mRNA-Mediated Intron Losses: Evidence from Extraordinarily Large Exons

Deng-Ke Niu, Wen-Ru Hou, and Shu-Wei Li

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

Multicellular eukaryotes that have high intron density have their introns almost evenly distributed within genes, but unicellular eukaryotes that are generally intron poor have their introns asymmetrically distributed toward the 5′ ends of genes. This was explained by homologous recombination of genomic DNA with the cDNA reverse transcribed from the 3′ polyadenylated tail of spliced mRNA. This paper is to study whether mRNA-mediated intron losses have ever occurred in multicellular eukaryotes. If intron losses were mRNA-mediated, adjacent introns should be commonly lost together. A direct result is fusion of several previously adjacent exons and producing a large exon. We found that extraordinarily large exons (ELEs) are common not only in unicellular eukaryotes but also in multicellular eukaryotes. The percentage of genes having ELEs is negatively correlated with intron abundance. In addition, the number of lost introns estimated from the relative lengths of ELEs is negatively correlated with the number of extant introns. These results support mRNA-mediated intron losses in all eukaryotes. Moreover, we found that the ELEs of intron-common eukaryotes (with more than 0.5 intron per gene on average) are not only located at 3′ ends but also at 5′ ends and the middle of genes. This is contrary to what would be expected if the involved cDNAs were reverse transcribed from the 3′ polyadenosine ends. A remarkable difference in intron distribution was revealed between intron-rare eukaryotes and intron-common eukaryotes. The intron-rare eukaryotes show very strong 5′-biased intron distribution, whereas the intron-common eukaryotes display even intron distribution or only weak 5′-biased distribution. We suspected that intron losses from 3′ end of genes may be limited in intron-rare eukaryotes. The intron losses from intron-common eukaryotes should have other priming mechanism, like self-primed reverse transcription.

Introduction

Reverse transcriptase encoded by retrotransposon occasionally produces cDNA from cellular mRNA (Luan et al. 1993; Esnault, Maestre, and Heidmann 2000). If reverse transcriptases begin from the 3′ ends of mRNA molecules and dissociate in a length-dependent manner, most of the cDNAs would include 3′ ends of the coding sequences but very few would extend completely to 5′ ends. Homologous recombination between these cDNAs and the genomic sequences would thus preferentially remove introns at the 3′ ends of genes (fig. 1*A*). This model was initially proposed to explain the paucity of introns in *Saccharomyces cerevisiae* genes as well as the 5′-biased localization of introns in the rare genes that have retained them (Fink 1987). Later, the cDNAs produced by the reverse transcriptase of retrotransposon *Ty* were detected in *S. cerevisiae* to recombine with homologous chromosomal regions, resulting in intron loss. In addition, intronless pseudogenes with polyadenosine were produced, indicating that the reverse transcription was started from the 3′ end of the mRNA template (Derr, Strathern, and Garfinkel 1991; Derr and Strathern 1993).

A recent study indicated that mRNA-mediated intron losses from 3′ ends might have occurred in all unicellular eukaryotes (Mourier and Jeffares 2003). These organisms with low intron density display 5′-biased intron distributions within coding sequences.

By contrast, multicellular eukaryotes are commonly intron rich. Their introns are almost evenly distributed throughout the coding sequences (Sakurai et al. 2002; Mourier and Jeffares 2003). There are three hypotheses for multicellular eukaryotes. The first hypothesis is that very few mRNAs exist in the germ line cells of multicellular eukaryotes, so most genes have no chance to lose their introns by the mRNA-mediated mechanism. However, some evidence suggested that germ line cells may transcribe much more genes than required (Majewski 2003; Yanai, Graur, and Ophir 2004). The second hypothesis is, as Fink (1987) suggested, that cDNAs may be more likely to integrate into multicellular genomes by nonhomologous recombination and thus produce processed pseudogenes. The intron losses in multicellular eukaryotes revealed by phylogenetic studies may be simple genomic deletion (Banyai and Patthy 2004) or precise in-frame deletions involving nonhomologous recombination between short direct repeats in or near the 5′ and 3′ splicing sites (Robertson 1998). Apparently, most researchers do not think so. The intron losses from multicellular genomes revealed by phylogenetic analyses were often explained by the model of mRNA-mediated intron losses from 3′ ends (Drouin and Moniz de Sá 1997; Frugoli et al. 1998; Feiber, Rangarajan, and Vaughn 2002; Krzywinski and Besansky 2002; Cho et al. 2004; Roy and Gilbert 2005). The last hypothesis is that introns were lost not only from 3′ ends of genes. Assume a gene with five introns (*a*, *b*, *c*, *d*, and *e* from 5′ to 3′ end as shown in fig. 1) equally distributed. Loss of 5′-end introns (e.g., *a*, or *a* and *b*) or middle introns (e.g., *b*, *c*, and *d*) would not result in 5′-biased intron distribution. In fact, as revealed by phylogenetic analyses in multicellular eukaryotes, losses of middle introns or 5′-end introns are as common as, or maybe more common than, losses of introns at 3′ ends (Krzywinski and Besansky 2002; Roy, Fedorov, and Gilbert 2003; Banyai and Patthy 2004; Cho et al. 2004; Kiontke et al. 2004). Recent analysis on four fungal genomes also indicated that losses of middle introns are more frequent than losses of 3′-end introns (Nielsen et al. 2004). A modified version of Fink's model is that the reverse transcription is self-primed by the 3′ terminus of mRNA (or partially

Key words: extraordinarily large exon, intron loss, mRNA self-primed reverse transcription, homologous recombination, intron position.

E-mail: dkniu@bnu.edu.cn; dengkeniu@hotmail.com.

