Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution

Scott William Roy and David Penny

Genome Res. 2006 16: 1270-1275
Access the most recent version at doi:10.1101/gr.5410606

Letter

Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution

Scott William Roy[1] and David Penny

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

The age of modern introns and the evolutionary forces controlling intron loss and gain remain matters of much debate. In the case of the apicomplexan malaria parasite *Plasmodium falciparum*, previous studies have shown that while the positions of two thirds of *P. falciparum* introns are not shared with surveyed non-apicomplexans (leaving open the possibility that they were relatively recently gained), 99.1% are shared with *Plasmodium yoelii*, which diverged from *P. falciparum* at least 100 Mya. We show here that 60.6% of *P. falciparum* intron positions in conserved regions are shared with the distantly related apicomplexan *Theileria parva*, whereas only 18.2% of introns in the more intron-rich *T. parva* are shared with *P. falciparum*. Comparison of 3305 pairs of orthologous genes between *T. parva* and *Theileria annulata* showed that 7089/7111 (99.7%) introns in conserved regions are shared between species. These levels of conservation imply significant differences in rates of intron loss and gain through apicomplexan history. Because transposable elements (TEs) and/or (often TE-encoded) reverse transcriptase are implicated in models of intron loss and gain, the observed low rates of intron loss and gain in recent *Plasmodium* and *Theileria* evolution are consistent with the lack of known TE in those groups. We suggest that intron loss/gain in some eukaryotic lineages may be concentrated in relatively short episodes coincident with occasional TE invasions.

Eukaryotic species vary dramatically in average number of introns per gene, from less than 0.2 intron/gene in *Cryptosporidium* species, hemiascomycetes fungi, *Encephalitozoon cuniculi*, red algae, and *Giardia lamblia* to more than one per gene in animals, most characterized fungi, most apicomplexans, amoebae, diatoms, paramecia, jakobids, land plants, and green algae (compiled in Jeffares et al. 2006; Roy and Gilbert 2006). Such a pattern clearly implies recurrent episodes of massive intron loss and/or gain, although the relative importance of these two processes remains hotly debated (Babenko et al. 2004; Qiu et al. 2004; Csuros 2005; Roy and Gilbert 2005b,c).

Many studies have shown a complex pattern of intron position sharing between species (e.g., Perler et al. 1980; Dibb and Newman 1989; Moriyama et al. 1998; Fedorov et al. 2002; Guiliano et al. 2002; Kent and Zahler 2000; Rogozin et al. 2003; Roy et al. 2003; Kiontke et al. 2004; Nielsen et al. 2004; Vanacova et al. 2005; Roy and Hartl 2006). In some cases, the vast majority of introns are found at identical positions of orthologous genes. For instance, only 15 human-specific introns were found among over 10,000 intron positions in 1560 human–mouse ortholog pairs (Roy et al. 2003). In other cases, intron positions have diverged significantly over relatively short times (Seo et al. 2001; Kent and Zahler 2000; Edvardsen et al. 2004).

The apicomplexan malaria parasite *Plasmodium falciparum* provides a particularly interesting case. In a study of 684 sets of orthologs between *P. falciparum* and seven non-apicomplexan eukaryotic species, only one third of *P. falciparum* intron positions were shared with another species, less than was found for any non-apicomplexan species, leaving open the possibility that the majority of *P. falciparum* introns have been relatively recently gained (Rogozin et al. 2003). By contrast, a study of 3789 pairs of orthologs between *P. falciparum* and the rodent parasite *P. yoelii* (diverged ≥100 Mya) found very high conservation, with 99.1% of *P. falciparum* introns shared with *P. yoelii* (and 99.6% of *P. yoelii* introns shared with *P. falciparum*), and at least three quarters and very likely at least 95% of the observed differences attributable to intron loss, not intron gain (Roy and Hartl 2006).

