Evolutionary Dynamics of Introns in Plastid-Derived Genes in Plants: Saturation Nearly Reached but Slow Intron Gain Continues

Malay Kumar Basu, Igor B. Rogozin,* Oliver Deusch,† Tal Dagan,† William Martin,† and Eugene V. Koonin**

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD; and
†Institute of Botany III, University of Düsseldorf, Düsseldorf, Germany

Some of the principal transitions in the evolution of eukaryotes are characterized by engulfment of prokaryotes by primitive eukaryotic cells. In particular, ~1.6 billion years ago, engulfment of a cyanobacterium that became the ancestor of chloroplasts and other plastids gave rise to Plantae, the major branch of eukaryotes comprised of glaucophytes, red algae, green algae, and green plants. After endosymbiosis, there was large-scale migration of genes from the endosymbiont to the nuclear genome of the host such that ~18% of the nuclear genes in *Arabidopsis* appear to be of chloroplast origin. To gain insights into the process of evolution of gene structure in these, originally, intronless genes, we compared the properties and the evolutionary dynamics of introns in genes of plastid origin and ancestral eukaryotic genes in *Arabidopsis*, poplar, and rice genomes. We found that intron densities in plastid-derived genes were slightly but significantly lower than those in ancestral eukaryotic genes. Although most of the introns in both categories of genes were conserved between monocots (rice) and dicots (*Arabidopsis* and poplar), lineage-specific intron gain was more pronounced in plastid-derived genes than in ancestral genes, whereas there was no significant difference in the intron loss rates between the 2 classes of genes. Thus, after the transfer to the nuclear genome, the plastid-derived genes have undergone a massive intron invasion that, by the time of the divergence of dicots and monocots (150–200 MYA), yielded intron densities only slightly lower than those in ancestral genes. Nevertheless, the accumulation of introns in plastid-derived genes appears not to have reached saturation and continues to this time, albeit at a low rate. The overall pattern of intron gain and loss in the plastid-derived genes is shaped by this continuing gain and the more general tendency for loss that is characteristic of the recent evolution of plant genes.

Introduction

Multiple spliceosomal introns interrupting protein-coding genes and the concurrent splicing machinery are among the defining features of eukaryotes (Doolittle 1978; Gilbert 1978; Mattick 1994; Deutsch and Long 1999). Origin and evolution of introns are often considered within the context of the debate between the introns-early and introns-late concepts, a conundrum that emerged shortly after the discovery of the exon–intron organization of eukaryotic genes (Darnell 1978; Doolittle 1978; Logsdon and Palmer 1994; Logsdon 1998). The introns-early hypothesis (Darnell 1978; Doolittle 1978; Gilbert 1978, 1987; Gilbert and Glynias 1993; Gilbert et al. 1997) (appearing also in the more recent guise of "introns first"[Jeffares et al. 2006]) holds that introns were an intrinsic element of the first protein-coding genes and actually facilitated the origin of proteins via recombination; under this view, the absence of spliceosomal introns in prokaryotes is attributed to complete loss of introns in the course of "genome streamlining." The introns-late hypothesis counters that prokaryotes never had spliceosomal introns and that the introns and the spliceosome emerged during eukaryotic evolution (Logsdon and Palmer 1994; Stoltzfus 1994; Stoltzfus et al. 1994; Logsdon et al. 1995; Logsdon 1998). At present, despite extensive efforts to identify traces of primordial introns, there seems to be no substantial empirical evidence in support of introns-early, and consequently, spliceosomal introns and the spliceosome are generally considered a key, early eukaryotic innovation (Stoltzfus et al. 1994; Cho and Doolittle 1997; Logsdon et al. 1998; Belshaw and Bensasson 2006; Koonin 2006).

Introns and the spliceosome are among those fundamental features of the eukaryotic cell (along with the nucleus, the cytoskeleton, and several others) that, beyond reasonable doubt, were already present in the last eukaryotic common ancestor. The origin of each of these innovations and the earliest steps in their evolution are shrouded in mystery, given that they are present in their (nearly) fully developed form in all extant eukaryotes (Mans et al. 2004; Collins and Penny 2005). The eukaryotic spliceosomal introns as well as the spliceosomal small RNAs that perform critical functions in splicing show significant similarities to self-splicing group II introns (retroelements), found in bacteria and organelles and are thought to have evolved from such elements (Zimmerly et al. 2001; Lambowitz and Zimmerly 2004; Toro et al. 2007). However, the process of the acquisition of spliceosomal introns by eukaryotic genomes is not understood beyond general schemes (Koonin 2006; Martin and Koonin 2006).

The dynamics and mechanisms of intron evolution are not well understood either. Generally, the abundance and turnover rate of introns in a genome are thought to be determined by the effective population size and the characteristic mutation rate of the respective species (Lynch and Richardson 2002; Lynch and Conery 2003); however, it has been argued that various selective forces could substantially affect the rates of intron gain and loss (Jeffares et al. 2006). Comparative genomic studies have revealed impressive conservation of intron positions in diverse animals (Raible et al. 2005) and have shown that the positions of many introns are shared by orthologous genes even in distant eukaryotes, such as animals and plants (Fedorov et al. 2002; Rogozin et al. 2003). However, several recent models of the processes of intron gain and loss have yielded widely contradicting scenarios. Some models lead to an evolutionary

Key words: intron gain, intron loss, plant evolution, plastid-derived genes.