Mol. Biol. Evol. 22(6):1475–1481. 2005
doi:10.1093/molbev/msi138
Advance Access publication March 23, 2005

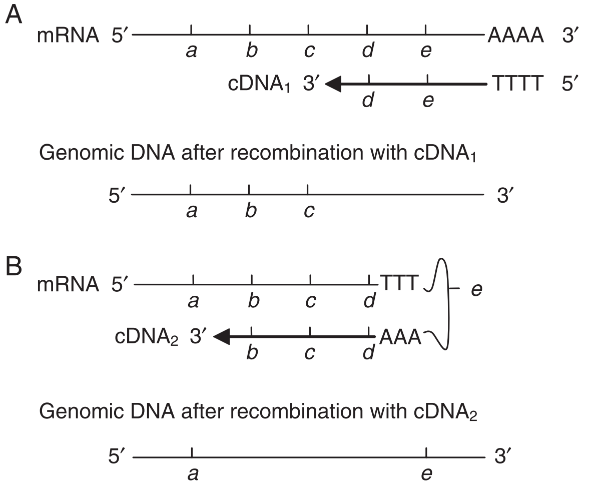

FIG. 1.—Schematic diagrams for mRNA-mediated intron loss. The small bars show the position of exon-exon junction sites in mRNA and cDNA (i.e., introns position in genomic DNA). (*A*) If reverse transcription is primed from the 3′ polyadenosine end of the mRNA, homologous recombination of cDNA with genomic DNA will preferentially remove introns at the 3′ end and produce a large exon at the 3′ end of the gene; (*B*) If reverse transcription is self-primed, the homologous recombination will remove introns at the 5′ end or 3′ end or in the middle of the gene depending on how long the mRNA 3′ terminus folds back. So large exon may be produced at any position. Here, only loss of middle introns and production of middle large exon are shown.

degraded mRNA) with stem-loop (Feiber, Rangarajan, and Vaughn 2002). Which introns are lost depends on how long the 3′ terminus folds back (fig. 1*B*). In this paper, we present evidence (1) for mRNA-mediated intron losses in multicellular eukaryotes as well as unicellular eukaryotes and (2) that the involved reverse transcriptions were unlikely primed from 3′ polyadenosine ends.

Besides the mRNA-mediated model, simple genomic deletion (Banyai and Patthy 2004) and in-frame intron deletion (Robertson 1998) were proposed to explain intron losses. A prediction of mRNA-mediated intron loss is that adjacent introns should be commonly lost together (Roy and Gilbert 2005). In mRNA-unrelated mechanisms, introns should be lost individually. A direct result of intron loss is exon fusion. Fusion of two adjacent exons may result from either mRNA-mediated intron loss (Fink 1987; Feiber, Rangarajan, and Vaughn 2002) or mRNA-unrelated intron deletion (Robertson 1998; Banyai and Patthy 2004), whereas fusion of three or more adjacent exons could only be attributed to mRNA-mediated intron losses (Fink 1987; Feiber, Rangarajan, and Vaughn 2002). So extraordinarily large exons (ELEs) are more likely to be produced by mRNA-mediated intron loss. In this study, we found that ELEs exist in all multicellular eukaryotes as well as in intron-common unicellular eukaryotes (*Aspergillus nidulans*, *Fusarium graminearum*, *Magnaporthe grisea*, *Neurospora crassa*, *Plasmodium falciparum*, *Schizosaccharomyces pombe,* and *Ustilago maydis*). Analyses of the lengths and abundance of ELEs support mRNA-mediated intron losses in multicellular eukaryotes and intron-common unicellular eukaryotes.

If intron losses were mediated by cDNAs reverse transcribed from 3′ polyadenosine ends as proposed by Fink (1987), the resulting large exons should exist at 3′ ends (fig. 1*A*). By contrast, if the cDNAs involved in intron losses were primed by other mechanisms (e.g., self-primed; Feiber, Rangarajan, and Vaughn 2002), the resulting large exons may exist at 5′ ends, middle parts, or 3′ ends depending on how long the 3′ terminus folded back (fig. 1*B*). We found that 5′-end ELEs and middle ELEs are as frequent as 3′-end ELEs in multicellular eukaryotes. Surprisingly, similar facts were also observed in the intron-common unicellular eukaryotes we surveyed. These results indicated that the mRNA-mediated intron losses in multicellular eukaryotes as well as intron-common unicellular eukaryotes might have priming mechanisms (Feiber, Rangarajan, and Vaughn 2002; Nielsen et al. 2004) other than primed from 3′ polyadenosine ends.

Materials and Methods

The present analysis covered 23 eukaryotes. The genomes of *Encephalitozoon cuniculi*, *Guillardia theta*, *Eremothecium gossypii*, *S. cerevisiae*, *S. pombe*, *Candida glabrata*, *Kluyveromyces lactis*, *Debaryomyces hansenii*, *Yarrowia lipolytica*, *P. falciparum*, *Anopheles gambiae*, *Drosophila melanogaster*, *Apis mellifera*, *Arabidopsis thaliana*, *Caenorhabditis elegans* (WS97), *Gallus gallus* (NCBI build 1 version 1), *Mus musculus* (NCBI build 32), *Rattus norvegicus* (NCBI build 2), *Pan troglodytes* (NCBI build 1 version 1), and *Homo sapiens* (NCBI build 34 version 3) were downloaded from the National Center for Biotechnology Information (NCBI) GenBank database (ftp://ftp.ncbi.nih.gov). The genome annotation files of *A. nidulans* (Data Version March 7, 2003; *Aspergillus* Sequencing Project), *F. graminearum* (Data Version March 11, 2003; *F. graminearum* Sequencing Project), *M. grisea* (Release 2.3; *Magnaporthe* Sequencing Project), *N. crassa* (Assembly Version 3; Galagan et al. 2003), and *U. maydis* (Data Version April 1, 2004; *U. maydis* Sequencing Project) were downloaded from Broad Institute of MIT and Harvard (http://www.broad.mit.edu/). Genes with alternative splicing sites, with obvious annotation errors, or overlapping with other genes were excluded from our analyses. So the values of intron paucity and relative intron position we calculated from some genomes may be slightly different from previous results (Mourier and Jeffares 2003).

