J Mol Evol (2006) 63:136–141
DOI: 10.1007/s00239-005-0261-6

JOURNAL OF MOLECULAR EVOLUTION

A Gradient in the Distribution of Introns in Eukaryotic Genes

A. Ruvinsky, W. Ward*

The Institute for Genetics and Bioinformatics, University of New England, Armidale 2351 NSW, Australia

Received: 3 November 2005 / Accepted: 13 February 2006 [*Reviewing Editor*: Dr. Manyuan Long]

Abstract. The majority of eukaryotic genes consist of exons and introns. Introns can be inserted either between codons (phase 0) or within codons, after the first nucleotide (phase 1) and after the second (phase 2). We report here that the frequency of phase 0 increases and phase 1 declines from the 5′ region to the 3′ end of genes. This trend is particularly noticeable in genomes of *Homo sapiens* and *Arabidopsis thaliana*, in which gains of novel introns in the 3′ portion of genes were probably a dominant process. Similar but more moderate gradients exist in *Drosophila melanogaster* and *Caenorhabditis elegans* genomes, where the accumulation of novel introns was not a prevailing factor. There are nine types of exons, three symmetric (0,0; 1,1; 2,2) and six asymmetric (0,1; 1,0; 1,2; 2,1; 2,0; 0,2). Assuming random distribution of different types of introns along genes, one can expect the frequencies of asymmetric exons such as 0,1 and 1,0 or 1,2 and 2,1 to be approximately equal, allowing for some variation caused by randomness. The gradient in intron distribution leads to a small but consistent and statistically significant bias: phase 1 introns are more likely at the 5′ ends and phase 0 introns are more likely at the 3′ ends of asymmetric exons. For the same reason, the frequency of 0,0 exons increases and the frequency of 1,1 exons decreases in the 3′ direction, at least in *H. sapiens* and *A. thaliana*. The number of introns per gene also affects the distribution and frequency of phase 0 and 1 introns. The gradient provides an insight into the evolution of intron-exon structures of eukaryotic genes.

Key words: Intron phase distribution — Asymmetric exons — Eukaryotic genes — Genomes

Introduction

The majority of eukaryotic genes consist of exons and introns. Introns are located either between codons (phase 0) or within codons, after the first nucleotide (phase 1) and after the second (phase 2). This creates nine types of exons, three symmetric (0,0; 1,1; 2,2) and six asymmetric (0,1; 1,0; 1,2; 2,1; 2,0; 0,2). It remains unclear how and when introns were inserted in eukaryotic genes, however, it is known that phase 0 introns are more common and phase 2 introns are rare (Fedorov et al. 1992). The variety of intron phase proportions can be generally described as being at a ratio of 5:3:2 for phases 0, 1, and 2, respectively, certainly with broad deviations in different species (Qui et al. 2004). Several explanations were proposed for this pattern, including exon shuffling (Fedorov et al. 1992), sequence conservation of splice signals in exons (Long and Deutsch 1999), correlation with regions of amino acid conservation for phase 1 and 2 introns (Endo et al. 2002), specific elements of protein structure (Gilbert et al. 1997), and intron sliding (Lynch 2002).

Recent simulation experiments showed that species specific codon usage frequencies significantly affect intron phase distribution, making it surprisingly similar to the observed in a particular species (Ru-

**Current address:* School of Mathematics, Statistics and Computer Sciences, University of New England, Armidale 2351 NSW, Australia
Correspondence to: A. Ruvinsky; *email:* aruvinsk@une.edu.au

vinsky et al. 2005). Thus, the conclusion can be drawn that the observed bias in the intron phase distribution is at least in part caused by codon usage frequencies. Another feature relevant to the distribution of intron phases is the symmetry of exons. Symmetric exons are overrepresented in eukaryotic genes and hence asymmetric exons are underrepresented (Long et al. 1995; Long and Rosenberg 2000; Kaessmann et al. 2002). Assuming random distribution of introns along genes, one can expect the frequencies of asymmetric exons such as 0,1 and 1,0 to be approximately equal, allowing for some variation caused by randomness. Here we reported that in four model species there is a small but consistent and statistically significant bias: phase 1 introns are more likely located at the 5′ ends and phase 0 introns are more likely located at the 3′ ends of asymmetric exons. This trend is particularly noticeable in the genomes of *Homo sapiens* and *Arabidopsis thaliana*.