E-mail: koonin@ncbi.nlm.nih.gov.

Mol. Biol. Evol. 25(1):111–119. 2008
doi:10.1093/molbev/msm234
Advance Access publication October 31, 2007

landscape dominated by intron gain (Qiu et al. 2004), whereas others suggest dominance of intron loss (Roy and Gilbert 2005a, 2005b) or offer intermediate solutions, which imply that intron gain and loss both made important and comparable contribution to the evolution of eukaryotic genes (Rogozin et al. 2003; Csuros 2005; Nguyen et al. 2005). A recent, comprehensive maximum likelihood reconstruction of intron gain and loss in eukaryotic evolution not only supports the latter point of view but also reveals a significant (~1.8-fold) overall excess of losses and a distinctly nonuniform distribution of gains and losses over the history of eukaryotes (Carmel et al. 2007b). In particular, substantial excess of intron gains over intron losses has been detected only for some of the early intervals of eukaryotic evolution that are associated with major evolutionary innovations, such as the origin of animals.

Somewhat unexpectedly, it is extremely hard to detect concrete evidence of intron gain events. It would seem plausible that new introns, at least, in part, are gained via transposition of preexisting introns; however, with a single exception noted below, the attempts to detect homologous introns in different genes within the same genome so far have failed (Fedorov et al. 2003). In complementary studies, comparison of orthologous genes from mammalian genomes failed to reveal any gains at all, suggesting that all introns currently contained in mammalian genes were already present at the time of the radiation of mammalian orders (Roy et al. 2003), and similar results have been reported in a detailed analysis of several plant gene families (Teich et al. 2007). A reconstruction of intron gain and loss in paralogous gene families from animals and plants similarly indicated no significant gain during the last several hundred million years of evolution (Babenko et al. 2004). Recently, a focused attempt was undertaken to identify intron gain in genes of the protist *Entamoeba histolytica* that appear to have been acquired via horizontal gene transfer from bacteria; any introns present in these genes would be relatively recent gains (Roy et al. 2006). However, Roy et al. detected very few introns in the horizontally transferred genes, indicating a low gain rate. About the only purported showcase of intron gain is the report on an apparent gain of 122 introns in the nematodes of the genus *Caenorhabditis* since the divergence of the 2 nematodes with sequenced genomes, *Caenorhabditis elegans* and *Caenorhabditis briggsae* (Coghlan and Wolfe 2004). However, a recent reanalysis has suggested that most of these purported intron gains might actually represent losses of ancestral introns (Roy and Penny 2006). In addition to the virtual lack of comparative genomic evidence in support of recent intron gain, the mechanisms of new intron insertion into genes are not understood; this contrasts the case of intron loss that is widely believed to be mediated, primarily, by recombination between intron-containing genes and intronless cDNA copies of the corresponding mRNAs (Fink 1987; Mourier and Jeffares 2003).

Given the elusiveness of intron gain, evolutionary scenarios are seriously considered under which most introns that are currently present in eukaryotic genes have been inserted at very early stages of the evolution of eukaryotes, whereas subsequent evolution had been dominated by intron loss (Roy and Gilbert 2005a, 2005b; Roy 2006). However, the logic of the aforementioned study of introns in the relatively few horizontally transferred genes of *Entamoeba* (Roy et al. 2006) applies also to more momentous events in the evolution of eukaryotes. In particular, plants have been shown to carry thousands of genes that have been transferred from the genome of the cyanobacterial endosymbiont to the nuclear genome (Martin et al. 1998, 2002; Timmis et al. 2004). There is no doubt that, at the time of the transfer, each of these genes was intronless, so any introns found in genes of plastid (chloroplast) derivation would result from de novo insertion (gain), even if occurring soon after the symbiosis event. In an attempt to infer the salient features of the intron gain process, we compared the characteristics of introns found in plastid-derived and ancestral eukaryotic genes present in plant genomes. We found that plastid-derived genes have a slightly but significantly lower intron density than ancestral eukaryotic genes in the same plants and that the process of intron acquisition leading to the increase of intron density in plastid-derived genes is still ongoing, albeit at a low rate.

Materials and Methods

Identification of Orthologous Genes

Rice genome annotation and sequences were downloaded from Rice Annotation Project DataBase (http://rapdb.lab.nig.ac.jp/), *Arabidopsis* genome annotation and sequences were from National Center for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nih.gov/genomes/), and poplar sequences were from the Joint Genome Institute (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html). Orthologs in rice, poplar, and *Arabidopsis* orthologs were identified using a modified InParanoid algorithm (O'Brien et al. [2005] and Basu MK, unpublished data). Pairs of orthologous proteins that differed by >25% in length were discarded from the analysis to avoid the potential impact of inaccurate gene annotations. Altogether, ~2,500 clusters containing exactly 1 orthologous gene from Arabidopsis, poplar, and rice each were identified. Each of the 3 pairwise comparisons between plant species yielded a greater number of orthologs because of lineage-specific duplications, particularly, the whole-genome duplication in poplar (Tuskan et al. 2006). The *Arabidopsis*–poplar comparison produced ~7,000 orthologous gene pairs, the *Arabidopsis*–rice comparison ~5,500 pairs, and the poplar–rice comparison ~5,000 pairs.

