A Very High Fraction of Unique Intron Positions in the Intron-Rich Diatom Thalassiosira pseudonana Indicates Widespread Intron Gain

Scott William Roy and David Penny
Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

Although spliceosomal introns are present in all characterized eukaryotes, intron numbers vary dramatically, from only a handful in the entire genomes of some species to nearly 10 introns per gene on average in vertebrates. For all previously studied intron-rich species, significant fractions of intron positions are shared with other widely diverged eukaryotes, indicating that 1) large numbers of the introns date to much earlier stages of eukaryotic evolution and 2) these lineages have not passed through a very intron-poor stage since early eukaryotic evolution. By the same token, among species that have lost nearly all of their ancestral introns, no species is known to harbor large numbers of more recently gained introns. These observations are consistent with the notion that intron-density genomes have arisen only once over the course of eukaryotic evolution. Here, we report an exception to this pattern, in the intron-rich diatom Thalassiosira pseudonana. Only 8.1% of studied T. pseudonana intron positions are conserved with any of a variety of divergent eukaryotic species. This implies that T. pseudonana has both 1) lost nearly all of the numerous introns present in the diatom–apicomplexan ancestor and 2) gained a large number of new introns since that time. In addition, that so few apparently inserted T. pseudonana introns match the positions of introns in other species implies that insertion of multiple introns into homologous genomic sites in eukaryotic evolution is less common than previously estimated. These results suggest the possibility that intron-rich lineages may have arisen multiple times in evolution. These results also provide evidence that multiple intron insertion into the same site is rare, further supporting the notion that early eukaryotic ancestors were very intron rich.

Introduction

Spliceosomal introns are sequences that interrupt coding sequences in eukaryotes and which are removed from RNA transcripts by the spliceosome. Introns are specific and common to eukaryotes, however genomic intron number varies by more than 4 orders of magnitude, from hundreds of thousands of introns per genome in many metazoans and plants to fewer than 100 introns in a wide variety of protists, algae, and fungi (compiled in Logsdon 1998; Jeffares et al. 2006; Roy and Gilbert 2006; Rodríguez-Trelles et al. 2006b). The timing, causes, and mechanisms of intron origin have long been matters of debate (e.g., Hickey and Benkel 1986; Stoltzfus 1994; Elder 1991, 2000; Giroux et al. 1994; Poole et al. 1998; Venkatesh et al. 1999; Lynch 2002; Fedorov et al. 2003; Mourier and Jeffares 2003; de Roos 2004; Fedorov and Fedorova 2004; Roy 2004; Sverdlov, Babenko, et al. 2004; Collins and Penny 2005; Niu et al. 2005; Lin and Zhong 2005; Permal et al. 2005; Fedorov and Fedorova 2006; Knowles and McLysaght 2006; Logsdon 1991).

Recent genomic studies have shown 2 commonalities among diverse intron-rich species. First, in all studied intron-rich species (arbitrarily defined hereafter as at least 0.5 introns per gene on average), a significant fraction of intron positions match the position of introns in distantly related species (Fedorov et al. 2002; Rogozin et al. 2003). For instance, 60% of introns in the fission yeast Schizosaccharomyces pombe were found to match an intron position in one or more species out of 6 studied nonfungal species (Rogozin et al. 2003). Around a quarter of intron positions in conserved coding regions are conserved between the flowering plant Arabidopsis thaliana and humans (Rogozin et al. 2003). This pattern indicates that at least a significant fraction of introns in these species are retained from much earlier stages of eukaryotic evolution. Second, genome-wide comparisons of closely related species in a wide variety of intron-rich lineages have shown a surprising dearth of recent intron gains (Roy et al. 2003; Nielsen et al. 2004; Stajich and Dietrich 2006; Roy and Hartl 2006; Lin et al. 2006; Roy and Penny 2006a, 2006b; Roy et al. 2006). Rates of intron gain in the past tens to hundreds of million years in mammals, Plasmodium, Theliera, basiidomyctes and eusacomyctes fungi, Entamoeba, and land plants have been very low, equating to less than 1 intron gain per gene per 2 billion years. Two apparent exceptions involve nematodes and Oikopleura urochordates. In both taxa large numbers of lineage-specific intron positions suggest higher average rates of intron gain over deeper evolutionary distances (i.e., >100 MYA), however, both lineages retain at least around 15% of their ancestral introns, and rates of gain in the past 100 Myr of Caenorhabditis appear to also be low (Seo et al. 2001; Giuliani et al. 2002; Edvardsen et al. 2004; Coghlan and Wolfe 2004; Raible et al. 2005; Roy and Penny 2006b).