Both mRNA-mediated intron loss and mRNA-unassociated intron deletion may lead to fusion of two adjacent exons, whereas only mRNA-mediated intron loss may result in the fusion of three or more adjacent exons. Thus, ELEs are more likely to be produced by mRNA-mediated intron loss. For assurance, only exons longer than six times the median of all exons in the same gene were selected as ELEs. An ELE in this study is an exon that should be much longer than expected from random variation. If a gene has both ELEs and extraordinarily small exons, the ELEs may come from random variations. So genes with both large exons six times longer than the median and small exons three times shorter than the median were excluded. For example, in *S. pombe*, gene *rec8* (with five exons having 56, 188, 97, 1,250, and 95 bp) was selected as having one ELE, but gene *sec61* (with six exons having 10, 68, 113, 33, 134, and 1,082 bp) was not believed to have any ELE. The exon length histograms of genes having

Table 1
The Abundance and Position of ELEs

	Number of Genes Studied[a]	Percentage of Genes Having ELEs (%)	Correlation Between Intron Number and Intron Losses (Spearman's Rho)[b]	ELE Position		
				5′ End (%)	Middle (%)	3′ End (%)
Yarrowia lipolytica	60	38.3	—	0	4.3	95.7
Plasmodium falciparum	1,475	27.1	−0.177**	43.1	31.7	25.2
Schizosaccharomyces pombe	1,151	18.4	—	5.6	19.1	75.3
Ustilago maydis	1,172	14.2	−0.200**	26.0	20.2	53.8
Fusarium graminearum	6,364	11.2	−0.271***	13.3	46.4	40.3
Neurospora crassa	4,484	11.7	−0.259***	11.1	46.7	42.2
Magnaporthe grisea	5,391	10.4	−0.169***	12.2	47.0	40.8
Aspergillus nidulans	6,033	10.1	−0.268***	12.8	46.8	40.4
Drosophila melanogaster	5,471	9.05	−0.288***	13.8	58.3	27.8
Arabidopsis thaliana	14,998	7.43	−0.210***	32.5	42.4	25.2
Anopheles gambiae	916	7.42	−0.224*	10.7	69.3	20.0
Apis mellifera	5,253	4.64	−0.281***	9.1	67.4	23.6
Caenorhabditis elegans	15,261	3.04	−0.340***	5.8	78.6	15.7
Gallus gallus	14,841	6.68	−0.069**	15.0	57.0	28.0
Rattus norvegicus	15,232	6.91	−0.141***	16.5	51.4	32.1
Mus musculus	14,379	6.13	−0.177***	17.7	45.8	36.5
Pan troglodytes	17,394	5.87	−0.185***	13.1	54.7	32.1
Homo sapiens	14,239	6.88	−0.185***	18.0	43.6	38.4
Human GL genes	6,728	6.73	−0.102**	20.9	49.2	30.0
Human OS genes	684	8.19	—	12.5	28.1	59.4
Mouse GL genes	3,551	6.20	−0.155**	19.0	44.8	36.2
Mouse OS genes	501	4.99	—	14.8	51.9	33.3

[a] Only genes with three or more exons were counted.

[b] $*0.05 \leq P < 0.1$; $**10^{-4} \leq P < 0.05$; $***P < 10^{-4}$, Spearman's rho is not shown if the correlation is not even marginally significant.

ELEs are right skewed or have noticeable outliers on the right side (fig. S1 of Supplementary Material online), indicating that the exon lengths do not have normal distribution. Some exons are much longer than expected from random variation. If ELEs were produced by fusion of previous adjacent exons, the number of introns lost from a gene can be approximately estimated by the relative lengths of ELEs:

$$\sum_{i=1}^{n}(Ei/M-1),$$

where n is the number of ELEs in a gene, Ei is the length of ELE i, and M is the median of the lengths of all the exons in a gene. Analyses of ELEs and the number of lost introns involved only genes with three or more exons (thus only genes with two or more introns).

We used GNF GeneAtlas Version 2 (Su et al. 2004) in determining whether a gene is expressed in germ line. We chose this database because it showed excellent reproducibility (Huminiecki, Lloyd, and Wolfe 2003). A gene was identified to be expressed in a tissue if it has an average difference value greater than 200 in that tissue, while a gene was identified to be not expressed in a tissue if it has an average difference value lower than 100 in that tissue. All the genes expressed in oocyte or fertilized egg of mouse samples were categorized as germ line–expressed (GL) genes in mouse. In the same way, all the genes expressed in testis germ cells of human samples were categorized as human GL genes. Genes expressed only in the soma (OS) of mouse were defined as the genes that are expressed in any somatic tissue but not expressed in germ line cells (oocyte or fertilized egg), reproductive organs (testis or ovary), or early developmental stages (blastocysts, embryo day 6.5, embryo day 7.5, embryo day 8.5, embryo day 9.5, or embryo day 10.5). Similarly, human OS genes were defined as the genes that were expressed in any somatic tissue but not expressed in testis germ cells, testis, or ovary.

The relative intron positions were measured and presented by the methods described in previous studies (Sakurai et al. 2002; Mourier and Jeffares 2003).