The clusters of orthologous plant genes were partitioned into 3 classes, namely, chloroplast derived, ancestral eukaryotic, and plant specific. The plastid-derived genes were identified by phylogenetic affinity to cyanobacterial genes using a modification of the previously described procedure (Martin et al. 2002). Specifically, each *Arabidopsis* protein sequence was compared with a data set that included proteins from 200 eubacteria, 24 archaebacteria, and 13 nonphotosynthetic eukaryotes (extracted from the NCBI Genome database, National Institutes of Health [NIH], Bethesda, MD; supplementary table S1, Supplementary Material online) using the BlastP program. The Blast hits with an E value $>10^{-10}$ and <25% amino acid identity were discarded, and the remaining, highly significant hits were ranked by the percent identities normalized by the fraction of aligned residues. The best hit from each phylum was selected for further analysis. The *Arabidopsis* protein sequences were aligned with the homologous sequences from other species using ClustalW (Thompson et al. 1994), and sites containing gaps were removed. Phylogenetic trees were built with the Neighbor-Joining using the PROTDIST program of the PHYLIP package (Felsenstein 1996) with the Jones-Thornton-Taylor substitution matrix for the calculation of protein distances and the Neighbor tree reconstruction. The trees in Newick format were parsed using a PERL script, and *Arabidopsis* genes that had a cyanobacterial homolog as a sister taxon were defined as chloroplast-derived genes. *Arabidopsis* genes that had homologues only in cyanobacterial genomes were classified as chloroplast-derived genes as well. The plastid-derived genes in poplar and rice were defined as orthologs of the plastid-derived genes of *Arabidopsis* that were identified in pairwise genome comparisons.

The rest of the clusters were classified as ancestral or plant specific using the following criteria. The *Arabidopsis* proteins from the uncategorized clusters were compared with the nonredundant protein sequence database (NCBI, NIH) using the BlastP program with the composition-based statistical adjustment (Altschul et al. 1997; Schaffer et al. 2001). All clusters containing proteins with highly significant sequence similarity (E value $< 10^{-10}$) to at least 1 homologous protein in both fungi and metazoa were classified as ancestral. In addition, a subset of more liberally defined ancestral genes was identified using a relaxed criterion (E value $< 10^{-4}$). The remaining clusters, that is, those for which, under the strict or the relaxed criterion, no significant similarity to nonplant proteins was detected, were classified as plant specific. This procedure partitioned the identified single-gene orthologous clusters from 3 plants into 304 chloroplast-derived, 1,104 (1,279) ancestral, and 545 (650) plant-specific clusters (numbers obtained with the relaxed criterion given in parentheses). All the analyses described below involved comparisons between chloroplast-derived and ancestral genes; the plant-specific genes were disregarded.

Conservation of Intron Positions

Intron conservation was determined, essentially, as previously described (Rogozin et al. 2003). Amino acid sequences of orthologous plant proteins were aligned using the MUSCLE program (Edgar 2004), and intron positions were then mapped onto the alignments. All the positions that contained 1 or more gaps within 5 amino acid upstream and downstream were discarded from the analysis.

Intron Gain and Loss

Intron gain and loss in the 2 dicot species, *Arabidopsis* and poplar, were calculated using Dollo parsimony (Rogozin et al. 2005), with rice as an outgroup. Introns that are shared by 1 of the dicot species and rice are counted as being lost in the second dicot species, whereas introns present in 1 of the dicot species but missing in the other and in rice are counted as gained in the respective dicot species. The same alignment quality criterion was applied as for the analysis of intron position conservation.

Information Content of the Splice Junction and Intron Phase

Information content at the splice junctions was calculated for the −6 to +4 positions of the donor and acceptor sites of each reliable (no gaps in the alignment within a 5 amino acid window on either side) intron positions in the clusters of orthologous plant genes. The information content was calculated using the formula:

$$I_k = 2 + \sum_{i=1}^{4} p_{ik} \log_2(p_{ik}),$$

where, p_{ik} is the nucleotide frequency of ith nucleotide in position k (Stephens and Schneider 1992; Sverdlov et al. 2003).

Evolutionary Rate

Nonsynonymous (Ka) and synonymous (Ks) substitution rates for each protein-coding gene sequence were calculated as follows. For each protein sequence in every alignment, the corresponding coding sequences (CDSs without introns) were extracted from the database. These sequences were aligned using the protein sequence alignment as guide. The calculation of the Ka, Ks, and Ka/Ks were performed using the method of Nei and Gojobori (1986) as implemented in the program yn00 of the PAML package (Yang 2007).