We studied conserved regions of 1279 orthologous gene pairs between *P. falciparum* and the distantly related apicomplexan parasite *Theileria parva*. A total of 335/553 (60.6%) *P. falciparum* intron positions were shared with *T. parva*; 335/1842 (18.2%) *T. parva* intron positions were shared with *P. falciparum*. We then compared 3305 pairs of orthologous genes between *T. parva* and *T. annulata* (diverged ~82 Mya). Among 7111 intron positions in conserved regions, 7089 were shared between species, whereas only 11 (0.15%) were specific to each species. Implied intron loss/gain rates between *Theileria* species and between *Plasmodium* species are orders of magnitude smaller than estimated rates between *T. parva* and *P. falciparum*.

Proposed mechanisms of intron loss and gain involve transposable elements (TE) or TE-encoded reverse transcriptases. Thus, as noted previously for *Plasmodium* (Roy and Hartl 2006), a dearth of intron loss and gain in *Theileria* is consistent with the lack of transposable elements in modern *Theileria* species. We suggest that in some lineages including apicomplexans most intron loss/gain may be confined to relatively rare episodes of transposable element invasion.

Results and Discussion

Very little intron loss/gain in *Theileria*

We compared intron–exon structures for 3305 putative pairs of orthologous genes for *T. parva* and *T. annulata*. In 3.24 Mb of conserved amino acid level alignment, there were only 22 intron

[1]Corresponding author.
E-mail scottwroy@gmail.com; fax (617) 496-5854.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5410606.

16:1270–1275 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

positions specific to one species out of 7111 total intron positions. Eleven introns were specific to each species (Table 1; Fig. 1). The rate of divergence at synonymous positions between these species (d_S) has been estimated as 0.82 (Pain et al. 2005). Assuming that rates of divergence at synonymous positions reflects that found in *Plasmodium* (~5×10^{-9}; Castillo-Davis et al. 2004; Tanabe et al. 2004; Mu et al. 2005; Neafsey et al. 2005), gives an estimated divergence time of ~82 Mya. The *Theileria* genes experiencing intron loss/gains are summarized in Table 2.

Intron gain/loss between *T. parva* and *Plasmodium falciparum*

We next analyzed 2060 intron positions in 596 kb of conserved regions of 1279 putative pairs of orthologous genes for *T. parva* and *P. falciparum*. A total of 335/553 (60.6%) of *P. falciparum* intron positions were shared with *T. parva*, whereas 335/1842 (18.2%) *T. parva* intron positions were shared with *P. falciparum* (Table 1; Fig. 1). The results were nearly identical in a subset of 597 gene pairs for which conservation of synteny makes particularly confident orthology assignment possible. In a comparison of a small set of conserved regions of 464 possibly orthologous sequences between *T. parva* and *Paramecium tetraureli*, 5/17 *P. tetraurelia* introns were found in *T. parva* and 5/12 *T. parva* introns were found in *P. tetraurelia*. However, these results are difficult to interpret because of small sample numbers, and the lack of a full *P. tetraurelia* genome for analysis prohibits confident orthology assignments.

Discordant *Theileria* introns

We BLASTed each of the 22 introns found in only one of the two *Theileria* species against the genome in which it is found (e.g., *T. parva* introns against the *T. parva* genome). None of the intronic sequences gave hits with convincing sequence similarity to any intergenic, coding, or intronic sequence. Twenty-one of 22 were exact changes, without loss or gain of adjacent coding sequences. Comparisons with other apicomplexans for which genomic sequence was available confirmed intron presence in 10 cases, mostly in *Babesia bigemina*, strongly suggesting intron loss. In many cases intron positions that are conserved between the two *Theileria* species are absent in these outgroups, thus absence of discordant *Theileria* introns in these outgroups does not imply intron gain. Full understanding of the history of these introns will have to await the availability of fully annotated genome sequences for additional apicomplexan species.

Interestingly, most discordant *Theileria* introns (77.2%) including most confirmed losses (80.0%) fell in phase one, much higher than the fraction of all *Theileria* introns (33.9%, $P < 0.005$ by a binomial test). This is particularly surprising since previously a bias toward loss of phase zero introns was found for a variety of eukaryotic lineages (Roy and Gilbert 2005a). In the manner of Roy and Hartl (2006), we identified discordant introns that were 5′ or 3′ of the median intron in conserved regions and found no difference (6 vs. 7 for all discordant introns; 3 vs. 2 for confirmed intron losses), thus there is no clear bias toward a 3′ biased loss. In addition, in only one case were two discordant introns found in the same gene, and these were not adjacent. Thus, in these ways the data does not provide evidence for mRNA-mediated intron loss, although this could reflect the small number of intron losses identified.