Results and Discussion

Extraordinarily Large Exons

We found that ELEs exist in all the eukaryotes we surveyed including multicellular eukaryotes (table 1). Similar to the relative intron positions previously reported

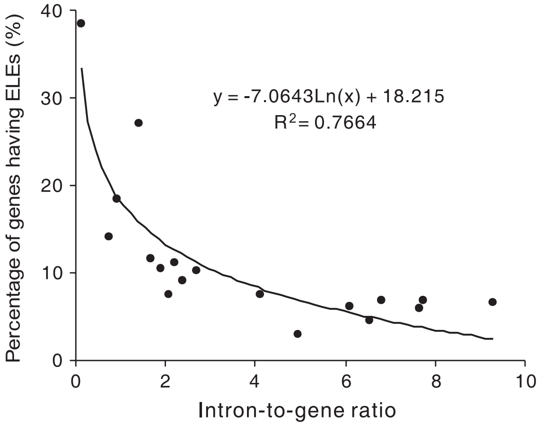

FIG. 2.—Significant negative correlation between intron abundance and percentage of genes having ELEs in eukaryotic genomes.

(Mourier and Jeffares 2003), there is a significant negative correlation between the percentage of genes having ELEs and the intron-to-gene ratio (Spearman's rho = -0.868, $P < 10^{-5}$; fig. 2). Genomes with more ELEs have fewer introns. Our stringent criteria in selecting ELEs may underestimate the frequency of mRNA-mediated intron loss but ensured to exclude the mRNA-unrelated intron deletions (Robertson 1998; Banyai and Patthy 2004).

Negative Correlation Between Intron Number and Estimated Number of Lost Introns

Because ELEs are likely to be produced by fusion of previous adjacent exons, we roughly estimated the number of introns lost by the relative length of ELEs. As shown in table 1, the estimated number of lost introns is negatively correlated with the number of extant introns in all the species except *S. pombe* and *A. gambiae*. Even in *A. gambiae*, the correlation is marginally significant (Spearman's rho = -0.224, $P = 0.067$). If we adopt a less stringent criterion in selecting ELEs (e.g., excluding genes with small exons six times shorter than the median of all the exons of a gene in selecting ELEs), significant negative correlation was found in *S. pombe* (Spearman's rho = -0.220, $P = 0.00018$). All the 23 genes with ELEs have three exons, so no result on the correlation between the number of extant introns and the number of lost introns was obtained in *Y. lipolytica*.

Intron gain in evolution has also been well documented (Cho and Doolittle 1997; Kiontke et al. 2004; Nielsen et al. 2004; Qiu, Schisler, and Stoltzfus 2004; Sadusky, Newman, and Dibb 2004; Sverdlov et al. 2004). Some results showed that introns were inserted into coding sequences randomly (Cho and Doolittle 1997), while some other evidence suggested that introns may be preferentially accumulated in the 3′ side of genes (Sverdlov et al. 2004). If introns were frequently inserted into coding sequences, the exons would become smaller. And if one exon was exceptionally selected against intron insertion, it should become a comparatively large exon and even an ELE. But as we know, there is no evidence for selection against intron insertion into a specific coding region. Certainly, this possibility cannot be completely excluded. Apparently, intron losses are more likely the cause for the existence of ELEs.

Genes with fewer introns can be explained by more introns lost. By contrast with a previous study based on intron position (Mourier and Jeffares 2003), we found that *P. falciparum* is similar to other unicellular eukaryotes if ELEs and estimated number of lost introns are considered (table 1). In addition, our results suggested that mRNA-mediated intron losses happened not only in unicellular eukaryotes but also in multicellular eukaryotes. This consisted of previous phylogenetic analyses on specific genes (Drouin and Moniz de Sá 1997; Frugoli et al. 1998; Feiber, Rangarajan, and Vaughn 2002; Krzywinski and Besansky 2002; Cho et al. 2004). While this paper was under review, losses of adjacent introns were reported in some multicellular eukaryotes (Roy and Gilbert 2005).

Some introns contain functional elements (Croft et al. 2000; Hare and Palumbi 2003; Kolb 2003), and natural selection would operate against their deletions. Many unrecognized alternatively spliced exons exist in sequences annotated as introns (Croft et al. 2000). The differences in intron losses we observed may be explained by the fact that multicellular eukaryotes have more frequent alternative splicing of introns than unicellular eukaryotes (Ast 2004).

Position of ELEs Within Genes

Some unicellular eukaryotes (*E. cuniculi*, *C. glabrata*, *G. theta*, *K. lactis*, *E. gossypii*, *S. cerevisiae,* and *D. hansenii*) have very few, or even no, genes with three or more exons. Thus, statistically enough ELEs were not obtained from these organisms by our methods. The 5′-end ELEs and middle ELEs were observed in other eukaryotes except that *Y. lipolytica* has no 5′-end ELE (table 1). In most cases, middle ELEs are more abundant than 3′-end ELEs, and 3′-end ELEs are more abundant than 5′-end ELEs. It seems that mRNA-mediated losses of 5′-end introns and middle introns were as common as that of 3′-end introns. Recent phylogenetic analysis of four filamentous fungal genomes (*A. nidulans*, *F. graminearum*, *M. grisea,* and *N. crassa*) revealed more intron losses from the middle of genes than from 3′ ends (Nielsen et al. 2004). These are obviously contrary to what would be expected if intron loss primarily involved homologous recombination of polyadenosine-primed reverse transcripts. A likely mechanism is that the involved cDNAs were produced by self-primed reverse transcription (Feiber, Rangarajan, and Vaughn 2002).