Results and Discussion

Plastid-Derived Genes Have Slightly but Significantly Lower Intron Densities than Ancestral Genes and Fewer Conserved Intron Positions

In an attempt to characterize the process of intron gain in plastid-derived genes, we compared a variety of features of these genes with the corresponding features of ancestral eukaryotic genes in the same plants. The plastid-derived genes represented an updated set of genes that showed a phylogenetic affinity with cyanobacteria and were identified, essentially, as described previously (Martin et al. [2002] and see Materials and Methods). The ancestral eukaryotic genes were defined as those that are conserved between plants and representatives of other eukaryotic kingdoms and hence appear to antedate the origin of Plantae (disregarding the unlikely possibility of horizontal gene transfer between phylogenetically remote eukaryotes); 2 levels of stringency were employed to delineate a more permissive and a more strictly defined sets of ancestral genes (for details see Materials and Methods). Clearly, the density of introns in plastid-derived genes from all 3 plant genomes, calculated either per gene or per unit length of the CDS and either by comparing the means or the entire density distributions across the plastid-derived and ancestral gene sets,

Table 1
Intron Density in Plastid-Derived Genes and Ancestral Genes of Plants

		Plastid Derived[a]	Ancestral (relaxed ancestral)[a]	*P* value (*P* value with relaxed ancestral)
Arabidopsis	Genes	1,543	6,609 (7,933)	
	Introns	7,904	44,989 (50,073)	
	Amino acids	701,238	3,292,488 (3,823,834)	
	Mean protein length, aa	454	498	
	Intron content (/gene)	5.12	6.80 (6.31)	$<2.2 \times 10^{-16}$ ($<2.2 \times 10^{-16}$)[b]
	Mean intron density (/aa)	0.0113	0.0137 (0.0131)	$<2.2 \times 10^{-16}$ ($<2.2 \times 10^{-16}$)[b]
	Median intron density (/aa)	0.011	0.0139 (0.013)	$<2.2 \times 10^{-16}$ (1.895×10^{-09})[c]
Poplar	Genes	831	3,059 (3,630)	
	Introns	4,509	21,244 (23,506)	
	Amino acids	349,197	1,503,459 (1,735,468)	
	Mean protein length, aa	420	491	
	Intron content (/gene)	5.42	6.94 (6.47)	3.126×10^{-13} (1.859×10^{-07})
	Mean intron density (/aa)	0.0129	0.0141 (0.0135)	3.679×10^{-08} (3.4×10^{-03})
	Median intron density (/aa)	0.0129	0.0147 (0.0140)	9.73×10^{-03} (0.34)
Rice	Genes	767	2,716 (3,129)	
	Introns	4,488	20,322 (22,299)	
	Amino acids	332,527	1,361,110 (1,539,890)	
	Mean protein length, aa	434	501	
	Intron content (/gene)	5.85	7.48 (7.12)	4.412×10^{-16} (5.34×10^{-11})
	Mean intron density (/aa)	0.0135	0.0149 (0.0145)	8.39×10^{-10} (1.764×10^{-05})
	Median intron density (/aa)	0.0136	0.0160 (0.0153)	7.46×10^{-06} (7.6×10^{-04})

NOTE.—aa, amino acid.

[a] The gene sets in poplar and rice were derived by orthology with the respective *Arabidopsis* genes determined in pairwise genome comparison (see Materials and Methods); the numbers differ because not all *Arabidopsis* genes had one-to-one orthologs in poplar and rice.

[b] Calculated using Fisher's exact test.

[c] Calculated by comparing intron density distributions across the respective gene sets using the *t*-tests.

was slightly but significantly lower than the respective densities in the ancestral genes (table 1).

A potential complication to this conclusion is that the proteins encoded by plastid-derived genes are, on average, shorter than those encoded by ancestral genes (table 1). It has been reported that intron density positively correlates with the length of a gene's CDS (Lynch and Kewalramani 2003); accordingly, it could not be ruled out that the plastid-derived genes had lower intron density than ancestral genes solely because they encoded, on average, shorter proteins. To address this problem, we compared the intron densities in the plastid-derived genes of each of the 3 plant species with those in a randomly generated subset of ancestral genes with CDS length distributions statistically indistinguishable from those in the plastid-derived genes. The difference in intron densities between plastid-derived genes and ancestral genes was detected in these comparisons as well and remained highly significant for *Arabidopsis* and rice although not for poplar (supplementary table S2, Supplementary Material online). Thus, we conclude that the lower intron density in plastid-derived genes is not, primarily, explained by the differences in the lengths of the encoded proteins.

By contrast, a comparison of the distributions of intron lengths in plastid-derived genes and in ancestral genes did not reveal any significant differences, with the median intron lengths being 114 and 113 nt, respectively (fig. 1). Thus, plastid-derived genes and ancestral genes in the same plants have significantly different intron densities but, essentially, the same characteristic intron length. It has been reported that intron length is linked to gene expression level, but although the signs of the correlation appear to be opposite in animals and plants such that highly expressed animal genes have longer introns than lowly expressed ones (Castillo-Davis et al. 2002), whereas, in plants, highly expressed genes have longer introns (Ren et al. 2006). However, the near identity of the distributions of intron lengths in plastid-derived and ancestral genes suggests that the lower intron density in the plastid-derived genes is unrelated to expression but rather is rooted in the history of the 2 classes of genes.

We further compared the level of conservation (the amount of variation) of intron positions in the plastid-derived and ancestral genes. In agreement with the previous observations (Roy and Penny 2007), over 95% of the intron positions were found to be conserved in both classes of genes in all 3 pairwise comparisons of plant genomes. However, the plastid-derived genes consistently showed a lower level of conservation and the difference was highly significant for the comparisons of each of the dicots with rice (table 2), suggesting that lineage-specific loss and/or gain of introns were more prominent in plastid-derived than in ancestral genes.