Recent studies have also confirmed that all known extremely intron-poor lineages have arisen by massive loss of relatively large numbers of ancestral introns. For example, a large fraction of S. pombe’s roughly 6,000 total introns were present in the fungal ancestor (because 60% of introns in conserved regions share positions with introns in nonfungi), implying that Saccharomyces cerevisiae’s much smaller number of introns (~300 in the entire genome) reflects massive intron loss (Rogozin et al. 2003). Similar conclusions can be drawn for all other known extremely intron-poor lineages. Recently, Slamovits and Keeling (2006) confirmed that some species of excavates, a diverse and potentially very early branching group of eukaryotes that includes a variety of extremely intron-poor lineages, contain a large numbers of ancestral intron positions, indicating massive (and likely
intron loss in very intron-poor excavates (Archibald et al. 2002). Thus, 1) all previously studied intron-rich eukaryotes retain significant numbers of ancestral introns and thus have never gone through a nearly intronless stage, whereas 2) all species that have lost nearly all of their ancestral introns have not replaced these ancestral introns with more recently gained introns but instead remain nearly intronless.

Here, we report an exception to this pattern in the diatom *Thalassiosira pseudonana* (Armburst et al. 2004). *Thalassiosira pseudonana* genes contain an average of 1.4 introns per gene, of which only around 8.1% were found to match intron positions in any of a variety of distantly related eukaryotes. *Thalassiosira pseudonana* has thus lost the vast majority of the numerous introns present in the chromalveolate ancestor and has also gained a large number of introns. These results have important implications for the ongoing debate on the relative importance of intron gain and loss (Rogozin et al. 2003, 2005; Babenko et al. 2004; Qiu et al. 2004; Csuros et al. 2005; Nguyen et al. 2005; Roy and Gilbert 2005a, 2005b), and support the notion that early eukaryotic ancestors were very intron rich, with intron losses outnumbering intron gains through subsequent evolution over a wide variety of eukaryotic lineages (Roy and Gilbert 2005a, 2005b).

**Materials and Methods**

**Gene Sequences and Intron Positions**

Predicted coding sequences and corresponding intron positions for full genome sequences of the studied chromalveolates were downloaded from the following sources: *T. pseudonana* (http://www.plasmodb.org/plasmo/home.jsp; version 1), *Plasmodium falciparum* (PlasmoDB.org; version 2), *Theileria parva* (National Center for Biotechnology Information [NCBI]; accession number AAGK01000001.1), *Toxoplasma gondii* (http://www.toxodb.org/toxo/home.jsp; version 4.0).

**Intron Prediction in *T. pseudonana***

We studied the predicted introns in the *T. pseudonana* genome. In all, 98.7% of introns begin with the canonical GT (and 99.2% begin with G[T/C]); 99.5% end with the canonical AG. However, we found 2 patterns suggesting that some introns in *T. pseudonana* are in fact coding sequences that have incorrectly been predicted to be intronic. First, there is a pronounced excess of introns that are a multiple of 3 bases (3n; 9,571 total) versus introns that are one more than a multiple of 3 bases (3n + 1; 3,037) or 2 more than a multiple of 3 bases (3n + 2; 3,029). Second, among introns that are a multiple of 3 bases, a larger fraction (75.2%) of 3n introns lack in-frame stop codons. This is a much larger fraction than for 3n + 1 or 3n + 2 introns. We can make separate estimates of the number of predicted introns that are instead protein encoding (i.e., not spliced out) from these 2 observations. First, assuming that there should be equal numbers of introns of the 3 different classes (an assumption which works very well for 3n + 1 and 3n + 2, which have 3,037 and 3,029 predicted introns, respectively), there are 6,540 “too many” 3n introns. Second, assuming that equal fractions of introns in the 3 classes should lack stop codons (which again works well for 3n + 1 and 3n + 2 introns, which have 29.1% and 28.4% stop-lacking introns, respectively), we estimate based on the 2,368 stop-containing 3n introns that there should be 972 stop-lacking introns, 6,233 fewer than currently predicted. The similarity of these 2 estimates suggests that around 6200 to 6600 (~90%) of the 7,205 predicted 3n in-frame stop codon–lacking introns are in fact coding sequence. We therefore excluded all 3n introns that lack in-frame stop codons from further consideration.