In genes with four or more exons, middle ELEs are a collection of ELEs at any position except two ends, so more middle ELEs do not necessarily mean middle introns were lost more frequently. Instead, there might be a decline in the frequency of intron losses from 3′ end to 5′ end. A recent analysis of 684 groups of orthologous genes revealed that introns closer to 3′ ends of genes are lost more frequently (Roy and Gilbert 2005). The first introns are more conserved than other introns (Keightley and Gaffney 2003). Numerous reports described functional elements in first introns (Chan et al. 1999; Majewski and Ott 2002; Kolb 2003). The selective constraint on first introns may partially explain the relatively lower frequency of intron loss at the 5′ end of genes. A noticeable exception is *P. falciparum*, which has more 5′-end ELEs than 3′-end ELEs. A similar fact has also been revealed by relative intron position (Mourier and Jeffares

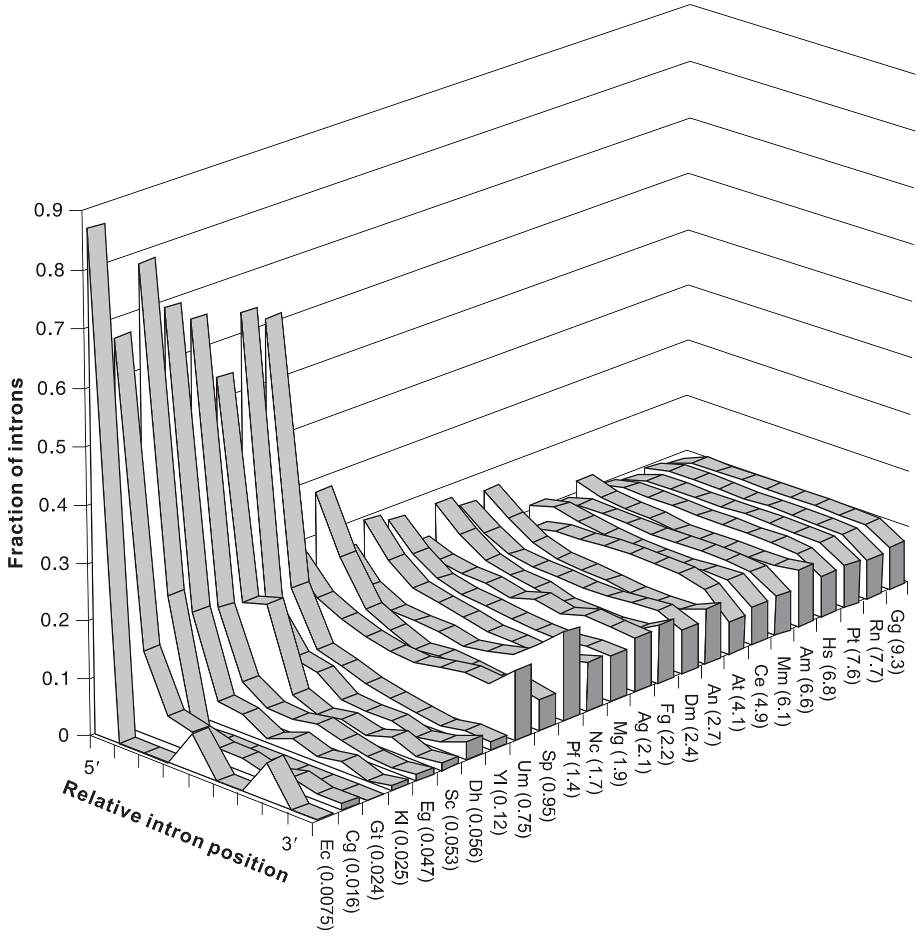

FIG. 3.—The relative intron positions. Species are sorted according to their intron-to-gene ratios (in parentheses). Species abbreviations: Gg, *Gallus gallus*; Rn, *Rattus norvegicus*; Pt, *Pan troglodytes*; Hs, *Homo sapiens*; Am, *Apis mellifera*; Mm, *Mus musculus*; Ce, *Caenorhabditis elegans*; At, *Arabidopsis thaliana*; An, *Aspergillus nidulans*; Dm, *Drosophila melanogaster*; Fg, *Fusarium graminearum*; Ag, *Anopheles gambiae*; Mg, *Magnaporthe grisea*; Nc, *Neurospora crassa*; Pf, *Plasmodium falciparum*; Sp, *Schizosaccharomyces pombe*; Um, *Ustilago maydis*; Yl, *Yarrowia lipolytica*; Dh, *Debaryomyces hansenii*; Sc, *Saccharomyces cerevisiae*; Eg, *Eremothecium gossypii*; Kl, *Kluyveromyces lactis*; Gt, *Guillardia theta*; Cg, *Candida glabrata*; Ec, *Encephalitozoon cuniculi*.

2003; fig. 3). Although the artifact of gene prediction for *Plasmodium* genomes cannot be excluded (Mourier and Jeffares 2003), *P. falciparum* is more likely to have frequent intron losses from both 3′ ends and 5′ ends of genes.

Although the model of mRNA-mediated intron losses was demonstrated by experiments in *S. cerevisiae* (Derr, Strathern, and Garfinkel 1991; Derr and Strathern 1993), no convincing evidence was obtained for other eukaryotes to lose introns by the same mechanism. Reverse transcription and the consequent generation of processed pseudogenes were well demonstrated in metazoans (Luan et al. 1993; Jurka 1997; Esnault, Maestre, and Heidmann 2000). But in the mechanisms that produce processed pseudogenes in insects and mammals, mRNA reverse transcription and cDNA integrating into the genome are tightly coupled by using the nick at the chromosomal target site of integration as the primer of reverse transcription (Luan et al. 1993; Jurka 1997; Ostertag and Kazazian 2001). So diffusible cDNAs for homologous recombination would not be produced. If mRNA-mediated intron losses have really occurred in the evolution of metazoans, the cDNAs should arise from a mechanism different from that producing pseudogenes. That is, the reverse transcription should have priming mechanisms other than that producing pseudogenes, thus did not necessarily start from the 3′ polyadenosine end of the template mRNA.