The average length of intergenic regions flanking genes experiencing an intron loss/gain were very similar to values for all genes for *Plasmodium* (2221 nucleotides in all genes vs. 2257 for genes experiencing intron loss/gains) and *Theileria* (1618 vs. 2034, $P = 0.48$ by a Monte Carlo simulation).

Gain and loss in *Theileria* and *Plasmodium*

To get a sense of the relative importance of intron loss and gain in *Theileria* and *Plasmodium* since the common ancestor, we assessed intron presence in homologous genes in the distantly related apicomplexan species *Toxoplasma gondii* for 30 *Theileria*-specific and 30 *Plasmodium*-specific introns. In both cases, 19/30 were present in *T. gondii*, suggesting intron loss. Levels of conservation of *Theileria/Plasmodium* introns in *T. gondii* are unknown, thus for the introns that are absent in *T. gondii* intron loss/gain is uncertain. This small sample suggests that intron losses have outnumbered intron gains in these species; however, full understanding of intron evolution in the deeper branches of apicomplexans will have to await availability of a fully annotated *T. gondii* genome.

Rates of intron loss and gain

We next estimated rates of intron loss and gain from the data. Consider two species A and B that diverged t million years ago, in which the ancestor contained N introns in the c Mb of studied orthologous coding sequence, and in which both species have experienced constant rate of intron gain g/Mb/My and intron loss l/My since their divergence. We assume that all observed intron positions shared between the two species represent truly ancestral introns (i.e., no multiple insertions into homologous sites). If all introns are lost at equal rate, the chance that an ancestral intron has not been lost by the present time along a single lineage is given by the exponential distribution e^{-lt}. The expected number of introns shared between the species is the number of ancestral introns N times the probability that an ancestral intron has not been lost in species A (e^{-lt}) times the probability that it has not been lost in species B (also e^{-lt}), or Ne^{-2lt}. The probability that an ancestral intron is lost in a given species is $1-e^{-lt}$. An intron may thus be retained in species A (with probability e^{-lt}) but lost in species B (with probability $1-e^{-lt}$), with total probability $e^{-lt}(1-e^{-lt})$, or lost in A but retained in B also with probability $e^{-lt}(1-e^{-lt})$; thus the expected total number of ancestral introns that are present in only one species is $N \times e^{-lt}(1-e^{-lt}) \times 2$.

The rate of intron gain across the entire sequence is simply cg per My. However, an intron that is gained at time t_g before the present may be subsequently lost, with probability given by the exponential distribution e^{-lt_g}. Integrating, we get an expecta-

Table 1. Summary of species comparisons

Species 1	Species 2	Genes	Intron Positions			
			Total	Shared[a]	Sp. 1[b]	Sp. 2[b]
T. parva	*T. annulata*	3305	7111	7089 (99.7%)	11	11
P. yoelii[c]	*P. falciparum*	3479	2212	2185 (98.8%)	8	19
P. falciparum	*T. parva*	1293	2060	335 (16.3%)	218	507
T. parva	*P. tetraurelia*	87	24	5 (20.8%)	7	12

[a]Shared column indicates the number of intron positions shared between species.
[b]Sp. 1 and Sp. 2 columns indicate the number of intron positions that are specific to one species.
[c]Data from Roy and Hartl (2006).

