Widespread Intron Loss Suggests Retrotransposon Activity in Ancient Apicomplexans

Scott William Roy* and David Penny*

*Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

Several facets of spliceosomal intron in apicomplexans remain mysterious. First, intron numbers vary across species by 2 orders of magnitude, indicating massive intron loss and/or gain. Second, previous studies have shown very different evolutionary patterns over different timescales, with 100-fold higher rates of intron loss/gain between genera than within genera. Third, the timing and dynamics of nearly complete intron loss in Cryptosporidium species, as well as reasons for retention of the few remaining introns, remain unknown. We compared intron positions in 785 orthologous genes between 3 moderate to intron-rich apicomplexan species. We estimate that the Theileria-Plasmodium ancestor had 4.5 times as many introns as modern Plasmodium species and 38% more than modern Theileria species, and that subsequent intron losses have outnumbered intron gains by 5.8 to 1 in Theileria and by some 56 to 1 in Plasmodium. Several patterns suggest that these intron losses occurred by recombination with reverse-transcribed mRNAs. Intriguingly, this finding suggests significant retrotransposon activity in the lineages leading to both Theileria and Plasmodium, in contrast to the dearth of known retrotransposons and intron loss within modern species from both genera. We also compared genomes from Cryptosporidium parvum and C. hominis and found no evidence of ongoing intron loss, nor of intron gain. By contrast, Cryptosporidium introns are less evolutionary conserved with Toxoplasma than are introns from other apicomplexans; thus the few remaining introns are not simply indispensable ancestral introns.

Introduction

Large-scale sequencing of a broad diversity of eukaryotic species has revealed pronounced differences in a variety of features: genome size, gene number, activity and number of transposable elements, degree of alternative splicing, and the density and lengths of spliceosomal introns. The increasingly broad phylogenetic sampling of sequenced species has also revealed the lack of a simple relationship between these various aspects of genome organization. While some species’ genomes seem quite generally “simple” (Encephalitozoon cuniculi has a compact genome with few genes and very few short introns; Katinka et al. 2001) or “complex” (Homo sapiens has a 3-Gb genome, over 20,000 genes, more than 8 introns per gene on average, extensive alternative splicing, and abundant transposable elements derived sequence; Lander et al. 2001), other species defy such clear characterization. For instance, the Bigelowiaella natans nuclemorph, an enslaved algal nucleus, represents an extreme of genomic compaction in most genomic features but still retains more than 3 introns on average per gene (Gilson and McFadden 1996; Gilson et al. 2006). The genome of Takifugu rubripes contains numerous introns but relatively few transposable elements; the genome of Oryza sativa contains abundant transposable elements but has generally short introns (Aparicio et al. 2002; Yu et al. 2002). The explanations for these patterns remain obscure.

Here we further investigate an example of partial genome compaction in the phylum Apicomplexa. Apicomplexans are single-celled parasites of animals with complex life cycles. The large numbers of introns in most species are perhaps surprising in light of the generally reductive mode of evolution observed in various parasitic lineages (Katinka et al. 2001; Gardner et al. 2002, 2005; Abrahamsen et al. 2004; Sakharkar, Dhar, and Chow 2004). While species of the apicomplexan genus Cryptosporidium have a very small number of introns (only around 5% of genes in the fully sequenced C. parvum genome contain introns; Abrahamsen et al. 2004), in keeping with a variety of other unicellular parasites (Giardia lamblia, Trypanosoma sp., Trichomonas vaginalis, Encephalitozoon cuniculi), most characterized apicomplexan lineages exhibit moderate to high intron densities (roughly 1, 2, and 5 introns per gene on average in Plasmodium and Theileria species and in Toxoplasma gondii, respectively; compiled in Jeffares, Mourier, and Penny 2006; Roy and Gilbert 2006).