This raises the question of spatial distribution of different phase introns along the genes. Little is known about this except that the 5′ regions of human genes are saturated with phase 1 exons (Torday and Patthy 2004). In this paper we presented data on the distribution of different phase introns along the genes and showed that the frequency of phase 0 increases and that of phase 1 declines from the 5′ region to the 3′ end of genes. The existence of such gradients also leads to frequency bias among pairs of asymmetric exons as well as to different trends in the distribution of symmetric exons. The latest publications suggest that novel inrons are unevenly distributed along the genes (Sverdlov et al. 2004; Rogozin et al. 2005). It has also been shown that novel introns are more frequent in phase 0 (Rogozin et al. 2003; Qui et al. 2004; Coglan and Wolfe 2004). This provides a possible explanation for the observed gradient.

Materials and Methods

Gene Data

Information relevant to *A. thaliana*, *C. elegans*, *D. melanogaster*, and *H. sapiens*, was extracted from the Exon-Intron Database (EID, version 112), which was compiled in the W. Gilbert laboratory, Department of Molecular and Cellular Biology, Harvard University (Saxonov et al. 2000). The initial database was extensively purged by J. Chamary (University of Bath, Bath, UK). The removal of potential duplicates was done after performing an all-against-all BLAST, with an expected value of $p < 0.001$ (Lercher et al. 2004) and creating clusters of duplicated genes. The longest of the duplicate genes were left in the database. This procedure was based on the assumption that, in the case of alternative transcripts, the longest is the constitutive form. Even if this is not the case, it is just an arbitrary way of selecting one duplicate. Then one from the "longest" duplicates, if several are of the same length, was randomly selected. The total numbers of studied genes and introns are reported in Table 1. The authors will provide the dataset on request.

Table 1. Numbers of genes and exons studied in genomes of model species

Species	No. of genes	No. of exons	Exons per gene
H. sapiens	2,033	18,465	9.08
D. melanogaster	8,502	37,856	4.45
C. elegans	10,312	69,180	6.71
A. taliana	9,914	65,766	6.63

Table 2. Observed numbers of asymmetric exons in genomes of model species and χ^2 comparisons

	Comparison of asymmetric exons					
Species	0,1	1,0	0,2	2,0	1,2	2,1
Hs						
Observed	1745	**1812**	1273	1364	**1069**	946
χ^2	*1.26*		*3.14*		***7.51***	
p	0.2617		0.0764		0.0061	
Dm						
Observed	2495	**2613**	2313	2261	**1757**	1624
χ^2	*2.73*		*0.59*		***5.23***	
p	0.0985		0.4424		0.0222	
Ce						
Observed	5483	**5831**	5923	5896	**3215**	3144
χ^2	***10.70***		*0.06*		*0.79*	
p	0.0011		0.8065		0.3741	
At						
Observed	4868	**5313**	5478	5622	**2085**	1977
χ^2	***19.45***		*1.87*		*2.87*	
p	< 0.0001		0.1715		0.0902	

Note. Hs, *H. sapiens*; Dm, *D. melanogaster*; Ce, *C. elegans*; At, *A. thaliana*. Boldface numbers show higher values for 1, 0 and 1, 2 exons. Italic boldface numbers show significant χ^2 values.

Databases and Software

From the purged databases, four separate exon databases, for each of the model species (Hs [*Homo sapiens*], Dm [*Drosophila melanogaster*], Ce [*Caenorhabditis elegans*], and At [*Arabidopsis thaliana*]), were created using a Perl script. The sequences were read into objects using modules from the BioPerl toolkit and regular expressions were used to iterate through the gene sequences identifying each exon. In addition to the exon sequences, the exon length, the nucleotide number for the starting position of each exon, its 5′ and 3′ phases, and the gene ID numbers were extracted and included as fields in the databases. The primary keys were created by appending the exon number to the ID number of each gene ID. Outputs from the program were loaded into the Postgres DBMS and SQL and Perl scripts, which were used to generate the statistical data. Analysis was then performed using Microsoft Excel, R statistical software packages, Spearman rank correlation (http://www.wessa.net/rankcorr.wasp), and other standard statistical procedures.