Table 2. *Theileria* genes experiencing intron loss/gains

Gene (# of introns)		Shared introns	Unshared introns[a]	Gene function
T. annulata	*T. parva*			
CAI76020.1 (0)	EAN30663.1 (2)	0	P1	Unknown
CAI72925.1 (1)	EAN34406.1 (3)	1	P3	Unknown
CAI76889.1 (4)	EAN32247.1 (3)	3	$A2^{B}$	S-adenosylmethionine synthetase, putative
CAI75691.1 (7)	EAN31214.1 (9)	6	$P4^{B}$,P9	Origin recognition protein; microtubule associated
CAI75529.1 (13)	EAN31053.1 (12)	11	A2	Unknown
CAI73278.1 (6)	EAN34060.1 (6)	5	$P2^{B}$	DEAD box RNA helicase, putative
CAI75338.1 (10)	EAN30845.1 (8)	7	$A1^{B}$	Thioredoxin reductase, putative
CAI72901.1 (10)	EAN34429.1 (7)	3	A7	Unknown
CAI74423.1 (1)	EAN32842.1 (1)	0	$P1^{B,F}$	GTP-binding protein rab11
CAI73790.1 (1)	EAN33503.1 (1)	0	$P1^{B}$	Unknown
CAI72993.1 (1)	EAN34341.1 (0)	0	A1	Cyclophilin-type peptidyl-prolyl
CAI73274.1 (10)	EAN34064.1 (4)	2	$P2^{B}$	Nucleic-acid binding
CAI74816.1 (7)	EAN32452.1 (6)	4	$P3^{B,E}$	Dihydroorotate dehydrogenase
CAI73423.1 (1)	EAN33921.1 (1)	0	A1	Cytochrome oxidase subunit II precursor
CAI75266.1 (9)	EAN30758.1 (5)	5	$A3^{B}$	Unknown
CAI75761.1 (5)	EAN31286.1 (4)	4	A5	Unknown
CAI75445.1 (1)	EAN30965.1 (3)	1	$P3^{B,E,T}$	Succinate dehydrogenase flavoprotein subunit
CAI72908.1 (8)	EAN34421.1 (8)	7	P8	Pre-mRNA splicing factor
CAI74742.1 (5)	EAN32526.1 (4)	3	A5	Heat shock protein 110
CAI74412.1 (1)	EAN32857.1 (6)	0	A1	Unknown
CAI75330.1 (4)	EAN30832.1 (2)	1	A3	RhoGAP protein

[a]"A" and "P" denote *T. annulata* and *T. parva*, respectively, thus P1 indicates that *T. parva* intron 1 is absent in *T. annulata*. "B," "E," "F," and "T" denote intron presence in distantly related apicomplexans *B. bigemina, E. tenella, P. falciparum,* and *T. gondii*, respectively.

tion of $\int_0^t cge^{-lt_g}dt_g = (cg/l)(1 - e^{-lt})$ extant intron gains per lineage since the time of divergence, thus a total of twice that number of species-specific intron gains (since there are two species).

In the case of the *T. parva–T. annulata* comparison, there are 22 species-specific and 7089 shared intron positions in 3.24 Mb of sequence. Ten of the species-specific introns are known to be due to intron loss. The remaining 12 may be due to intron loss or gain. Assuming that all 22 species-specific introns are attributable to intron loss, we estimate $22/7089 = 2Ne^{-lt}(1-e^{-lt})/Ne^{-2lt}$, which gives $lt = 0.0015$, or $l = 1.9 \times 10^{-5}$/My for $t = 82$ My. Assuming that all 12 species-specific introns of unknown origin are due to intron gain gives estimates of $l = 8.6 \times 10^{-6}$/My and $g = 0.023$/Mb/My. Figure 2 shows the entire range of estimates assuming between 0 and 12 of the species-specific introns are due to intron gain. For a previous *P. falciparum–P. yoelii* comparison (Roy and Hartl 2006), at least 19 of the 27 species-specific introns among 2212 total introns in 3.5 Mb of conserved regions are due to intron loss. Estimated rates of loss, assuming $t = 100$ Mya, are therefore $5.4–6.2 \times 10^{-5}$/My, and the estimated gain rate is between zero and $g = 0.011$/Mb/My (Fig. 2).

Rates of loss and/or gain implied by the *T. parva–P. falciparum* divergence are much higher. The *Theileria–Plasmodium* divergence very likely postdates the deepest splits within known apicomplexans, estimated to be 350–824 Mya (Escalante and Ayala 1995). These values are unlikely to be a significant underestimate, since the apicomplexan ancestor was very likely to have been an animal parasite and therefore to not predate the origin of animals (Zilversmit and Hartl 2005). Rates of intron gain and loss similar to those estimated for the *P. falciparum–P. yoelii* and *T. parva–T. annulata* divergence would therefore predict that between 3 and 10% of introns in a *T. parva–P. falciparum* comparison should be species specific. Instead, 84% of intron positions are species specific.