Fortunately, problems with intron prediction are unlikely to yield gene structures in which flanking coding sequences on both sides of the intron show strong sequence similarity to corresponding sequences in orthologous genes because incorrect intron prediction will lead to frameshifts and/or large alignment gaps (e.g., Roy et al. 2003; Roy and Hartl 2006). Indeed, exclusion of the excess 3n introns did not substantially change the total numbers of introns in conserved regions for the studies reported here or cause large changes in the results. In particular, that predicted intron positions in conserved regions are in fact true intron positions is underscored by the lack of gaps in nearby coding sequence: for each of the 3 pairwise comparisons with apicomplexan species, more than 80% of species-specific intron positions in conserved regions were in alignment regions with no gap within 5 codons.

**Comparison with Apicomplexans**

For each apicomplexan species, BlastP searches against *T. pseudonana* yielded best reciprocal hits. Each protein sequence pair was then aligned in ClustalW using default parameters and intron positions mapped onto resultant alignments. We defined “conserved regions” of the alignment as 33% amino acid identity over windows of 25 amino acid alignment positions (including gaps) both upstream and downstream of the intron position, as previously described (Roy and Hartl 2006).

**Comparison with 684 Eukaryotic Orthologous Groups**

We downloaded intron positions and sequences from the 684 sets of orthologous groups (Rogozin et al. 2003) from the NCBI Web site (ftp.ncbi.nlm.nih.gov/pub/koonin/intron_evolution). For each of the 8 species from the original data set, we performed BlastP searches of the 684 proteins against the predicted *T. pseudonana* proteome. Putative *T. pseudonana* orthologs were assigned in cases in which the orthologous gene from all the 8 species had a best hit to the same *T. pseudonana* gene. This yielded 493 putative sets of orthologs. Protein sequences were then aligned in ClustalW using default parameters and intron positions were mapped onto the corresponding alignments. Following Rogozin et al. (2003), we excluded alignment positions that contained gaps in one or more species, as well as flanking alignment positions.

**Estimation of Ancestral Intron Numbers**

We downloaded Csuros’ intron evolution program from his Web site (www.iro.umontreal.ca/~csuros/...
introns/) and applied his method (Csuros 2005) as well as that of Roy and Gilbert (2005a, 2005b) to the 493 sets of eukaryotic orthologs among the 9 species.

Branch Length Estimation

Branch lengths for the eukaryote phylogeny were inferred in PAUP* 4.0b10 (Swofford 2002) under the minimum evolution criterion. The distance matrix was determined under the JTT model of Jones et al. (1992) protein evolution in Splitstree 4.0 (Huson 1998) with gapped sites excluded. Importantly, the relative pattern of branch lengths among the taxa was either maintained or exaggerated when nonhomogenous rates parameters (invariant sites and gamma distribution) were included, regardless of whether the tree topology was constrained or estimated from the data. This was also the case when various alternative models of protein evolution were used.

Results

Phylogenetic Relationships of Species Studied

Figure 1 shows the most likely phylogenetic relationship between the species studied. It is unknown whether plants are more closely related to chromalveolates or to fungamals (fungi + animals) (Sogin 1991; Phillippe et al. 2000; Stechmann and Cavalier-Smith 2002).

Pairwise Comparison of T. pseudonana to 3 Apicomplexan Species

We compared intron positions in conserved regions of alignment between T. pseudonana and available genomic sequences from 3 apicomplexans (the closest relatives to diatoms for which full genome sequences were available): P. falciparum, T. parva, and T. gondii. Only between 1.9% and 4.2% of T. pseudonana intron positions were shared with each species (table 1).

Analysis of 493 Sets of Eukaryotic Orthologs

We next studied 684 sets of eukaryotic orthologs between 8 eukaryotic species previously compiled by Rogozin et al. (2003). We identified putatively orthologous T. pseudonana genes for 493 of the 684 sets. There were 321 T. pseudonana introns in conserved regions of alignment. Only 26/321 T. pseudonana introns (8.1%) were found at positions shared with introns from any of the other 8 species. This is a much smaller fraction than for each of the other 8 species (fig. 2). For instance, a 4 times higher fraction of P. falciparum intron positions (31.6%, 92/291) were shared with another species (this distribution is different from the T. pseudonana distribution at the $P = 8.9 \times 10^{-9}$ level by a Fisher Exact test). Figure 3 summarizes the pattern of intron position sharing with nonchromalveolates for T. pseudonana and P. falciparum.