We previously showed mysterious contrasting patterns of intron evolution in apicomplexans over different timescales. Rates of intron loss/gain are extremely low within genera (less than 1.5% intron loss/gain over ~100 My for both Plasmodium and Theileria), but are 100- to 1,000-fold higher between genera (only one-sixth of intron positions are shared between Plasmodium and Theileria; Roy and Hartl 2006; Roy and Penny 2006). At a deeper evolutionary level only one-third of Plasmodium falciparum intron positions are shared with surveyed nonapicomplexans (Rogozin et al. 2003).

A second evolutionary mystery involves near-complete genome-wide intron loss, a surprisingly common occurrence in diverse eukaryotic lineages. At least roughly one-third of P. falciparum introns are ancestral to apicomplexans (intron positions shared with nonapicomplexans; Rogozin et al. 2003), indicating near-complete intron loss in Cryptosporidium. Analogous cases can be made for each nearly intronless eukaryotic lineage (Rogozin et al. 2003; Slamovits and Keeling 2006). The dynamics of these dramatic reductions are not well understood: first whether loss in these lineages is episodic or ongoing, and second why these species retain their few remaining introns.

We report 2 studies of intron position conservation in apicomplexans. First, we compared orthologous gene trios between P. falciparum, T. parva, and T. gondii. We estimate that the Plasmodium-Theileria ancestor had 4.5 times as many introns as P. falciparum and 38% more introns than T. parva. We further estimate that losses have outnumbered gains by 5.8 to 1 in T. parva, and by a striking 56 to 1...
in *P. falciparum*. Nearly all observed intron losses are exact, occurring without loss/gain of adjacent coding sequences. Intron losses are 3’ biased, and loss of introns that are adjacent along the gene is more common than expected by random single loss. These patterns suggest intron loss by recombination with reverse transcribed copies of mRNAs. This in turn suggests significant retrotransposon activity in lineages leading to *Plasmodium* and *Theileria*, whereas the dearth of intron loss (and retrotransposons) within those genera suggest more recent, convergent (near) extinction of retrotransposons in both lineages.

Second, we report a genome-wide comparison of introns in *Cryptosporidium* species. We find no intron loss or gain since the *C. parvum-C. hominis* divergence (Abrahamsen et al. 2004; Xu et al. 2004). However, over much longer evolutionary timescales, *C. parvum* intron positions are relatively not highly conserved: only 36.7% of *C. parvum* intron positions are shared with *T. gondii*, less than for more intron-rich *P. falciparum* (71.6%) and *T. parva* (70.6%). Thus the small subset of introns present in *Cryptosporidium* species do not simply represent a broadly evolutionary-conserved set of functionally important ancestral introns.

These results indicate that the relatively high intron densities among parasites found in *Plasmodium* and *Theileria* species in fact represent reductions from the ancestral genomes of ancestors (which were also parasitic), and that there is no evidence for significant ongoing intron loss in *Cryptosporidium*. We discuss the implications for the evolutionary forces controlling genome structure.

**Material and Methods**

We downloaded the annotated genome sequences for *T. parva*, *C. parvum*, and *C. hominis* (Genbank accession numbers AA0K0100001.1, AAEE01000000, and AAEL0100001.1, respectively) and for *P. falciparum* (http://www.plasmodb.org, version 10.03.2002.v2) and *T. gondii* (http://www.toxodb.org). Putatively orthologous gene pairs/trios were determined by reciprocal BLASTP searches. Intron positions were mapped onto ClustalW alignments and computationally filtered. Intron positions within 5 codons were considered to be conserved (which applied in less than 3% of cases for the *T. gondii-T. parva-P. falciparum* alignments). Conserved positions were conservatively defined as 50% conservation in the flanking 10 amino acids between each pair of species, as previously described (Roy and Hartl 2006). Alignments were further analyzed by eye. The programs used are available from the authors.