A Higher Rate of Intron Gain but Not Intron Loss in Plastid-Derived Genes Compared with Ancestral Genes

We further sought to directly compare the rates of intron gain and loss in plastid-derived and ancestral genes of plants. Given the availability of the complete genomes of 3 relatively close plant species with an unequivocally resolved phylogeny, the number of intron gains and losses in each of the 2 dicotyledons (*Arabidopsis* and poplar) could be determined in a straightforward fashion using the Dollo parsimony method (Rogozin et al. 2003,

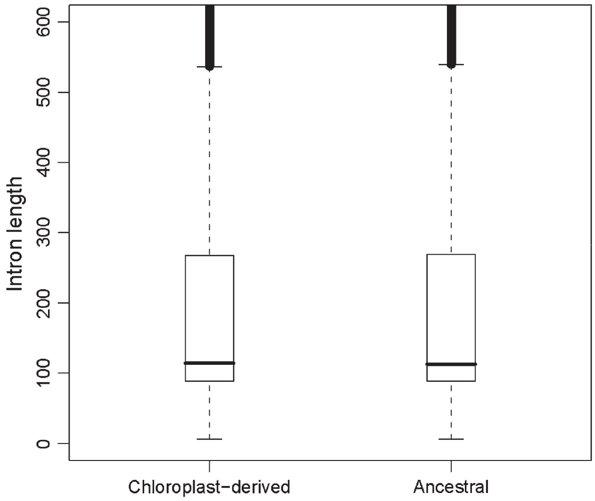

FIG. 1.—Distributions of intron lengths in chloroplast-derived and ancestral genes of plants. The median values were 114 and 113 nt, respectively. The *y* axis is shown only partially to emphasize the median values. The box extends to the first quartile above and below the median, and the thin lines extend to the third quartile. The thick black line shows the data beyond the third quartile.

2005). Introns that are shared by 1 of the dicot species and rice are counted as being lost in the second dicot species, whereas introns present in 1 of the dicot species but missing in the other and in rice are counted as gained in the respective species. The use of Dollo parsimony as opposed to more complex maximum likelihood models (Carmel et al. 2005, 2007b; Csuros 2005; Nguyen et al. 2005) is justified, in this case, because, at such short evolutionary distances, the contribution of parallel gain of introns in the sites in independent lineages is negligible (Rogozin et al. 2005; Sverdlov et al. 2005); maximum likelihood analysis was not practical because of the small number of potential gains and losses. In a general agreement with a previous report (Roy and Penny 2007), we found that only a few introns were lost or gained after the divergence of the 2 dicots (table 3). In *Arabidopsis*, there was a substantial excess of losses over gains, again, as previously reported (Roy and Penny 2007); no such excess of intron loss was detected in poplar, which has a much less compact genome than *Arabidopsis* (Tuskan et al. 2006) (table 3). Although the compared numbers of introns lost and gained were small, a significant excess of gains in plastid-derived genes was detected as compared with ancestral genes (table 3); by contrast, there was no significant difference in the rates of intron loss between the 2 classes of plant genes (table 3). Overall, in *Arabidopsis*, even in plastid-derived genes, slightly more introns were lost than gained but in poplar, a small net gain of introns was seen (table 3). Although some overestimate of intron gain is possible due to parallel losses (even as not many parallel losses are expected given the small overall number of lost introns), this effect would apply equally to plastid-derived and ancestral genes.

The emerging interplay between intron loss and gain in plant genes is fairly complex. In ancestral eukaryotic genes, intron loss decidedly dominates over gain, in agreement with the findings of Roy and Penny (2007). The pattern of intron loss and gain in the plastid-derived genes can be plausibly conceived of as a superposition of 2 opposite trends: 1) the gain of introns that continues, albeit at a low rate, since the time of the original invasion and 2) the general trend of intron loss that is particularly pronounced in the Arabidopsis lineage, probably, due to the evolution toward genome contraction.

Differences in Intron Phase Distribution and Splice Sites Information Content between Chloroplast-Derived and Ancestral Genes

We further compared additional salient features of introns in chloroplast-derived and ancestral genes, including intron phase distributions and the patterns of information content in and around the donor and acceptor splice sites. It is well known that introns in all species show a substantial excess of phase 0 (i.e., introns located between codons) over phases 1 and 2 (introns located after the first and second bases of a codon, respectively) (Fedorov et al. 1992;

Table 2
Conservation of Intron Positions in Plastid-Derived and Ancestral Genes in Plants

	Conserved Intron Positions	Variable Intron Positions	Fraction of Variable Positions
Arabidopsis versus Rice			
Plastid derived	2,384	106	0.042
Ancestral (relaxed ancestral)	12,216 (13,132)	415 (461)	0.032 (0.033)
P values (Fisher's exact test)			
Plastid derived versus ancestral		0.0188	
Plastid derived versus relaxed ancestral		0.0333	
Arabidopsis versus Poplar			
Plastid derived	2,450	61	0.024
Ancestral (relaxed)	12,432 (13,362)	255 (280)	0.020 (0.020)
P values (Fisher's exact test)			
Plastid derived versus ancestral		0.192	
Plastid derived versus relaxed ancestral		0.226	
Rice versus Poplar			
Plastid derived	2,420	85	0.033
Ancestral (relaxed)	12,412 (13,360)	328 (357)	0.025 (0.026)
P values (Fisher's exact test)			
Plastid derived versus ancestral		0.026	
Plastid derived versus relaxed ancestral		0.027	