Assuming that all 1725 species-specific introns out of 2060 total intron positions in 0.60 Mb of conserved sequences are due to intron loss gives an estimate of $l = 0.0015–0.0075$. These values are two orders of magnitude larger than the *P. falciparum–P. yoelii* and *T. parva–T. annulata* estimates. Attributing all 1725 species-specific introns to intron gain yields estimates of $g = 1.8–4.1$/Mb/My (assuming t of 824 My and 350 My, respectively), two to three orders of magnitude higher than the *P. falciparum–P. yoelii* and *T. parva–T. annulata* estimates. Figure 2 gives the entire range of estimates assuming that between 0 and 1725 of the species-specific introns are attributable to intron gain, showing clearly higher estimates than the *T. parva–T. annulata* and *P. falciparum–P. yoelii* estimates.

Rates of intron evolution through time

Other results suggest significant variations in intron loss/gain rate through time. While intron position divergence between mouse and humans is only 0.1%, divergence between mammals and flies or worms, reflecting a time of divergence perhaps tenfold earlier, is not 1% but ~80% (Rogozin et al. 2003; Roy et al. 2003). While intron position divergence between euascomycetous fungi appears to be around one quarter, divergence between euascomycetes and hemiascomycetes is near complete, apparently due to hemiascomycetes having lost nearly all of their ancestral introns (Rogozin et al. 2003; Nielsen et al. 2004). Similarly, while divergence between *Theileria* and *Plasmodium* is 84%, their divergence from the slightly more distantly related *C. parvum*, with only one intron per 10 genes, is nearly complete (Abrahamsen et al. 2004).

At the rates of intron loss estimated from the *T. parva–T. annulata* comparison, it would take 35 billion years for a lineage to lose half of its ancestral introns, yet many lineages have lost much higher fractions of their introns over much shorter times (Rogozin et al. 2003; Roy and Gilbert 2005b,c). At the rates of intron gain estimated from the *T. parva–T. annulata* comparison, it would take hundreds of billions of years to reach the intron densities of 5–8 introns per gene observed in a variety of eukaryotic lineages, even ignoring subsequent intron loss. Thus, intron loss and gain rates have clearly varied substantially through eukaryotic evolution.

Transposable elements and rate variation

Why have intron loss and gain rates varied so significantly through the history of these species? Differences in intron number have traditionally been attributed to differences in selection based on biological differences, with for instance fast-replicating species experiencing more selection against excess DNA. Such an explanation does not readily present itself in this case. Both *Plasmodium* and *Theileria* are intracellular parasites of vertebrates transmitted by an arthropod vector. Both groups have a complex life cycle with multiple different asexual stages and a single meiotic cycle per transmission.

Instead, the rate of intron loss and gain mutations themselves may be a central part of the explanation (Roy and Gilbert 2006; Roy and Hartl 2006). Intron loss and gain may not follow a classical mode of near-constant rate but may instead be largely confined to dramatic episodes (Fedorov et al. 2003). Such a scenario is in fact expected in lineages in which the number of TEs varies through time. The most likely mechanism of intron loss is reverse transcription (likely largely by TE-encoded reverse transcriptase) of a spliced mRNA and subsequent recombination with the genomic copy (Mourier and Jeffares 2003; Roy and Gilbert 2005a). Intron gain is likely also dependent on TEs, either via reverse splicing of spliced-out RNA introns into previously intronless sites of mRNAs, followed by reverse transcription and subsequent double recombination with the genomic copy, or by simple conversion of coding sequence-interrupting TE insertions into new introns (Crick 1979; Sharp 1985; Roy 2004). It is well established that rates of TE insertion vary through time (e.g., Lander et al. 2001; Blumenstiel et al. 2002; Salem et al. 2005), and that individual TE families can go extinct in species that previously harbored them (Lander et al. 2001), or can be introduced to species that previously lacked these families (Robertson and Lampe 1995; de Almeida and Carareto 2005; Sanchez-Gracia et al. 2005; Diao et al. 2006). Such differences in TE numbers and insertion patterns through time might therefore lead to dramatic differences in both rates of intron loss and gain through time, leading to apparent non-linearity of rates of intron position change.