Level of Intron Loss in T. pseudonana

Comparison of levels of intron loss/gain highlights the discrepancy between T. pseudonana and other lineages. For instance, of the 51 introns shared between P. falciparum and both fungamals and plants (and thus very likely to have

![Diagram](https://academic.oup.com/mbe/article-abstract/24/7/1447/986012/247147896012?download=1)

**Table 1** Summary of Pairwise Comparisons between Thalassiosira pseudonana and 3 Apicomplexan Species

<table>
<thead>
<tr>
<th>Species</th>
<th>Gene Pairs Studied</th>
<th>Apicomplexan T. pseudonana Introns</th>
<th>Shared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theileria parva</td>
<td>1016</td>
<td>1226</td>
<td>397</td>
</tr>
<tr>
<td>Plasmodium</td>
<td>903</td>
<td>339</td>
<td>371</td>
</tr>
<tr>
<td>falciparum</td>
<td></td>
<td></td>
<td>7 (1.9%)</td>
</tr>
<tr>
<td>Toxoplasma</td>
<td>1265</td>
<td>2664</td>
<td>595</td>
</tr>
<tr>
<td>gondii</td>
<td></td>
<td></td>
<td>5 (4.2%)</td>
</tr>
</tbody>
</table>

**Note.**—The “intron” columns gives the number of introns in the apicomplexan species and in T. pseudonana within conserved regions, and the number and percentage of T. pseudonana intron positions that are shared with the apicomplexan species.

![Diagram](https://academic.oup.com/mbe/article-abstract/24/7/1447/986012/247147896012?download=1)
been present in the T. pseudonana–P. falciparum (P. falciparum ancestor), only 4 (7.8%) are shared with T. pseudonana; by contrast, more than a third (4/14) of T. pseudonana introns shared with both fungamals and plants are shared with P. falciparum \( (P = 8.0 \times 10^{-8}) \) by a Fisher Exact test. Assuming a plant–chromalveolate sister relationship, A. thaliana and T. pseudonana are equally closely related to fungamals. However, A. thaliana introns are 4 times more likely to be shared with a fungamal species (32.7% in A. thaliana vs. 7.5% in T. pseudonana). Similarly, 51/80 P. falciparum introns shared with fungamals are present in A. thaliana, a fraction 13 times higher than the fraction shared with T. pseudonana, despite the fact that T. pseudonana is more closely related to P. falciparum.

Shared T. pseudonana Intron Positions Are Broadly Shared

Are the few introns that are shared between T. pseudonana and other eukaryotes ancestral? Intron positions common to T. pseudonana and other species could either be ancestral introns retained in T. pseudonana or could be introns that have been gained in T. pseudonana at positions corresponding to intron positions in the other species (parallel insertion). If the shared T. pseudonana introns are due to parallel insertion, they would not be expected to show any particular phylogenetic distribution. For instance, an otherwise Homo sapiens-specific intron position would be as likely to be shared with T. pseudonana as would an intron position common to many species. If instead the shared T. pseudonana intron positions largely reflect ancestral introns, we would expect the intron positions to be broadly shared, reflecting their ancestral nature.

Among the 4,961 intron positions that are present in at least one of the nondiatom species, 704 (14.2%) are shared between 2 or more broad phylogenetic groups (plants, fungamals, and apicomplexans; table 2 and fig. 4). If the intron positions that are shared between T. pseudonana and other species were attributable to parallel intron gain, we would therefore expect 14.2% of introns to be shared with at least 2 groups. Instead, fully half (13/26) of T. pseudonana introns that are shared with another species are shared with at least 2 groups \( (P = 1.7 \times 10^{-5}) \) by a Fisher Exact test. This pattern is not expected if the shared intron positions between T. pseudonana and other species are due to secondary insertion in

![Diagram showing intron positions shared between T. pseudonana and other species.]