Table 3
Intron Gain and Loss in Plastid-Derived and Ancestral Genes of Dicotyledons

		Plastid-derived (P)	Ancestral (A)	Ancestral Relaxed (R)
Introns		1,310	6,563	7,069
	Gain (fraction)	15(0.011)	29(0.004)	32(0.005)
Arabidopsis	Loss (fraction)	26(0.020)	142(0.022)	160(0.023)
P values (Fisher's exact test)	P and A	0.0092		
	P and R	0.0087		
Poplar	Gain (fraction)	12(0.009)	40(0.006)	42(0.006)
	Loss (fraction)	8(0.006)	44(0.007)	46(0.007)
P values (Fisher's exact test)	P and A	0.456		
	P and R	0.458		
	Total gain	27	69	74
	Total loss	34	186	206
P values (Fisher's exact test)	P and A	0.0126		
	P and R	0.0082		

de Souza et al. 1998). Subsequently, it has been noticed that relatively new introns have an even greater excess of phase 0 than ancient introns, and this has been explained by the apparent migration of information from the exonic splice signals to the sequence of the intron itself, in the course of an intron's evolution (Sverdlov et al. 2003). Under this hypothesis, introns preferentially fix in phase 0 but, subsequently, some of them slide to phases 1 and 2, thanks to the relaxed selective constraints on flanking nucleotides after the migration of information into the intron. The results of the comparison of the phase distributions of introns in the chloroplast-derived and ancestral genes of plants appeared to be in line with this interpretation of intron evolution. In particular, the introns in chloroplast-derived genes showed a significantly greater excess of phase 0 than ancestral genes (table 4). The comparison of the information content of splice sites also conformed with the previously established pattern. The sites in chloroplast-derived genes showed a higher information content in exonic positions flanking the splice junctions, especially, in the −1 position upstream of the donor site and the +1 position downstream of the acceptor site, and conversely, ancestral genes showed a stronger signal in intronic positions adjacent to the splice junctions (fig. 2).

Conclusions

The comparisons of the intron densities and features in plant genes derived from plastids and ancestral eukaryotic genes from the same plants lead to 2 principal, complementary conclusions. First, the chloroplast-derived genes that contained no introns when they resided in the cyanobacterial genome immediately after the symbiosis have accumulated a large number of introns such that the intron density in these genes became close to that in the ancestral genes. The current collection of genomes from the plant lineage is insufficient to reconstruct the dynamics of intron invasion into the plastid-derived genes. However, almost 96% of the intron positions in the plastid-derived genes are conserved between dicotyledons and monocotyledons, indicating that the great majority of introns were acquired by these genes prior to the dicot–monocot divergence 150–200 MYA (Wolfe et al. 1989; Chaw et al. 2004). Previous analyses have suggested that extensive intron gain occurred only during short intervals of eukaryotic evolution, possibly, coinciding with major evolutionary innovations (Babenko et al. 2004; Carmel et al. 2007b). Thus, it seems likely that the bulk of the invasion occurred over a relatively short time span such that, in this respect, the occupation of the plastid-derived genes with introns resembled the original invasion of introns into the genes of the emerging eukaryotic cell, presumably, following the mitochondrial endosymbiosis (Koonin 2006; Martin and Koonin 2006). These results unequivocally show that massive intron gain occurred on at least 1 occasion during eukaryotic evolution other than the original invasion. At present, both demonstrable waves of intron invasion are associated with primary endosymbiosis; it remains yet uncertain whether or not extensive intron gain also accompanies secondary endosymbiotic events.

Table 4
Phase Distribution of Introns in Chloroplast-Derived and Ancestral Genes of Plants

		Phase 0[a]	Phase 1[a]	Phase 2[a]	Total
Arabidopsis	Chloroplast derived	4,735 (0.599)	1,662 (0.210)	1,507 (0.190)	7,904
	Ancestral	26,693 (0.593)	8,639 (0.192)	9,657 (0.214)	44,989
	Ancestral relaxed	29,593 (0.590)	9,783 (0.195)	10,697 (0.213)	50,073
P values[b]	Chloroplast derived versus Ancestral	2.769×10^{-07}			
	Chloroplast derived versus Ancestral relaxed	2.86×10^{-06}			
Poplar	Chloroplast derived	2,713 (0.601)	889 (0.197)	907 (0.201)	4,509
	Ancestral	12,733 (0.599)	4,021 (0.189)	4,490 (0.211)	21,244
	Ancestral relaxed	13,980 (0.594)	4,546 (0.193)	4,980 (0.211)	23,506
P values[b]	Chloroplast derived versus Ancestral	0.21			
	Chloroplast derived versus Ancestral relaxed	0.26			
Rice	Chloroplast derived	2,706 (0.602)	907 (0.202)	875 (0.194)	4,488
	Ancestral	12,040 (0.592)	3,979 (0.195)	4,303 (0.211)	20,322
	Ancestral relaxed	13,173 (0.590)	4,401 (0.197)	4,725 (0.211)	22,299
P values[b]	Chloroplast derived versus ancestral	0.04			
	Chloroplast derived versus ancestral relaxed	0.03			

[a] Number in the parenthesis indicated fraction of total.

[b] Calculated using chi-square 2 × 3 table with degrees of freedom = 2.