Ancestral intron numbers and numbers of subsequent intron losses and gains were estimated using previously described methods (Roy and Gilbert 2005b, c). Given the small number of species, other methods that attempt to account for parallel insertion are not applicable (Csuros 2005; Nguyen et al. 2005; Carmel et al. 2007). Fortunately, parallel gains are very unlikely to be a significant issue in the lineages studied. First, we estimate only 15 gains in *Plasmodium* and 65 gains in *Theileria* over 785 genes. Assuming equal rates of insertion across genes, we therefore expect only around 1 gene to have experienced a gain in both lineages (15 × 65 / 785), and even assuming very restricted insertion, the chance that both gained introns would insert into the identical site is very small. Second, using greater taxonomic sampling, Nguyen and coauthors (2007) reached similar conclusions about relative rates of loss and gain in these lineages while attempting to account for parallel intron gain. As such, we are confident that our estimates are not significantly affected by parallel gain.

For the *C. parvum-C. hominis* comparison, we performed BLASTN searches of coding sequences from intron-containing *C. parvum* genes against the *C. hominis* genome, and evaluated intron presence/absence on the presence/absence of an alignment gap. Of the 180 annotated *C. parvum* introns, 3 were discarded due to implausibly short lengths (<5 bp), and 1 was excluded because proximity to the 5’ of the gene made intron presence/absence impossible to ascertain. Orthologous sequence was identified for 164/176 of the remaining introns, all of which showed intron presence.

**Results**

**Intron Loss and Gain in Apicomplexans**

To characterize the relative contributions of intron loss and gain in the evolutionary history of *Plasmodium* and *Theileria*, we compared intron positions in trios of orthologous genes between relatively intron-poor *Plasmodium falciparum*, more moderate *Theileria parva*, and very intron-rich outgroup *Toxoplasma gondii*. Within highly conserved regions of 785 ortholog trios, there were 249 introns in *P. falciparum*, 814 in *T. parva*, and 1,331 in *T. gondii*. In total, there were 1,604 intron positions. The relationship between apicomplexans with fully sequenced genomes used in this and other studies is given in figure 1.

Some genes showed very highly conserved intron-exon structures across the species (fig. 2). Others showed several introns that were shared between the outgroup *T. gondii* and 1 (but not the other) of the other 2 species, suggesting intron loss (fig. 3). In total, 7.4% of intron positions were shared across all 3 species, 34.4% were shared between exactly 2 species (82.6% of these were present in both outgroups).

![Fig. 1.—Relationship between studied species. Divergence times are drawn from a variety of heterogeneous analyses and methods of inference (Escalante and Ayala 1995; Roy and Penny 2006; current results), and should be considered very approximate.](https://academic.oup.com/mbe/article-abstract/24/9/1926/2925641)
T. gondii-T. parva introns), and 58.2% were species-specific (74.6% in T. gondii, only 3.6% in intron-poor P. falciparum; fig. 4A).

We used previously described methods (Nielsen et al. 2004; Roy and Gilbert 2005b, c) to estimate the number of introns in the analyzed regions for the T. parva-P. falciparum ancestor and the subsequent numbers of intron losses and gains (fig. 4B). The T. parva-P. falciparum ancestor is estimated to have had 1,227 introns in analyzed regions, which is 38% more introns than T. parva and 4.5 times as many introns as P. falciparum. P. falciparum is estimated to have experienced 16 intron gains and 893 intron losses since the Plasmodium-Theileria divergence, or 56 times more losses than gains. T. parva is estimated to have experienced 5.8 times as many losses as gains (378 vs. 65). Of the 753 introns that are species-specific between P. falciparum and T. parva, only around 11% (81/753) are estimated to be due to intron gain. Thus intron loss has been the dominant mode of evolution since the Theileria-Plasmodium divergence.