**Table 2**

<table>
<thead>
<tr>
<th>KOG Gene Intron Shared with</th>
<th>KOG Gene Intron Shared with</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chaperonin gamma chain</td>
<td>Chaperonin gamma chain</td>
</tr>
<tr>
<td>Ubiquitin-conjugating enzyme</td>
<td>Ubiquitin-conjugating enzyme</td>
</tr>
<tr>
<td>RNA helicase</td>
<td>RNA helicase</td>
</tr>
<tr>
<td>DNA ligase I</td>
<td>DNA ligase I</td>
</tr>
<tr>
<td>Clathrin heavy chain</td>
<td>Clathrin heavy chain</td>
</tr>
<tr>
<td>Histone H2A</td>
<td>Histone H2A</td>
</tr>
<tr>
<td>Small nuclear ribonucleoprotein F</td>
<td>Small nuclear ribonucleoprotein F</td>
</tr>
<tr>
<td>Vacuolar ATP synthase subunit F</td>
<td>Vacuolar ATP synthase subunit F</td>
</tr>
<tr>
<td>60s Ribosomal protein L36</td>
<td>60s Ribosomal protein L36</td>
</tr>
<tr>
<td>Prefoldin</td>
<td>Prefoldin</td>
</tr>
</tbody>
</table>

**NOTE.**—Pf, At, Sp, Hs, Ce, Ag, and Dm indicate Plasmodium falciparum, Arabidopsis thaliana, Schizosaccharomyces pombe, Homo sapiens, Caenorhabditis elegans, Anopheles gambiae, and Drosophila melanogaster, respectively.

* This intron position is shared only with H. sapiens but falls in the same gene as a broadly distributed intron position.

T. pseudonana and suggesting that at least perhaps half the shared T. pseudonana introns are ancestral retained introns.

Further, the 13 T. pseudonana introns that are shared with multiple groups also show a very high degree of conservation across species within those groups. For instance, whereas only 19.2% (126/656) of all intron positions in the data set that are shared between T. pseudonana and a plant/fungus/apicomplexan (and thus presumably present in the animal ancestor) are retained in Drosophila melanogaster, 6/11 of these introns that are also shared with T. pseudonana are retained in D. melanogaster \( (P = 0.010) \) by a Fisher Exact test. Whereas only 13.3% (90/675) of introns shared between an animal and A. thaliana and/or P. falciparum are retained in S. pombe, fully three-fifths (8/13) of these introns that are also shared with T. pseudonana are retained in S. pombe \( (P = 9.1 \times 10^{-5}) \).

These patterns are consistent with these few ancestral introns having been selectively retained in T. pseudonana because they encode some important function or otherwise
experience lower rates of intron loss. The 13 *T. pseudonana* introns that are shared with species from only one major phylogenetic group may be shared ancestral introns that have been lost from other species or may represent cases of parallel insertion.

Patterns of Shared *T. pseudonana* Intron Positions

The 13 intron positions shared broadly between *T. pseudonana* and other species are not randomly distributed across the 493 studied genes. Instead, 2 genes contain 2 of the broadly shared introns each (table 2, fig. 5). In a 3rd case, one of the broadly shared *T. pseudonana* intron positions is found in the same gene as one of the 13 *T. pseudonana* intron positions that is more narrowly shared (in this case between *T. pseudonana* and *H. sapiens*). This pattern is expected if *T. pseudonana* has largely retained the ancestral intron–exon structure in a small number of genes, but is not predicted from parallel insertion.

Low Rate of Parallel Insertion in *T. pseudonana*

These data inform the more general debate about the incidence of parallel intron insertion (Tarrío et al. 2003; Qiu et al. 2004; Sadusky et al. 2004; Stoltzfus 2004; Csuros 2005; Nguyen et al. 2005; Sverdlov et al. 2005). Central to this debate has been analysis of 488 kb of conserved coding sequence in 684 sets of orthologous genes compiled by Rogozin et al. (2003). The most sophisticated estimates come from 2 independent analyses, which reconstructed intron loss and gain evolution in the data set under the assumption that there was a constant number of fixed possible intron insertion sites. Under this assumption, both analyses estimated that there were approximately 35,000–41,000 possible intron sites (Csuros 2005; Nguyen et al. 2005). Because there are 7,236 total intron positions in the data set, this would predict that approximately 18–21% (7,236 divided by the number of sites) are occupied by an intron in the data set. Therefore, we would expect that 18–21% of new intron insertions would match an existing intron.
intron position from a species in this data set. Instead, we found that only 8.1% of introns in orthologous genes from *T. pseudonana* match intron positions from another species, suggesting that parallel insertion is less frequent than the previous analyses estimated (Csuros 2005; Nguyen et al. 2005). Because some half of these introns are broadly shared intron positions that are likely ancestral (see above), it is very likely that the rate of parallel insertion is lower still. There are 13 shared *T. pseudonana* intron positions that are not broadly shared (and thus are stronger candidates for parallel insertions), and 295 that are specific to *T. pseudonana*. Allowing that all 13 of the not-broadly shared introns represent parallel insertions, this would imply that some 4.2% of new intron insertions match positions of other introns, some 5-fold lower than expected from the previous analyses. We note that these estimates are more in keeping with an earlier analysis by Sverdlov et al. (2005), which suggested that parallel insertion accounted for perhaps 5% of observed shared intron positions.