Results

Frequencies of Asymmetric Exons

Assuming random distribution of different types of introns along genes, one can expect the frequencies of asymmetric exons such as 0,1 and 1,0 or 1,2 and 2,1 to be approximately equal, allowing for some variation

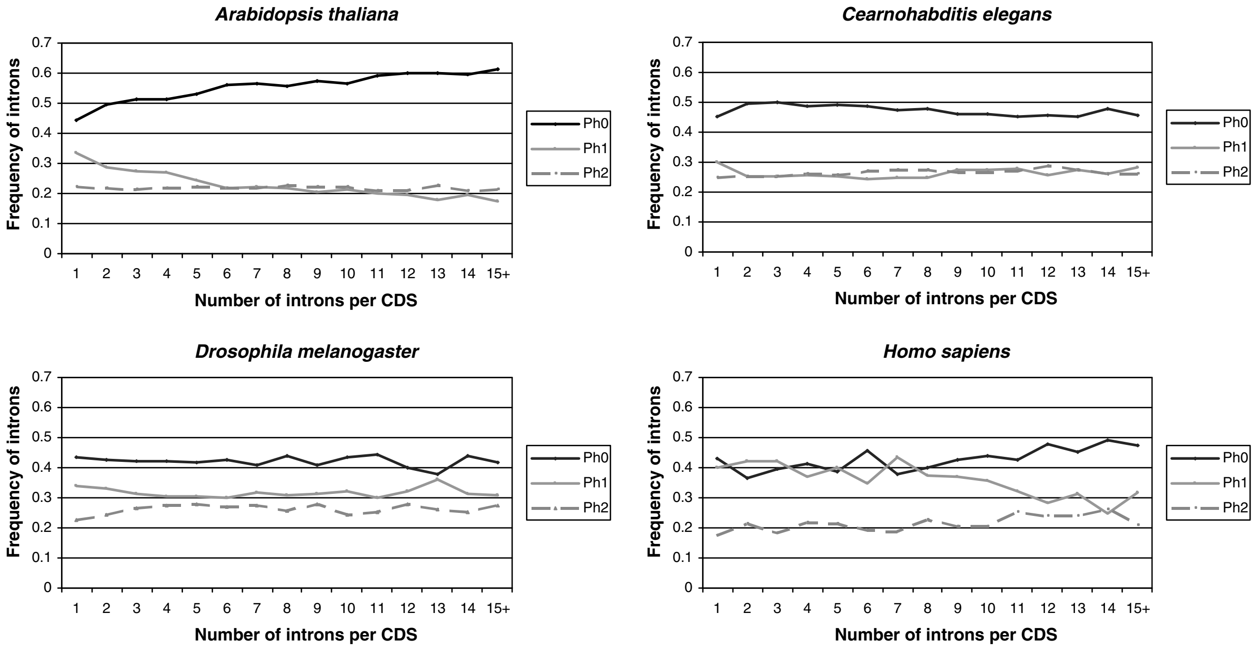

Fig. 1. Frequency of intron phases in the genomes of model species depending on the number of introns per coding sequence.

caused by randomness. However, as Table 2 shows, this is not the case. In all four compared species 1,0 exons are more frequent, and the differences are statistically significant in *C. elegans* and *A. thaliana*. Exons 1,2 are also more common in all studied species and the differences are statistically significant in *H. sapiens* and *D. melanogaster*. In the studied species (*H. sapiens*, *D. melanogaster*, *C. elegans*, and *A. thaliana*), there is a relatively small (~2–4%) but consistent bias in favor of asymmetric exons having phase 1 introns at their 5′ ends rather than at their 3′ ends (Hs, $\chi^2 = 6.48$, $p = 0.0109$; Dm, $\chi^2 = 7.42$, $p = 0.0065$; Ce, $\chi^2 = 9.93$, $p = 0.0016$; At, $\chi^2 = 21.47$, $p < 0.0001$). Conversely, phase 0 introns are more common at the 3′ end of asymmetric exons in the studied species except *D. melanogaster* (Hs, $\chi^2 = 4.03$, $p = 0.0447$; Ce, $\chi^2 = 4.45$, $p = 0.0349$; At, $\chi^2 = 16.3$, $p < 0.0001$).