Evidence for Reverse Transcriptase–Mediated Intron Loss

Three predictions of reverse transcriptase (RT) mediated intron loss were fulfilled among the observed intron losses in P. falciparum and T. parva (i.e., positions at which introns are present in 2 out of 3 species). First, intron losses are biased toward the 3‘ end of the gene. Among 26 genes with at least 1 three-way-conserved intron position as well as at least 1 loss in P. falciparum, 18 showed a bias toward 3‘ loss (P = 0.038; see Roy and Gilbert 2005a for a thorough discussion). For T. parva losses, 6/8 genes showed 3‘-biased loss (not statistically significant). Second, genes experiencing multiple intron losses showed a bias toward loss of adjacent introns. For all 10 genes with at least 1 intron position conserved across all 3 species and at least 2 intron losses in P. falciparum, the lost introns were adjacent relative to the conserved introns (P = 0.0015; see Roy and Gilbert 2005a for a thorough discussion). For T. parva multiple losses were adjacent in 1/3 cases (not statistically significant). Third, 99% (451/456 in P. falciparum; 59/60 in T. parva) of observed intron losses were exact (not associated with coding sequence alignment gaps between intron-containing and -lacking homologs), as expected by RT-mediated intron loss.

No Intron Loss or Gain Between C. parvum and C. hominis

We also studied intron evolution in Cryptosporidium. All 164 predicted C. parvum introns in conserved regions of C. parvum and C. hominis were present in C. hominis, the closest relative of C. parvum. No intron loss or gain was detected in the C. parvum-C. hominis alignments.
were found to be present in closely related *C. hominis*; thus there has been no intron loss in *C. hominis* nor intron gain in *C. parvum* since the species’ divergence. Silent site divergence between *C. parvum* and *C. hominis* is approximately 2% (Xu et al. 2004), suggesting a date of divergence of roughly 2 Mya (assuming the estimated divergence rate in *Plasmodium* of around 5/10^9 per year; Neafsey, Hartl, and Berriman 2005; Tanabe et al. 2004). Lack of intron gain in 2 My in 3,396 *C. parvum* genes suggests a rate of less than 1 intron gain per gene per 7 By, within the range of estimates for other eukaryotic lineages (e.g., Roy, Fedorov, and Gilbert 2003; Nielsen et al. 2004; Lin et al. 2006; Stajich and Dietrich 2006).

It is not known why nearly intronless species retain their few remaining introns. One possibility is that retained introns encode essential functions (e.g., Jeffares, Mourier, and Penny 2006), in which case we might expect them to be highly conserved evolutionarily. Interestingly, of 109 *C. parvum* intron positions in conserved regions, a much smaller fraction (36.7%) were shared with *T. gondii* than for either of the more intron-rich species (70.6% in *T. parva*, 71.6% in *Plasmodium*; *P* < 10^-11 by a Fisher Exact test), even when *C. parvum* introns that fell near (within 5 codons of) *T. gondii* intron positions were considered shared (51.4%; *P* = 1 × 10^-3). An example of a highly diverged *C. parvum* intron-exon structure is given in figure 5. Thus, *C. parvum* introns are less evolutionarily conserved than are introns in other apicomplexans.

**Discussion**

**Intron Loss/Gain Evolution in the Evolutionary History of *Plasmodium* and *Theileria***

We report large-scale conservation of intron positions in *Plasmodium* and *Theileria* with the outgroup *T. gondii* and pronounced excesses of intron loss over intron gain in both *Plasmodium* and *Theileria*. For both *P. falciparum* and *T. parva*, more than 70% of introns in conserved regions were also present in *T. gondii*; thus the vast majority of introns in these species predate the *Plasmodium-Theileria* divergence. The *Plasmodium-Theileria* ancestor was more intron rich than species from either genera, with an estimated 4.5 times as many introns as *P. falciparum* and 38% more introns than *Theileria*. Overall, we estimate that intron losses have outnumbered gains in the lineage leading to *T. parva* by some 5.8 to 1, and in *P. falciparum* by an astounding 56 to 1. These results are in good qualitative agreement with a previous analysis based on a smaller number of genes, which estimated that the ancestor was 5.5 times as intron rich as *P. falciparum* and had 30% more introns than *T. parva*, with a pronounced excess of losses in *Theileria* and an overwhelming excess in *Plasmodium* since the...
Evidence for Retrotransposons in Ancient Apicomplexans