**Phase of *T. pseudonana* Introns**

Introns may fall in phase 0 (between 2 codons), phase 1 (between the 1st and 2nd bases of a codon), or phase 2 (between the 2nd and 3rd bases of a codon). A wide variety of studied species exhibit an excess of phase 0 introns over introns in the other 2 phases (Fedorov et al. 1992; Long et al. 1995; Ruvinsky et al. 2005). However, it is not known whether this is due to partial retention of a phase-biased ancestral set of introns or to phase-biased intron insertion (Dibb and Newman 1989; Fedorov et al. 1992; Long et al. 1995, 1998; Souza et al. 1998; Long and Rosenberg 2000; Paquette et al. 2000; Fedorov et al. 2001; Roy et al. 2001; Wolf et al. 2001; Lynch 2002; Fedorov et al. 2003; Coglian and Wolfe 2004; Quet al. 2004; Sverdlov, Rogozin, et al. 2004; Roy and Gilbert 2005c; Ruvinsky et al. 2005; Vibranovskii et al. 2005; Ruvinsky and Ward 2006; Nguyen et al. 2006). These questions relate to the fundamental questions about the ultimate origin of introns (e.g., Stoltzfus et al. 1994; Logsdon et al. 1995; Rzhetsky et al. 1997; Roy et al. 1999; Roy 2003; de Souza 2003; Elder 1991, 2000; de Roos 2004; Koonin 2006). Previous studies have shown that putatively recently gained introns also exhibit a bias toward insertion into phase 0, however, in many cases the methods by which these “recently inserted” introns were identified may be unable to confidently distinguish between recently gained and multiply lost introns (Krzywinska and Besansky 2002; Kiontkovski et al. 2004; Roy and Penny 2006b). Because the vast majority of *T. pseudonana* introns appear to have been gained since early eukaryotic evolution, the phase distribution of *T. pseudonana* introns is informative about patterns of intron insertion and the possibility of phase-biased intron gain.

Among the entire set of *T. pseudonana* genes, there were 3,394 (31.5%) phase 0 introns, 4,296 (39.9%) phase 1 introns, and 3,081 (28.6%) phase 2 introns (these are different at a $P < 10^{-5}$ level by a chi-square test). Thus, although *T. pseudonana* shows a phase bias, supporting the notion of biased intron insertion, the pattern does not follow the general eukaryotic pattern in which the largest number of introns are in phase 0. Nonetheless, these data support the notion that introns may insert differentially into the 3 phases, in turn supporting the notion that the observed phase zero bias in other species could largely or completely reflect insertional biases.

**Rates of Sequence and Intron Evolution**

What explains the higher rates of intron loss and gain evolution in the *T. pseudonana* lineage? One possibility is that high rates of intron loss/gain in *T. pseudonana* simply reflect a generally higher rate of molecular evolution (a possibility suggested by an anonymous reviewer). We investigated this possibility by looking at 2 sources of data. First, we used unappapped positions in the 493 sets of orthologs among 9 species. Using standard models of protein sequence evolution, we estimated branch lengths (see Supplementary Material online). The estimated length of the branch leading to *P. falciparum* was roughly 1.5 times longer than that leading to *T. pseudonana*. Thus, if these estimates are accurate, rates of protein evolution in the lineage leading to *T. pseudonana* have been slower than in that leading to *P. falciparum*, not faster.

However, branch length estimation for such deeply diverged species with such sparse sampling of taxa poses a host of difficulties. Therefore, we undertook a survey of a variety of previous phylogenetic studies that included both taxa, analyzing branch lengths from published trees (e.g., Budin and Philippe 1998; Saldarriaga et al. 2001; Bapteste et al. 2002; Fast et al. 2002; Kuvaridina et al. 2002; Harper et al. 2005). This literature survey showed no clear bias toward longer branch length in either apicomplexans or stramenopiles, even in instances where the much greater sampling density of stramenopiles is expected to lead to artificially longer estimated branch lengths. Thus, although the long evolutionary distances involved still make it difficult to make conclusive statements about relative sequence rate evolution, there does not seem to be evidence that the rapid rate of intron loss/gain evolution in *T. pseudonana* corresponds to a generally rapid rate of sequence evolution.