A direct evidence of self-primed reverse transcription–mediated intron loss is the reverse complement between the mRNA 3′ terminus and exon sequence as revealed in polymorphic alleles of *Drosophila 4f-rnp* genes (Feiber, Rangarajan, and Vaughn 2002). But, except that the stem-loops have some functions, constantly occurring nucleotide substitution, insertion, and deletion after intron loss would make the past intramolecular reverse complement undetectable.

Germ Line Expression and Intron Loss

In multicellular eukaryotes, only genes having transcripts in germ cells would be susceptible to mRNA-mediated intron loss (Krzywinski and Besansky 2002). We used the recent version of an excellent database of human and mouse gene expression, GNF GeneAtlas Version 2 (Su et al. 2004), to determine whether a gene is expressed in germ line. Significant negative correlations were found between the number of extant introns and the number of lost introns estimated from the relative length of ELEs in GL genes but not in

OS genes (table 1). Unexpectedly, OS genes have comparable ELEs (and thus comparable estimated intron losses) to GL genes (table 1).

Cell specialization of multicellular organisms was hypothesized to have some advantage by protecting genes that are not transcribed in germ line from transcription-associated damages (Niu and Chen 1997). If transcription levels and locations were fixed by positive Darwinian selection, genes are transcribed only when their functions are required to be expressed. Our selection of OS genes seems to be stringent enough. But if most changes in gene transcription were fixed by random drift without functional significance as suggested by recent studies (Khaitovich et al. 2004; Yanai, Graur, and Ophir 2004), many genes may be ectopically expressed (Chelly et al. 1989; Sarkar and Sommer 1989; Kimoto 1998). By analyzing the strand compositional asymmetries, Majewski (2003) suggested that 71%–91% of all human genes may be transcribed in germ line. Thus, the reliable criterion in selecting OS genes is that the genes should be transcribed in somatic tissues but not in any germ cells or any embryonic stages that have some cells doomed to develop into germ cell. Obviously, limited by data availability, we did not exclude enough genes. Especially for humans, only genes transcribed in testis germ cells, testis, and ovary were excluded. If the OS genes we selected are actually ectopically transcribed in some stage of germ line, it is reasonable for them to have comparable ELEs and lost introns. In addition, if ectopic transcriptions are common (Chelly et al. 1989; Sarkar and Sommer 1989; Kimoto 1998; Khaitovich et al. 2004; Yanai, Graur, and Ophir 2004), for most genes, OS and GL should not be conserved in evolution. An OS gene may have comparable ELEs and mRNA-mediated intron losses because it was previously expressed in germ line. Equally, a GL gene may have no ELEs and mRNA-mediated intron losses simply because it is previously OS.

Intron-Rare Eukaryotes and Intron-Common Eukaryotes

The mRNA-mediated intron loss from 3′ end was initially proposed to explain the paucity of introns in *S. cerevisiae* genes and the 5′-biased distribution of introns in the rare genes that have retained them (Fink 1987). Later, this hypothesis gained experimental supports in *S. cerevisiae* (Derr, Strathern, and Garfinkel 1991; Derr and Strathern 1993). Other unicellular eukaryotes that are usually intron poor show 5′-biased intron distribution within genes and thus were believed to have lost some introns by the same way as *S. cerevisiae* (Mourier and Jeffares 2003). A recent phylogenetic analysis of intron-poor fungi showed that intron losses from the middle of genes are common (Nielsen et al. 2004), suggesting other priming mechanisms in reverse transcription. Our survey of the ELEs indicated that mRNA-mediated losses of 5′-end introns and middle introns may be widespread in intron-common unicellular eukaryotes. Yeast *S. cerevisiae* was not included in our survey of ELEs because of methodological reasons. The most intron-poor eukaryote we analyzed is *Y. lipolytica*. It has no 5′-end ELE, very few middle ELEs, and up to 95.7% 3′-end ELEs (table 1). The intron losses in *Y. lipolytica* seem to follow the model of Fink (1987). As shown in figure 3, *Y. lipolytica* is a marginal species in intron distributions. There is a remarkable difference in intron distribution between intron-rare eukaryotes (*E. cuniculi*, *C. glabrata*, *G. theta*, *K. lactis*, *E. gossypii*, *S. cerevisiae*, *D. hansenii*, and *Y. lipolytica*) and intron-common eukaryotes (other organisms with intron-to-gene ratio more than 0.5). The intron-rare eukaryotes including *S. cerevisiae* show distinctively strong 5′-biased intron distribution, whereas in intron-common eukaryotes the introns are evenly distributed within genes or show only weak 5′-biased distribution. The differences between multicellular eukaryotes and unicellular eukaryotes are not so clear-cut. For example, *A. nidulans* has more introns than, and a similar intron distribution as, multicellular eukaryote *D. melanogaster* (fig. 3). We suspected that cDNAs reverse transcribed from 3′ end of the mRNA (Fink 1987) may have led to the intron losses in intron-rare eukaryotes while cDNAs from other priming mechanisms, like self-primed reverse transcription (Feiber, Rangarajan, and Vaughn 2002), may have caused the intron losses in intron-common eukaryotes.

Supplementary Material

Supplementary figure S1 is available at *Molecular Biology and Evolution* online (www.mbe.oupjournals.org).