The pattern of intron loss in *Plasmodium* and *Theileria* strongly suggests intron loss by recombination with reverse transcribed copies of spliced mRNAs. Most strikingly, the vast majority of intron losses in both lineages are exact, occurring without addition or deletion of adjacent coding sequence. This pattern is expected based on reverse-transcribed mRNA-mediated loss, but is not expected based on simple genomic deletion (Roy and Gilbert 2005a). Second, intron losses in both lineages tend to be 3’ biased, consistent with the 3’ bias of reverse-transcribed products (Mourier and Jeffares 2003; Sverdlov et al. 2004; Roy and Gilbert 2005a). Finally, introns in genes experiencing multiple losses in the same lineage tend to fall adjacent to each other along the gene, consistent with loss of multiple adjacent introns via double recombination involving non-adjacent exons (Niu, Hou, and Li 2005; Stajich and Dietrich 2006). In total, these patterns provide strong evidence for the presence and activity of retrotransposons in the ancestry of *Theileria* and *Plasmodium*. These findings are striking in light of the (near) absence of these elements in modern lineages (Nene, Morzaria, and Bishop 1998; Gardner et al. 2002; Figueiredo et al. 2005; Durand, Oelofse, and Coetzer 2006; S.W. Roy and D.L. Hartl unpublished results), as reinforced by the near absence of intron loss in the recent history of both genera (Roy and Hartl 2006; Roy and Penny 2006). The implied evolutionary history of these genera is shown in figure 6.

A Possible Cause for Retroelement Reduction in *Plasmodium* and *Theileria*

What explains the apparent reduction in retrotransposons in *Plasmodium* and *Theileria*? Theory has shown that transposable elements are expected to proliferate so long as the selection against a transposable element is less than the rate of transposition (Charlesworth, Sniegowski, and Stephan 1994). Selection against a transposable element may then be divided into such components as selection against the presence of the physical transposable element sequence in the genome, selection against ectopic recombination between transposable element copies, and selection against additional transpositions (Charlesworth, Sniegowski, and Stephan 1994).

For most species, this last factor may be rather minimal. For species with large fractions of nonfunctional sequence, most newly inserted transposable element copies are likely to have little functional consequence. In addition, for species with frequent outcrossing, meiotic recombination will rapidly dissociate the progenitor copy from a deleterious descendent insertion, diminishing the total selective effect on the progenitor copy (a more complete treatment will be presented elsewhere; Hickey 1982). However, for species with compact genomes and with infrequent outcrossing, such selection may be sufficient to prohibit proliferation of transposable elements.

*Plasmodium* and *Theileria* species may fit both of these criteria. *Plasmodium* species, for instance, undergo 1 mitotic event (giving rise to dozens of individuals from a single parent) every few days, but undergo only 1 meiotic event perhaps every 6 months. Thus there is only 1 sexual recombination event per 100–200 asexual “generations.” Moreover, due to the life cycle of the parasite, meiotic events often involve individuals of like genotype, which will not dissociate the progenitor transposable element locus from a descendent copy. In addition, *P. falciparum* has high gene density, with 9.7 Mb of predicted coding sequence out of a total genome sequence of 22 Mb (Gardner et al. 2002). Conditions are similar for *Theileria*. Including functional non-protein-coding sequences, likely more than half of *Plasmodium* genome is composed of functional sequence.

One possible explanation for the reduction of retrotransposons (activity) in *Plasmodium* and *Theileria*, then, could be reduction in noncoding sequences. Apicomplexans, like many intracellular parasites, have undergone genomic reduction of various kinds, including gene loss and intron loss (Keeling 2004). If in addition both *Plasmodium* and *Theileria* have undergone reduction in nonfunctional
sequences (or more precisely, to have increased the fraction of their sequence which is functional), this change could have moved both genera over the threshold to nonproliferation of transposable elements. This could have led to a stark reduction in retrotransposon activity and, thus, to the observed near-cessation of intron loss. Future theoretical work should explore this possibility.