There are no known TEs in *Plasmodium* or *Theileria*. The dearth of intron loss and gain in *Plasmodium* and *Theileria* could therefore be due to the lack of known TEs in those groups, while the much higher degree of change between the two groups could reflect one or more TE invasions since the common ancestor. Given the high rates of gene conversion in some apicomplexans, a relatively short spurt of high TE abundance could lead to a large amount of intron loss, leading to large deviations from clock-like behavior.

What could drive variation in TE abundance in the evolution of apicomplexan parasites? Two possibilities present themselves. First, given the close association between both groups studied here and mammalian hosts, which have experienced dramatic changes in TE number through time (Lander et al. 2001), it is possible that fluctuation in host TE number could influence fluctuation in parasite TE number. Second, differences in demographics and relationships of host, vector, and parasite could create fluctuations in TE number. During times of low rates of parasite transmission, most infections may contain only a single parasite genotype, while increases in transmission may increase rates of zygote formation between unlike genotypes (Ferreira et al. 1998; Razakandrainibe et al. 2005). Increases in transmission could accompany host or vector switches due to lack of immune defense against the newly introduced parasite or to increased virulence (Boyd 1949; Waters et al. 1991; Mu et al. 2005). Since theory predicts that proliferation of TEs under certain conditions requires sexual reproduction between individuals of unlike genotype (Hickey 1982), this could lead to transient conditions permissible to TE proliferation.

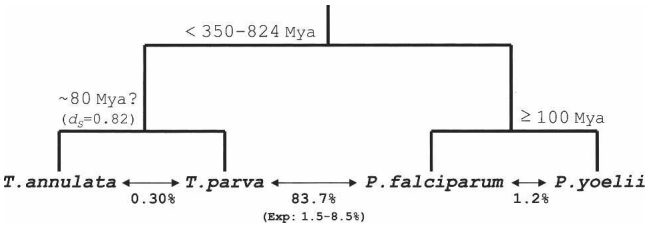

Figure 1. Intron divergences between and within *Theileria* and *Plasmodium* species. Percentages of total intron positions in conserved regions that are species specific are given for the three apicomplexan comparisons. Times of divergence are from Escalante and Ayala (1995) (*Theileria–Plasmodium* divergence), from an assumption of *P. yoelii–P. falciparum* speciation at least as deep as host divergence, and based on an estimated d_S = 0.82 divergence between *Theileria* species, assuming a nucleotide substitution rate of 5×10^{-9} (*T. annulata–T. parva* divergence).

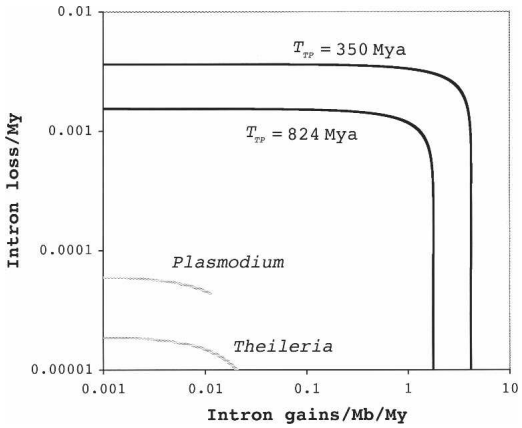

Figure 2. Estimated rates of intron loss and gain in three comparisons. The *Plasmodium* and *Theileria* traces give possible estimates of intron loss and gain for the *P. falciparum–P. yoelii* and *T. parva–T. annulata* comparisons, respectively, given the number of observed species-specific intron positions. The black traces give possible estimates derived from the *T. parva–P. falciparum* comparison, assuming either a divergence time of 350 Mya or 824 Mya. Estimates are derived as described in the text.