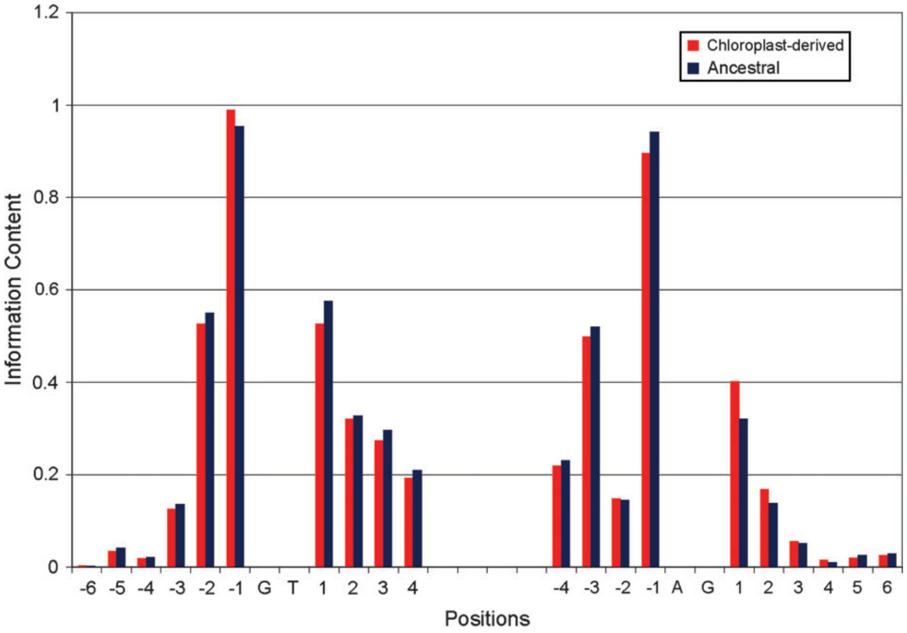

FIG. 2.—Information content of exonic and intronic positions flanking the splice junctions of chloroplast-derived and ancestral genes of plants.

Second, however, we observed that the plastid-derived genes continued to accumulate new introns at a slightly but significantly greater rate than ancestral genes even after the divergence of the 2 dicot species. This suggests that the plastid-derived genes still have not reached the saturation intron density. It is tempting to speculate that this saturation density, which, of course, differs between eukaryotic taxa, has an adaptive value, a notion that is in agreement with the recent demonstration of a positive correlation between intron gain rate and evolutionary conservation of genes (Carmel et al. 2007a).

The pattern of intron gain and loss in plastid-derived genes is further complicated by the superposition of the apparent continued gain and the more general trend of intron loss during the recent evolution of plant genes, at least, in the lineages where genome contraction is observed, such as *Arabidopsis*. The present conclusions give only a rough, first approximation reconstruction of the evolutionary dynamics of introns in the plastid-derived genes because of the small numbers of reliably identifiable intron losses and gains. For a more detailed reconstruction, a diverse sample of genomes from different branches of plants is required.

Supplementary Material

Supplementary tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Miklos Csuros for helpful discussions. This work was supported in part by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/Department of Health and Human Services.

Literature Cited

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Babenko VN, Rogozin IB, Mekhedov SL, Koonin EV. 2004. Prevalence of intron gain over intron loss in the evolution of paralogous gene families. Nucleic Acids Res. 32:3724–3733.

Belshaw R, Bensasson D. 2006. The rise and falls of introns. Heredity. 96:208–213.

Carmel L, Rogozin IB, Wolf YI, Koonin EV. 2005. An expectation-maximization algorithm for analysis of evolution of exon-intron structure of eukaryotic genes. Comp Genomics Lect Notes Comput Sci. 3678:35–46.

Carmel L, Rogozin IB, Wolf YI, Koonin EV. 2007a. Evolutionarily conserved genes preferentially accumulate introns. Genome Res. 17:1045–1050.

Carmel L, Wolf YI, Rogozin IB, Koonin EV. 2007b. Three distinct modes of intron dynamics in the evolution of eukaryotes. Genome Res. 17:1034–1044.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. Nat Genet. 31:415–418.

Chaw SM, Chang CC, Chen HL, Li WH. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol. 58:424–441.

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

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

Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. Mol Biol Evol. 22:1053–1066.

Csuros M. 2005. Likely scenarios of intron evolution. Comp Genomics Lect Notes Comput Sci. 3678:47–60.

Darnell JE Jr. 1978. Implications of RNA-RNA splicing in evolution of eukaryotic cells. Science. 202:1257–1260.

de Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W. 1998. Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proc Natl Acad Sci USA. 95:5094–5099.

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

Doolittle WF. 1978. Genes in pieces: were they ever together? Nature. 272:581–582.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Fedorov A, Merican AF, Gilbert W. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. Proc Natl Acad Sci USA. 99:16128–16133.

Fedorov A, Roy S, Fedorova L, Gilbert W. 2003. Mystery of intron gain. Genome Res. 13:2236–2241.

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

Felsenstein J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 266:418–427.

Fink GR. 1987. Pseudogenes in yeast? Cell. 49:5–6.

GenBank [Internet]. Bethesda (MD): National Center for Biotechnology Information; c2002 [cited 2007 Feb 15]. Available from: http://www.ncbi.nlm.nih.gov/Genbank/index.html

Gilbert W. 1978. Why genes in pieces? Nature. 271:501.