Causes and Effect in Genome Structure Evolution

These findings underscore the complex relationships between the evolution of different genomic structures. In this case, genomic reduction in 1 kind of noncoding element, retrotransposons, appears to have led to the retention (or recent lack of loss) of another kind of noncoding element, introns, likely due to reduction in the rate of retrotranscriptase-mediated intron loss mutations (Roy and Gilbert 2006; Roy and Hartl 2006; Roy and Penny 2006). This is a result of the web of dependency between elements in genome evolution: (1) retroelement activity is important for intron loss and perhaps intron gain, while intron sequences represent important sites for transposable element insertions; (2) recombination is necessary for intron loss, as well as for transposable element propagation; (3) transposable elements are an important source of intergenic and intronic sequences and are dependent on these noncoding sequences for propagation. These interconnections underscore the difficulties with a simple notion of “genomic complexity.”

The Causes of Intron Evolution

Following our previous arguments (Roy and Gilbert 2006; Roy and Hartl 2006; Roy and Penny 2006), we have put forward an essentially mutation-based explanation for observed heterogeneity of rates. This is in contrast to the suggestion of Nguyen, Yoshihama, and Kenmochi (2007), who suggested that periods of rapid intron loss/gain evolution may be associated with evolutionary transitions. While we find this an interesting possibility, we do not see much evidence for it currently. The previous authors’ data essentially shows a large number of gain/loss events over most of the tree, with the exception of 3 periods within apicomplexan evolution: recent Plasmodium and Theileria evolution and 1 deep branch. The vast majority of the branches studied by Nguyen and coauthors are very long with respect to intron loss/gain – of the outgroups to the Toxoplasma-Theileria divergence, they estimate that between 56% and 98% of ancestral introns have been lost along each branch. Under such circumstances it will be very difficult to accurately estimate loss/gain rates, particularly given uncertainties associated with divergence date estimates. Indeed, we have recently shown that a previous study that provided evidence for an association between intron loss and gain and epochs of rapid evolutionary change may have been unable to accurately reconstruct intron loss/gain due to the presence of very long branches in their sample (Roy and Penny 2007). Further work should seek to test this intriguing idea.

Intron Evolution in Cryptosporidium

We also report the first study of intron loss/gain evolution in nearly intronless Cryptosporidium species. We find no evidence for ongoing intron loss (no loss in 2 My, although the short evolutionary time makes firm conclusions difficult). We also find no intron gains in 2 My, suggesting a rate less than 1 gain per gene per 7 By.

By contrast, we find low levels of conservation of C. parvum introns with intron-rich T. gondii. An increasing number of instances of near-complete intron loss in eukaryotes have been documented; however not all eukaryotes have yet been found. It is not known why nearly intronless species retain their few introns. One possibility is that the few remaining introns represent a subset of...
functionally important introns whose loss is opposed by selection. That *C. parvum* introns are not highly evolutionarily shared suggests against this possibility.

How may this result be explained? Since both intron loss and gain are likely to be dependent on transposable element activity, rates of intron loss and gain might be expected to be directly correlated. In this case, lineages such as *Cryptosporidium* which have experienced a high degree of loss might also have experienced a relatively high rate of gain, leading to a generally lower level of intron conservation than for other lineages.

Conclusions
We report genome-wide analyses indicating: (1) a high degree conservation of *Theileria* and *Plasmodium* intron positions with the coccid *T. gondii*, (2) a striking dominance of intron loss over gain in *Theileria* and *Plasmodium* since their divergence, (3) complete conservation of *C. parvum* introns in *C. hominis*, and (4) low levels of conservation of *C. parvum* introns in *T. gondii*. These results reinforce the notion of intron number reduction as a very common evolutionary pattern and focus attention on the intriguing patterns of intron evolution in the many nearly intronless eukaryotic lineages.

Literature Cited
Durand PM, Oelofse AJ, Coetzer TL. 2006. An analysis of intron evolution in the many nearly intronless eukaryotic lineages.
Gilson PR, McFadden GI. 1996. The miniaturized nuclear genome of eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed, and the smallest known spliceosomal introns. Proc Natl Acad Sci USA. 93:7737–7742.


Geoffrey McFadden, Associate Editor

Accepted May 17, 2007