Implications to gene prediction and genome annotation

This is at least the third genome-wide study to show that in conserved regions of alignment, the vast majority of intron positions can be conserved over very long evolutionary times (Roy et al. 2003; Roy and Hartl 2006). In each of these pairwise comparisons of intron positions between pairs of putatively orthologous genes, a very large number of apparently discordant intron positions are cases in which an intron in one of the orthologous genes lies adjacent to a gap in the alignment of that same sequence. Such cases are very simply explained if a sequence that has been annotated as exonic in one of the genes is in fact an intron, or if the annotated intron sequence in the other gene is in fact exonic. These findings suggest that annotations could be improved, perhaps greatly, by comparison of apparently orthologous sequences, to determine whether corresponding sequence in the two genes may be called either both intronic or both exonic.

Conclusions

We show that intron loss and gain has been very scant between *T. parva* and *T. annulata*, that introns in *P. falciparum* are largely conserved in the distant genus *Theileria*, and that there are large apparent differences in the rate of intron loss/gain divergence between closely and distantly related apicomplexan species.

Methods

Sequences and orthologous gene pair definition

We downloaded the *T. parva* and *T. annulata* genome sequences and annotations from GenBank (accession numbers

AAGK01000001.1 and CR940346.1, respectively), and the *P. falciparum* genome sequence and annotation from PlasmoDB (plasmodb.org, version 10.03.2002.v2). We also downloaded the sequence and annotation for the largest chromosome of *P. aurelia* (Zagulski et al. 2004) from GenBank. Reciprocal BLASTP searches between *T. parva* and *T. annulata* and between *T. parva* and *P. falciparum* yielded 3305 and 1279 pairs of putatively orthologous genes, respectively. We used ClustalW with default parameters to align protein sequences of each pair and mapped the intron positions onto the resultant alignments using a custom Perl program. BLASTP searches of the *P. tetraurelia* sequences against *T. parva* yielded 464 putatively orthologous gene pairs.

Analysis of *T. parva–T. annulata* alignments

For each *T. parva–T. annulata* gene pair, we first excluded intron position discrepancies that were due to several obvious and recurrent annotation errors or concerned introns in doubtful regions of alignment. We excluded intron positions with less than 50% amino acid-level identity in the 15 aligned amino acids on each side (thus retaining positions near gaps). We excluded intron positions in one sequence opposite or within five positions of a 5-amino-acid or greater gap in the same sequence, as such cases are easily explained as an exonic stretch of sequence having been erroneously called an intron by the annotation (in the intron-containing sequence) or vice versa (in the other sequence). However, we retained intron positions adjacent to or within a 15-amino-acid or longer gap in the other sequence, as such cases are not explicable as simple annotation errors and could reflect intron loss or gain in which the adjacent sequences have been added/lost at the same time. We next excluded sequences that fell at the beginning or end of the alignment or in regions before/after the first/last region of good alignment. Custom Perl programs were written to perform these filters. Each apparent discordance was then analyzed by eye. The vast majority of remaining cases concerned a species-specific intron position near in the alignment to a shared intron position with an intervening gap and no intervening region of homology. Such cases are easily explained as an error in the prediction of the boundary of one intron or the other, such that an intron–exon–intron had been called a single intron or vice versa. A few other cases involved very slight (<5 bp) differences in intron positions between species, likely because of annotation error. This left 22 intron discordances, all in genes with well-conserved synteny. We performed BLASTN searches of each remaining intron against the corresponding genome. In no case was a clear sequence similarity found between the discordant intron and another genomic element.

Analysis of *T. parva–P. falciparum* and *T. parva–P. tetraurelia* alignments

Because of the much greater divergence between these species, we applied a different definition of "conserved regions." We required that one third of the 25 positions in the alignment on either side of an intron (including gaps) were conserved. For the *T. parva–P. falciparum* alignments, this yielded a total of 2060 introns. The analysis was also performed using a more restrictive criterion of 50% conservation over 15 alignment positions on each side, excluding intron positions with alignment gaps within five positions. Degree of intron position conservation under these conditions was only marginally higher (64.0% of *P. falciparum* introns present in *T. parva*; 20.0% of *T. parva* introns present in *P. falciparum*). We further identified 597 with evidence of conserved synteny between *T. parva* and *P. falciparum*. An orthologous pair was considered to be syntenic if each fell within 10 kb of the corresponding gene from a second putatively orthologous pair. The rate of intron conservation in this subset of genes was very similar (60.6% of *P. falciparum* introns present in *T. parva*; 17.0% of *T. parva* introns present in *P. falciparum*). For the *T. parva–P. tetraurelia* there were 24 intron positions in regions of conserved alignments.