In general we see no a priori reason that rates of intron loss/gain evolution should correlate with rates of protein evolution. The 2 types of evolution are dependent on entirely different set of mutations (recombination with reverse-transcribed mRNAs for intron loss and possibly intron gain, possibly repetitive element insertion for intron gain, vs. point mutation for sequence evolution) and are likely to be governed by fundamentally different selective forces. Nonetheless, previous results have found that metazoan lineages that have experienced high rates of intron loss have also experienced high rates of protein sequence change (Raihle et al. 2005). In addition, many of the eukaryotic lineages that were originally placed at the base of the eukaryotic tree apparently in part due to long-branch attraction arising from high rates of sequence change along these branches also have experienced high levels of intron loss (see, for instance, Logsdon 1998). The correspondence of sequence evolution to other types of genome evolution including intron loss/gain clearly requires further exploration.
Discussion

We show that the vast majority of introns in *T. pseudonana* have been gained since early eukaryotic evolution, unlike previously studied species. In all previously studied intron-rich eukaryotes, significant fractions of the intron positions are shared with distantly related species, implying that most of these introns have been retained from early eukaryotic ancestors (Rogozin et al. 2003; Slamovits and Keeling 2006). Here, we show that the diatom *T. pseudonana* constitutes a deviation from this pattern. In pairwise comparisons, no more than 4.2% of *T. pseudonana*’s introns were shared with each of 3 other chromalveolate species, the apicomplexans *P. falciparum*, *T. parva*, and *T. gondii*. Analysis of 493 sets of eukaryotic orthologs among plant, animal, fungal, and apicomplexan species further showed that only 8.1% of *T. pseudonana* introns in conserved coding regions are shared with any of the other species. Of these, half are broadly shared across widely diverged eukaryotic groups, suggesting that they are retained ancestral introns.

Large-Scale Intron Loss and Gain in *T. pseudonana*

Both *T. pseudonana* and *P. falciparum* belong to the broad phylogenetic group chromalveolates. In all, 31.3% of *Plasmodium falciparum*’s intron positions in conserved regions of 493 sets of putatively orthologous genes are shared with a nonchromalveolate species, most of which were presumably present in the chromalveolate ancestor. Of these only 5% (5/91) are also shared with *T. pseudonana*, thus *T. pseudonana* has apparently lost the vast majority of the ancestral chromalveolate introns.

On the other hand, only 8.1% of *T. pseudonana* introns are shared with any of the other species. That 31.3% of *P. falciparum* intron positions are shared with a nonchromalveolate implies that a significant number of introns present in the chromalveolate ancestor have both 1) been retained in *P. falciparum* (or more generally in apicomplexans) and 2) are represented in nonchromalveolates. If a significant fraction of *T. pseudonana* introns were present in the chromalveolate ancestor, we would similarly therefore expect a significant fraction to be shared with both *P. falciparum* and studied nonchromalveolate species. That only a small fraction of *T. pseudonana* introns are shared with nondiatomic thus strongly suggests that they have been largely gained since the chromalveolate ancestor. Thus, the current data indicates that *T. pseudonana* has experienced a large amount of both intron loss and gain since the chromalveolate ancestor.

Transition from Intron-Poor to Intron-Gain Genomes?

All previously studied intron-rich eukaryotic genomes share significant fractions of their intron positions with widely diverged species. This both implies that a large fraction of their introns are retained from early eukaryotic ancestors and, thus, that these lineages have never passed through a very intron-poor phase. Similarly, all known very intron-poor eukaryotic species are now known to be related to much more intron-rich taxa (which have retained large numbers of ancestral introns) and thus are known to be very intron poor due to nearly complete secondary loss of ancestral introns. These previous results imply that among previously studied taxa, “intron richness” arose only once (though subsequent fluctuation in intron number through intron-rich lineages is possible).

The current results open up the possibility of a 2nd emergence of an intron-rich genome from an ancestrally intron-poor genome. There are 2 possibilities for the history of intron number in the *T. pseudonana* lineage (fig. 6). The lineage could have remained relatively intron rich since the chromalveolate ancestor, with ongoing intron loss and gain gradually replacing the ancestral introns with more recently gained introns. Alternatively, the lineage could have experienced massive loss of ancestral introns leading to a nearly intronless genome (as implied for various known nearly intronless modern taxa [Rogozin et al. 2003]) with subsequent widespread intron insertion leading to modern intron richness. Further study of additional diatom species (some of which have ongoing genome projects) should help to answer this question.