Despite obvious differences between the compared genomes, the trend is similar. Thus, combined data that include different types of exons indicate that phase 1 introns are more likely found at the 5′ end of asymmetric exons rather than at the 3′ end, and alternatively, phase 0 introns are more frequently located at the 3′ end of asymmetric exons. It does not appear that a distribution of phase 2 introns is biased in favor of one of the alternative exonic ends in any of the studied species except *D. melanogaster*.

Distribution of Introns with Different Phases Along Genes

The observed facts may have a simple explanation if a gradient in the distribution of different phase introns along the gene also exists. It is essential to take into account that the compared genomes have quite different distributions of introns per gene (Supplementary Fig. 1). The proportion of genes with a few introns was higher in *D. melanogaster* and *C. elegans* than in *H. sapiens* and *A. thaliana*. In order to study the intron distributions in the genomes of the model species we calculated the frequencies of phase 0 and 1 introns in all positions along the genes of different lengths. Tables 3 and 4 present the results for *A. thaliana*. Phase 1 introns are frequent at the 5′ ends of exons and steadily decline in the 3′ direction. The difference in occurrence of phase 1 introns between the first and the last positions is highly significant ($t = 5.1$, $p = 0.00013$). The frequency of phase 0 introns, on the other hand, increases very significantly from the first to the last exon position ($t = 8.65$, $p =$ 8E-07). A very similar trend is found in *H. sapiens* genome (Supplementary Tables 1 and 2). The frequency of phase 1 introns is high at the first positions and declines significantly in the 3′ direction ($t = 3.95$, $p = 0.001$); phase 0 shows the alternative pattern ($t = 5.45$, $p =$ 7E-05).

In genomes of *C. elegans* and *D. melanogaster* the trend is similar but not as strong. The frequency of phase 1 introns in *C. elegans* declines significantly toward the ends of the genes ($t = 4.28$, $p = 0.00053$) and phase 0 rises in the opposite direction ($t = 2.8$, $p = 0.008$). However, the major contribution to these changes comes from the genes with a smaller number of introns (Supplementary Tables 3 and 4). In *D. melanogaster* there is a significant decline in the frequency of phase 1 introns in the 3′ direction ($t = 3.24$, $p = 0.004$), but an increase in the frequency of

Table 3. Frequency of phase 1 introns per position in genes of *Arabidopsis thaliana*

No. of introns per sequence	Position of intron in CDS													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.334													
2	0.299	0.276												
3	0.298	0.266	0.257											
4	0.342	0.249	0.236	0.245										
5	0.304	0.261	0.255	0.206	0.197									
6	0.248	0.252	0.213	0.229	0.194	0.18								
7	0.25	0.206	0.226	0.237	0.182	0.219	0.228							
8	0.297	0.243	0.241	0.184	0.201	0.182	0.199	0.205						
9	0.252	0.249	0.207	0.202	0.149	0.214	0.156	0.224	0.184					
10	0.271	0.259	0.218	0.227	0.196	0.199	0.215	0.186	0.192	0.151				
11	0.274	0.297	0.229	0.169	0.188	0.207	0.188	0.165	0.117	0.207	0.165			
12	0.289	0.239	0.203	0.188	0.168	0.198	0.198	0.198	0.137	0.147	0.188	0.188		
13	0.24	0.253	0.2	0.18	0.2	0.153	0.173	0.147	0.113	0.14	0.207	0.147	0.147	
14	0.28	0.258	0.159	0.167	0.197	0.212	0.174	0.197	0.182	0.22	0.197	0.129	0.167	0.212

Table 4. Frequency of phase 0 introns per position in genes of *Arabidopsis thaliana*

No. of introns per CDS	Position of intron in CDS													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.442													
2	0.483	0.505												
3	0.49	0.519	0.531											
4	0.445	0.546	0.563	0.502										
5	0.482	0.518	0.514	0.564	0.581									
6	0.537	0.512	0.544	0.572	0.608	0.605								
7	0.504	0.561	0.586	0.563	0.594	0.568	0.57							
8	0.479	0.529	0.538	0.582	0.584	0.584	0.582	0.561						
9	0.481	0.542	0.587	0.572	0.627	0.574	0.62	0.569	0.607					
10	0.495	0.53	0.549	0.565	0.562	0.593	0.562	0.599	0.571	0.631				
11	0.5	0.549	0.556	0.564	0.59	0.575	0.62	0.624	0.68	0.579	0.658			
12	0.523	0.553	0.589	0.579	0.64	0.563	0.594	0.619	0.655	0.645	0.614	0.604		
13	0.54	0.527	0.607	0.58	0.573	0.673	0.547	0.607	0.607	0.593	0.6	0.633	0.7	
14	0.492	0.591	0.598	0.621	0.538	0.606	0.606	0.545	0.614	0.583	0.644	0.621	0.606	0.644

phase 0 introns is not evident (Supplementary Tables 5 and 6).

