert,W. (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl Acad. Sci. USA*, **100**, 7158–7162.
19. Llopart,A., Comeron,J.M., Brunet,F.G., Lachaise,D. and Long,M. (2002) Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. *Proc. Natl Acad. Sci. USA*, **99**, 8121–8126.
20. Kent,W.J. and Zahler,A.M. (2000) Conservation, regulation, synteny, and introns in a large-scale *C.briggsae-C.elegans* genomic alignment. *Genome Res*., **10**, 1115–1125.
21. Banyai,L. and Patthy,L. (2004) Evidence that human genes of modular proteins have retained significantly more ancestral introns than their fly or worm orthologues. *FEBS Lett*., **565**, 127–132.
22. Qiu,W.G., Schisler,N. and Stoltzfus,A. (2004) The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol. Biol. Evol*., **21**, 1252–1263.