Acknowledgments

We thank Kui Lin and Da-Yong Zhang for discussion, an anonymous referee for comments, and the GeneAtlas research group of the Genomics Institute of the Novartis Research Foundation for providing gene expression database and suggestion. The GenBank accession numbers for this research are: *Encephalitozoon cuniculi* (Accession NC_003229 to NC_003238, NC003242), *Guillardia theta* (Accession AF 083031, AF165818 and AJ 010592), *Eremothecium gossypli* (Accession NC_005782 to NC_005788), *S. cerevisiae* (Accession NC_001133 to NC_001148), *S. pombe* (Accession NC_003421, NC_003423 and NC_003424), *Candida glabrata* (Accession CR380947 to CR380959), *Kluyveromyces lactis* (Accession CR382121 to CR 382126), *Debaryomyces hansenii* (Accession CR382133 to CR382139), *Yarrowia lipolytica* (Accession CR382127 to CR382132), *P. falciparum* (Accession to NC_000521, NC_000910, NC_004314 to NC_0043118, NC_004325 to NC_004331), *Anopheles gambiae* (Accession AAAB01000001 to AAAB01008987), *Drosophila melanogaster* (Accession NC_004353, NC_004354, NT_033777 to NT_033779, NT_037436), *Apis mellifera* (NCBI build 1 version 1), *Arabidopsis thaliana* (Accession NC_003070, NC_003071, NC_003074 to 003076), *Caenorhabditis elegants* (WS97), *Gallus gallus* (NCBI build 1 version 1), *Mus musculus* (NCBI build 32), *Rattus norvegicus* (NCBI build2), *Pan troglodytes* (NCBI build 1 version 1) and *Homo sapiens* (NCBI build 34 version 3). This work was supported by National Natural Science Foundation of China (grant no. 30270695) and Beijing Normal University.

Literature Cited

Ast, G. 2004. How did alternative splicing evolve? Nat. Rev. Genet. **5**:773–782.

Banyai, L., and L. Patthy. 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.

Chan, R. Y. Y., C. Boudreau-Lariviere, L. M. Angus, F. A. Mankal, and B. J. Jasmin. 1999. An intronic enhancer containing an N-box motif is required for synapse- and tissue-specific expression of the acetylcholinesterase gene in skeletal muscle fibers. Proc. Natl. Acad. Sci. USA **96**:4627–4632.

Chelly, J., J. P. Concordet, J. C. Kaplan, and A. Kahn. 1989. Illegitimate transcription: transcription of any gene in any cell type. Proc. Natl. Acad. Sci. USA **86**:2617–2621.

Cho, G., and R. F. Doolittle. 1997. Intron distribution in ancient paralogs supports random insertion and not random loss. J. Mol. Evol. **44**:573–584.

Cho, S., S.-W. Jin, A. Cohen, and R. E. Ellis. 2004. A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. Genome Res. **14**:1207–1220.

Croft, L., S. Schandorff, F. Clark, K. Burrage, P. Arctander, and J. Mattick. 2000. ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. Nat. Genet. **24**:340–341.

Derr, L. K., and J. N. Strathern. 1993. A role for reverse transcripts in gene conversion. Nature **361**:170–173.

Derr, L. K., J. N. Strathern, and D. J. Garfinkel. 1991. RNA-mediated recombination in *S. cerevisiae*. Cell **67**:355–364.

Drouin, G., and M. Moniz de Sá. 1997. Loss of introns in the pollen-specific actin gene subfamily members of potato and tomato. J. Mol. Evol. **45**:509–513.

Esnault, C., J. Maestre, and T. Heidmann. 2000. Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. **24**:363–367.

Feiber, A. L., J. Rangarajan, and J. C. Vaughn. 2002. The evolution of single-copy *Drosophila* nuclear *4f-rnp* genes: spliceosomal intron losses create polymorphic alleles. J. Mol. Evol. **55**:401–413.

Fink, G. R. 1987. Pseudogenes in yeast? Cell **49**:5–6.

Frugoli, J. A., M. A. McPeek, T. L. Thomas, and C. R. McClung. 1998. Intron loss and gain during evolution of the catalase gene family in angiosperms. Genetics **149**:355–365.

Galagan, J. E., S. E. Calvo, K. A. Borkovich et al. (77 co-authors). 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature **422**:859.

Hare, M. P., and S. R. Palumbi. 2003. High intron sequence conservation across three mammalian orders suggests functional constraints. Mol. Biol. Evol. **20**:969–978.

Huminiecki, L., A. T. Lloyd, and K. H. Wolfe. 2003. Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. BMC Genomics **4**:31.

Jurka, J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. USA **94**:1872–1877.

Keightley, P. D., and D. J. Gaffney. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. Proc. Natl. Acad. Sci. USA **100**:13402–13406.

Khaitovich, P., G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, and S. Paabo. 2004. A neutral model of transcriptome evolution. PLoS Biol. **2**:682–689.

Kimoto, Y. 1998. A single human cell expresses all messenger ribonucleic acids: the arrow of time in a cell. Mol. Gen. Genet. **258**:233–239.

Kiontke, K., N. P. Gavin, Y. Raynes, C. Roehrig, F. Piano, and D. H. A. Fitch. 2004. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proc. Natl. Acad. Sci. USA **101**:9003–9008.

Kolb, A. 2003. The first intron of the murine beta-casein gene contains a functional promoter. Biochem. Biophys. Res. Commun. **306**:1099–1105.

Krzywinski, J., and N. J. Besansky. 2002. Frequent intron loss in the white gene: a cautionary tale for phylogeneticists. Mol. Biol. Evol. **19**:362–366.

Luan, D. D., M. H. Korman, J. L. Jakubczak, and T. H. Eickbush. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell **72**:595–605.

Majewski, J. 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. Am. J. Hum. Genet. **73**:688–692.

Majewski, J., and J. Ott. 2002. Distribution and characterization of regulatory elements in the human genome. Genome Res. **12**:1827–1836.

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

Nielsen, C. B., B. Friedman, B. Birren, C. B. Burge, and J. E. Galagan. 2004. Patterns of intron gain and loss in fungi. PLoS Biol. **2**:e422.