Frequencies of Intron Phases in Genes of Different Lengths

Figure 1 demonstrates that in *A. thaliana* and *H. sapiens* the frequencies of intron phases depend on the number of introns per CDS. Phase 0 becomes more frequent and phase 1 less frequent as the number of introns per CDS increases. The frequencies of phase 1 and 0 introns are quite similar in those genes which have only a few introns and are dissimilar in multi-intron genes. In *A. thaliana* this trend is obvious across all gene lengths and the Spearman rank correlation is highly significant ($r_s = 0.76$, $t = 4.03$, $p = 0.00083$). In *H. sapiens* the trend becomes evident in those genes which have more than eight introns in the CDS and the Spearman rank correlation is also significant ($r_s = 0.69$, $t = 3.32$, $p = 0.003$). Phase 2 introns have a flat distribution in *A. thaliana* and rise steadily in *H. sapiens*. In *C. elegans* the frequencies of all intron types remain unchanged across all gene lengths, except for the first couple of positions. However, this difference is sufficient to cause higher frequencies of 1,0 exons compared with 0,1 exons. In *D. melanogaster* the frequency of phase 2 introns increases slightly in the genes with only a few introns and this leads to higher proportions of 1,2 exons compared with 2,1 exons. The Spearman rank correlation between phase 1 and phase 2 introns is also highly significant ($r_s = 0.99$, $t = 25.95$, $p < 0.000001$).

Table 5. Comparison of frequencies of old and new introns in the 5′ and 3′ halves of coding sequences in the genomes of model species[a]

	Number of old introns		Number of new introns		
	5′ region	3′ region	5′ region	3′ region	χ^2
A. thaliana	598 (53.2%)	510 (46.8%)	1637 (47%)	1847 (53%)	16.42 (p = 0.0001)
H. sapiens	1045 (54.3%)	880 (45.7%)	1797 (49.2%)	1860 (50.8%)	13.37 (p = 0.0003)
C. elegans	487 (56.1%)	381 (43.9%)	947 (52%)	874 (48%)	3.97 (p = 0.046)
D. melanogaster	440 (54.5%)	368 (45.5%)	277 (52.7%)	249 (47.3%)	0.41 (p = 0.52)

[a] Original data were taken from Sverdlov et al. (2004, Table 1).

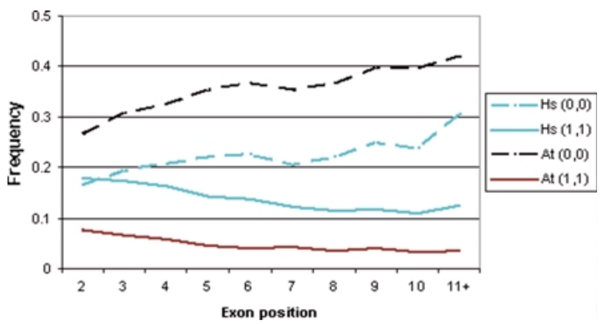

Fig. 2. Frequency of symmetric 0,0 and 1,1 exons in genomes of *Homo sapiens* and *Arabidopsis thaliana*.

Distribution of Symmetric Exons Along Genes

An excess of symmetric exons (particularly 0,0) has been known for some years (Long et al. 1995; Long and Rosenberg 2000). As frequencies of phase 1 and 0 introns change significantly along the genes, one would expect that this will also affect the frequency of symmetric exons. Figure 2 supports this view. In both *H. sapiens* and *A. thaliana* the occurrence of 0,0 exons increases dramatically and that of 1,1 exons declines in the 3′ direction. Spearman rank correlations are highly significant in both *H. sapiens* (r_s = 0.98, t = 18.0, p < 0.000001) and *A. thaliana* (r_s = 0.93, t = 7.01, p < 0.000001). The frequencies of other exon types are not significantly affected. In *C. elegans* and *D. melanogaster* the differences are marginal.