Gilbert W. 1987. The exon theory of genes. Cold Spring Harb Symp Quant Biol. 52:901–905.

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

Gilbert W, Glynias M. 1993. On the ancient nature of introns. Gene. 135:137–144.

Jeffares DC, Mourier T, Penny D. 2006. The biology of intron gain and loss. Trends Genet. 22:16–22.

Koonin EV. 2006. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biol Direct. 1:22.

Lambowitz AM, Zimmerly S. 2004. Mobile group II introns. Annu Rev Genet. 38:1–35.

Logsdon JM Jr. 1998. The recent origins of spliceosomal introns revisited. Curr Opin Genet Dev. 8:637–648.

Logsdon JM Jr, Palmer JD. 1994. Origin of introns—early or late? Nature. 369:526.

Logsdon JM Jr, Stoltzfus A, Doolittle WF. 1998. Molecular evolution: recent cases of spliceosomal intron gain? Curr Biol. 8:R560–R563.

Logsdon JM Jr, Tyshenko MG, Dixon C, D-Jafari J, Walker VK, Palmer JD. 1995. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. Proc Natl Acad Sci USA. 92:8507–8511.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science. 302:1401–1404.

Lynch M, Kewalramani A. 2003. Messenger RNA surveillance and the evolutionary proliferation of introns. Mol Biol Evol. 20:563–571.

Lynch M, Richardson AO. 2002. The evolution of spliceosomal introns. Curr Opin Genet Dev. 12:701–710.

Mans BJ, Anantharaman V, Aravind L, Koonin EV. 2004. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. Cell Cycle. 3:1612–1637.

Martin W, Koonin EV. 2006. Introns and the origin of nucleus-cytosol compartmentalization. Nature. 440:41–45.

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA. 99:12246–12251.

Martin W, Stoebe B, Goremykin V, Hapsmann S, Hasegawa M, Kowallik KV. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. Nature. 393:162–165.

Mattick JS. 1994. Introns: evolution and function. Curr Opin Genet Dev. 4:823–831.

Mourier T, Jeffares DC. 2003. Eukaryotic intron loss. Science. 300:1393.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

Nguyen HD, Yoshihama M, Kenmochi N. 2005. New maximum likelihood estimators for eukaryotic intron evolution. PLoS Comput Biol. 1:e79.

O'Brien KP, Remm M, Sonnhammer EL. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 33:D476–D480.

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

Raible F, Tessmar-Raible K, Osoegawa K, et al. (12 co-authors). 2005. Vertebrate-type intron-rich genes in the marine annelid Platynereis dumerilii. Science. 310:1325–1326.

Ren XY, Vorst O, Fiers MW, Stiekema WJ, Nap JP. 2006. In plants, highly expressed genes are the least compact. Trends Genet. 22:528–532.

Rice Annotation Project Database [Internet]. Japan: National Institute of Agrobiological Sciences and National Institute of Genetics; c2005 [cited 2006 Dec 20]. Available from: http://rapdb.lab.nig.ac.jp/

Rogozin IB, Babenko VN, Wolf YI, Koonin EV. 2005. Dollo parsimony and reconstruction of genome evolution. In: Albert VA, editor. Parsimony, phylogeny, and genomics. Oxford: Oxford University Press. p. 190–200.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, 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.

Roy SW. 2006. Intron-rich ancestors. Trends Genet. 22:468–471.

Roy SW, Fedorov A, 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.

Roy SW, Gilbert W. 2005a. Complex early genes. Proc Natl Acad Sci USA. 102:1986–1991.

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

Roy SW, Irimia M, Penny D. 2006. Very little intron gain in Entamoeba histolytica genes laterally transferred from prokaryotes. Mol Biol Evol. 23:1824–1827.

Roy SW, Penny D. 2006. Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses. Mol Biol Evol. 23:2259–2262.

Roy SW, Penny D. 2007. Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of O. sativa and A. thaliana. Mol Biol Evol. 24:171–181.

Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29:2994–3005.

Stephens RM, Schneider TD. 1992. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. J Mol Biol. 228:1124–1136.

Stoltzfus A. 1994. Origin of introns—early or late. Nature. 369: 526–527.

Stoltzfus A, Spencer DF, Zuker M, Logsdon JM Jr, Doolittle WF. 1994. Testing the exon theory of genes: the evidence from protein structure. Science. 265:202–207.

Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. 2003. Evidence of splice signal migration from exon to intron during intron evolution. Curr Biol. 13:2170–2174.

Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. 2005. Conservation versus parallel gains in intron evolution. Nucleic Acids Res. 33:1741–1748.

Teich R, Grauvogel C, Petersen J. 2007. Intron distribution in Plantae: 500 million years of stasis during land plant evolution. Gene. 394:96–104.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 5:123–135.

Toro N, Jimenez-Zurdo JI, Garcia-Rodriguez FM. 2007. Bacterial group II introns: not just splicing. FEMS Microbiol Rev. 31:342–358.

Tuskan GAS, Difazio S, Jansson J, et al. (110 co-authors). 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science. 313:1596–1604.

Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA. 86:6201–6205.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Zimmerly S, Hausner G, Wu X. 2001. Phylogenetic relationships among group II intron ORFs. Nucleic Acids Res. 29:1238–1250.

Aoife McLysaght, Associate Editor

Accepted October 22, 2007