Niu, D. K., and J. K. Chen. 1997. Evolutionary advantages of cell specialization: save and protect DNA. J. Theor. Biol. **187**:39–43.

Ostertag, E. M., and H. H. Kazazian Jr. 2001. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res. **11**:2059–2065.

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

Robertson, H. M. 1998. Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. Genome Res. **8**:449–463.

Roy, S. W., A. Fedorov, and W. Gilbert. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc. Natl. Acad. Sci. USA **100**:7158–7162.

Roy, S. W., and W. Gilbert. 2005. The pattern of intron loss. Proc. Natl. Acad. Sci. USA **103**:713–718.

Sadusky, T., A. J. Newman, and N. J. Dibb. 2004. Exon junction sequences as cryptic splice sites: implications for intron origin. Curr. Biol. **14**:505–509.

Sakurai, A., S. Fujimori, H. Kochiwa, S. Kitamura-Abe, T. Washio, R. Saito, P. Carninci, Y. Hayashizaki, and M. Tomita. 2002. On biased distribution of introns in various eukaryotes. Gene **300**:89–95.

Sarkar, G., and S. S. Sommer. 1989. Access to a messenger RNA sequence or its protein product is not limited by tissue or species specificity. Science **244**:331–334.

Su, A. I., T. Wiltshire, S. Batalov et al. (13 co-authors). 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA **101**:6062–6067.

Sverdlov, A. V., V. N. Babenko, I. B. Rogozin, and E. V. Koonin. 2004. Preferential loss and gain of introns in 3′ portions of genes suggests a reverse-transcription mechanism of intron insertion. Gene **338**:85–91.

Yanai, I., D. Graur, and R. Ophir. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS J. Integr. Biol. **8**:15–24.

Mark Ragan, Associate Editor

Accepted March 16, 2005

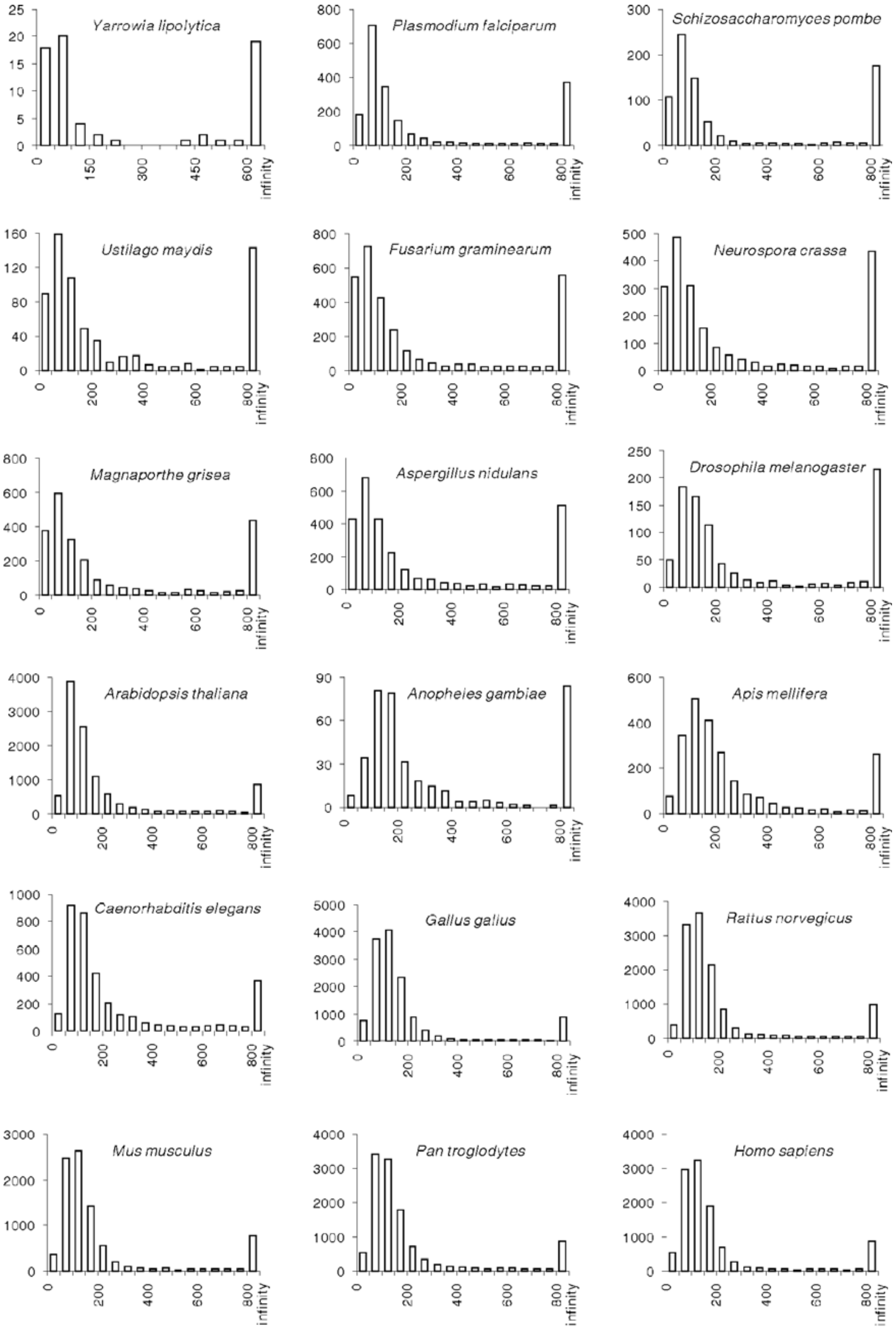

Figure S1. The exon length distribution of genes having extraordinarily large exons. The X-axes represent exon length (bp) while the Y-axes represent the number of exons.