Discussion

Our results demonstrate that phase 1 introns are more frequent in the 5′ region of genes, and their occurrence declines in the 3′ direction. Conversely, phase 0 introns are less frequent in the 5′ region but increase in the 3′ direction. The gradient in the distribution of different phase introns provides a likely explanation for the observed frequencies of asymmetric and symmetric exons. Possible factors causing such gradients are discussed below. Tordai and Patthy (2004) observed that in humans a high frequency of phase 1 introns at the 5′ end of the CDS is typical for the genes coding proteins with a signal peptide. This could be a contributing factor, however, the length of the N-terminal portion of the protein possessing a signal peptide is limited (~20–40 amino acids) and the corresponding section of a gene is also relatively short. It is unlikely that this explanation alone can provide a sufficient reason for the lengthy gradient we reported here. Another observation suggests that phase 1 introns are more frequent at the ends of genes because modern domains that frame proteins are frequently bounded by these types of introns (Vibranovski et al. 2005). Our data do not indicate that phase 1 is more frequent at the 3′ end of genes.

It is now commonly accepted that losses and gains of introns were an essential characteristic of gene evolution. There are, however, some differences in opinions. According to Roy and Gilbert (2005), in *D. melanogaster* and *C. elegans* intron losses are more prominent than gains, in *H. sapiens* there is a balance, and in *A. thaliana* intron gains are dominant events. Rogozin et al. (2005) summarized evidence that intron gains were the more common process in the compared species. Qui et al. (2004) expressed a similar view. Most recently Nguyen et al. (2005) concluded that *H. sapiens* and *A. thaliana* experienced many more gains than losses, but in *D. melanogaster* and *C. elegans* losses outweighed gains. It seems, however, that there is a consensus that *A. thaliana* and *H. sapiens* (by the majority of votes) have more gains and *D. melanogaster* has more losses. *C. elegans* holds a middle ground. This conclusion is somewhat relevant to the following discussion.

Recently published data on the distribution of introns are a useful source of information and might be helpful for understanding the nature of the observed gradient. Sverdlov et al. (2004) detected an excess of old introns in the 5′ regions of genes in all genomes described in this paper. In contrast, introns located at the 3′ end of genes are a mixed picture in different species. In genomes such as *H. sapiens* and *A. thaliana* there is a greater number of new introns located in the 3′ region.

Table 5, compiled and calculated using the original data presented in Table 1 of Sverdlov et al. (2004), demonstrates that the difference in the distribution of old and new introns in the opposite regions of genes is highly significant (Hs, χ^2 = 13.37, p = 0.0003; At, χ^2 = 16.42, p = 0.0001).

In *C. elegans* and *D. melanogaster* there are no large differences in the distribution of old and new introns in the 3′ and 5′ regions (Ce, $\chi^2 = 3.97$, $p = 0.046$; Dm, $\chi^2 = 0.41$, $p = 0.52$; data from Sverdlov at al. [2004, Table 1] were used for these calculations).

Coglan and Wolfe (2004) also reported similar results, indicating that there are no significant differences in the distribution of novel and old introns in *C. elegans* and *C. briggsae*. Nevertheless, these authors found that novel introns show a preference for phase 0 and also for the canonic splice site sequence AG↓G. Furthermore, Rogozin et al. (2003) have shown that the excess of phase 0 introns is greater in novel introns than in more ancient ones. Results published by Qui et al. (2004) support this argument.

Taken together these facts provide a reasonable explanation for our results. Namely, the gradient in distribution of phase 0 introns in *H. sapiens* and *A. thaliana* might be caused by preferential insertions of novel introns in the 3′ regions of the genes and such introns are biased toward phase 0. The opposite gradient of phase 1 introns might, at least in part, have a compensatory nature. Gains of novel introns rather than losses of old introns dominated the evolution of *H. sapiens* and *A. thaliana*. In *C. elegans* and *D. melanogaster* the disparity between gains and losses of introns is not great and the distribution of novel introns is rather uniform (Sverdlov et al. 2004; Coglan and Wolfe 2004). As a consequence, the gradients of phase 0 and 1 introns are weaker and their frequencies remain similar in genes with different numbers of introns (Fig. 1). In *H. sapiens* and *A. thaliana*, on the other hand, the differences between the 5′ and the 3′ portions of genes are obvious and the frequencies of phase 0 and 1 introns behave distinctly. However, we cannot rule out the possibility that the biased distribution of modern protein domains that correspond to exons flanked by phase 1 introns (Vibranovski et al. 2005) might contribute to the gradient reported in this paper.