It is indeed striking that all previously studied lineages that have lost all of their ancestral introns remain very intron poor, suggesting that in most cases the transition from an intron-rich to intron-poor genome is not easily reversible. Why should this be? One possibility is that large-scale intron loss enables changes in the spliceosome which prevent the spliceosome from recognizing new gene-interrupting insertions. In this case, the reduced spliceosomal flexibility...
would cause intron-poor taxa to experience low levels of successful intron insertion and thus to remain intron poor.

Intron-Rich Ancestral Eukaryotes

Previous analyses of the Rogozin et al. data set have disagreed as to the ancestral density of introns in early eukaryotic ancestors and as to the relative importance of subsequent intron loss and gain (reviewed in Rodriguez-Trelles et al. 2006b). Roy and Gilbert (2005a, 2005b) estimated that the fungal ancestor and the plant–fungamal ancestor had nearly as many introns as the most intron-rich modern species and that subsequent eukaryotic evolution has been dominated by an excess of intron loss over intron gain. By contrast, Csurös (2005) and Nguyen et al. (2005) independently concluded that ancestral intron densities were more modest and that the history of intron number had been more mixed, with some lineages experiencing dramatic increases in intron number (more gain than loss) while others experienced equally dramatic decreases (more loss than gain).

The difference in conclusions between these analyses reflects their different assumptions about the frequency of parallel insertion. Roy and Gilbert (2005a, 2005b) assumed that all intron positions that were shared between species reflected ancestral introns (essentially assuming that parallel insertions were infrequent enough to be ignored), whereas Csurös (2005) as well as Nguyen et al. (2005) assumed that there were a certain number of possible intron insertion sites, which (presumably to make the problem more tractable) were assumed to be constant through evolutionary history. The various impacts of the models’ assumptions have not been thoroughly tested.

These results provide an opportunity to put an upper limit on the importance of parallel intron gain on intron evolution. In all 4.2% (13/307) putative intron insertions in T. pseudonana coincide with intron positions from another species. Following the previous models in assuming a constant number of possible intron insertion sites, this would suggest that the 7,236 intron positions represented across the 8 studied species (excluding T. pseudonana) occupy around 4.2% of the possible sites, suggesting that there are roughly 171,000 “possible intron insertion sites” out of the 488 kb of studied sequence. Even assuming that all 26/320 shared T. pseudonana introns represent parallel insertions suggests roughly 89,000 sites (= 320/26 × 7236), still more than twice the previous estimates.

Underestimating the total number of sites will lead to overestimation of the probability that 2 inserted introns will occupy the same site (parallel insertion), presumably leading to underestimation of the number of truly ancestral introns. We used these directly estimated numbers of “possible intron sites” to reconstruct intron evolution of the 684 groups of eukaryotic orthologs using the method of Csurös. The results are shown in figure 7. For each ancestral node, the estimated number of ancestral introns is much closer to that previously estimated by Roy and Gilbert (2005b) than to that estimated by Csurös (2005) or Nguyen et al. (2005). This suggests that the methods of Csurös and Nguyen et al. may have greatly overestimated the importance of parallel insertion, and that Roy and Gilbert’s method, which ignores parallel insertion, may give more accurate estimates.

Conclusions

We show that, unlike all previously studied intron-rich eukaryotic species, the vast majority of introns in the intron-rich diatom T. pseudonana are found at unique positions, suggesting that T. pseudonana has lost nearly all of its ancestral introns and gained a large number of new introns. These results open up the possibility that genomes rich in spliceosomal introns may have arisen multiple times in eukaryotic evolution, however, further study of other diatom species will be necessary. The finding of a very low level of parallel insertion supports the notion that early eukaryotic ancestors were very intron rich and thus furthers the idea that intron density was very high early on in eukaryotic evolution and that most (or even all) eukaryotic lineages have since experienced a reduction in intron number.

Supplementary Material

Supplementary material is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Manuel Irimia for helpful discussions and constructive comments, Jason Stajich for help with the
ancient intron density estimates, and Matt Phillips for phylogenetic analyses, branch length estimation, and for general wise pronouncements.

Literature Cited


Fedorov A, Fedorova L. 2006. Where is the difference between the genomes of humans and anelids? Genome Biol. 7:203.


Arndt von Haeseler, Associate Editor

Accepted February 26, 2007