One can speculate that in the *H. sapiens* and *A. thaliana* genomes, two opposite regions of genes behaved a bit differently. The ancient introns saturated with phase 1 introns more often are located at the 5′ end of genes. On the contrary, the novel introns enriched by phase 0 introns are located at the 3′ end of genes. It is possible that mammalian genomes as well as genomes of flowering plants followed this evolutionary pathway. This may in part explain the surprising similarities between some genome features of these very distinct branches of eukaryotes, a fact emphasized in earlier publications (Rogozin et al. 2003). What could be the cause of the preferential insertions of novel introns in phase 0 in the 3′ region of such genomes? There are at least a couple of options: either a higher frequency of initial intron insertions in 0 phase or a higher retention of 0 phase introns due to positive selection. Further analysis may resolve this matter.

Acknowledgments. The authors are grateful to J.-V. Chamary for the purging of the database used in this study and to L. Hurst for the useful discussion.

References

Coghlan A, Wolfe KH (2004) Origins of recently gained introns in Caenorhabditis. Proc Natl Acad Sci USA 101:11362–11367

Endo T, Fedorov A, de Souza SJ, Gilbert W (2002) Do introns favor or avoid regions of amino acid conservation? Mol Biol Evol 19:521–525

Fedorov A, Suboch G, Bujakov M, Fedorova L (1992) Analysis of nonuniformity in intron phase distribution. Nucleic Acids Res 20:2553–2557

Gilbert W, de Souza SJ, Long M (1997) Origin of genes. Proc Natl Acad Sci USA 94:7698–7703

Kaessmann H, Zöllner S, Nekrutenko A, Li W-H (2002) Signatures of domain shuffling in the human genome. Genome Res 12:1642–1650

Lercher MJ, Chamary JV, Hurst LD (2004) Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. Genome Res 14:1002–1013

Long M, Deutsch M (1999) Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. Mol Biol Evol 16:1528–1534

Long M, Rosenberg C (2000) Testing the "proto-splice sites" model of intron origin:evidence from analysis of intron phase correlations. Mol Biol Evol 17:1789–1796

Long M, Rosenberg C, Gilbert W (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. Proc Natl Acad Sci USA 92:12495–12499

Lynch M (2002) Intron evolution as a population–genetic process. Proc Natl Acad Sci USA 99:6118–6123

Nguyen HD, Yoshihama M, Kenmochi N (2005) New maximum likelihood estimators for eukaryotic intron evolution. PloS Comput Biol 1:0631–0638 (www.ploscomp.biol.org)

Qui W-G, Schisler N, Stoltzfus A (2004) The evolutionary gain of spliceosomal introns:sequence and phase preferences. Mol Biol Evol 21:1252–1263

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BV, Koonin EV (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol 13:1512–1517

Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV (2005) Analysis of evolution of exon-intron structure of eukaryotic genes. Brief Bioinform 6:118–134

Roy SW, Gilbert W (2005) Rates of intron loss and gain:implications for early eukaryotic evolution. Proc Natl Acad Sci USA 102:5773–5778

Ruvinsky A, Eskesen ST, Eskesen FN, Hurst LD (2005) Can codon usage bias explain intron phase distributions and exon symmetry? J Mol Evol 60:99–104

Saxonov S, Daizadeh I, Fedorov A, Gilbert W (2000) EID: the Exon–Intron Database—an exhaustive database of protein-coding intron-containing genes. Nucleic Acids Res 28:185–190

Sverdlov AV, Babenko VN, Rogozin IB, Koonin EV (2004) Preferential loss and gain of introns in 3′ portion of genes suggests a reverse-transcription mechanism of intron insertion. Gene 338:84–91

Tordai H, Patthy L (2004) Insertion of spliceosomal introns in proto-splice sites: the case of secretory signal peptides. FEBS Lett 575:109–111

Vibranovski MD, Sakabe NJ, de Oliveira RS, de Souza SJ (2005) Signs of ancient and modern exon-shuffling are correlated to the distribution of ancient and modern domains along proteins. J Mol Evol 61:341–350