Presence/absence of discordant *Theileria* introns in other apicomplexans

To determine whether each of the 22 introns present in only one of the two *Theilieria* species was present in distantly related apicomplexans, we ran BLASTP searches of the corresponding protein sequences against the *Eimeria tenella* GeneDB database of preliminary gene predictions (http://www.genedb.org/genedb/etenella/) and ran TBLASTN sequences against *E. tenella* (http://www.sanger.ac.uk/Projects/E_tenella/) and *Babesia bigemina* contigs (http://www.sanger.ac.uk/Projects/B_bigemina/) to identify possible unannotated or misannotated sequences. For good GeneDB hits, we determined the presence/absence by inspecting the intron positions in the GeneDB entry; for contig hits, we determined the presence/absence of a gap in the alignment at a position corresponding to the intron position. In a few cases, sequence similarity with the contig ended abruptly at the intron position, although sequence homologous to the other flanking exon could not be found. Such cases are easily explained if the intron is present in the species, but only one flanking exon is currently represented in the genome assembly and were therefore scored as probable intron presence.

Intergenic distances

We calculated average flanking intergenic distance for all *T. parva* and all *P. falciparum* genes and for the 21 *Theileria* genes and 21 *Plasmodium* genes (Roy and Hartl 2006) experiencing intron loss/gains. We used Monte Carlo simulation to generate 10,000 random sets of 21 genes for *Theileria*. A total of 4811/10,000 sets had a higher average distance than the real set, yielding $P = 0.48$.

References

Abrahamsen, M.S., Templeton, T.J., Enomoto, S., Abrahante, J.E., Zhu, G., Lancto, C.A., Deng, M., Liu, C., Widmer, G., Tzipori, S., et al. 2004. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* **304:** 441–445.

Babenko, V.N., Rogozin, I.B., Mekhedov, S.L., and Koonin, E.V. 2004. Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res.* **32:** 3724–3733.

Blumenstiel, J.P., Hartl, D.L., and Lozovsky, E.R. 2002. Patterns of insertion and deletion in constrasting chromatin domains. *Mol. Biol. Evol.* **19:** 2211–2225.

Boyd, M.F. 1949. Historical review. In *Malariology* (ed. M.F. Boyd), pp. 3–25. Saunders, Philadelphia.

Castillo-Davis, C.I., Bedford, T.B., and Hartl, D.L. 2004. Accelerated rates of intron gain/loss and protein evolution in duplicate genes in human and mouse malaria parasites. *Mol. Biol. Evol.* **21:** 1422–1427.

Crick, F. 1979. Split genes and RNA splicing. *Science* **204:** 264–271.

Csuros, M. 2005. Likely scenarios of intron evolution. Third RECOMB Satellite Workshop on Comparative Genomics. *Lecture Notes Comp. Sci.* **3678:** 47–60.

de Almeida, L.M. and Carareto, C.M. 2005. Multiple events of horizontal transfer of the Minos transposable element between *Drosophila* species. *Mol. Phylogenet. Evol.* **110:** 583–594.

Diao, X., Freeling, M., and Lisch, D. 2006. Horizontal transfer of a plant transposon. *PLoS Biol.* **4:** e5.

Dibb, N.J. and Newman, A.J. 1989. Evidence that introns arose at proto-splice sites. *EMBO J.* **8:** 2015–2021.

Edvardsen, R.B., Lerat, E., Maeland, A.D., Flat, M., Tewari, R., Jensen, M.F., Lehrach, H., Reinhardt, R., Seo, H.C., and Chourrout, D. 2004. Hypervariable and highly divergent intron–exon organizations in the chordate *Oikopleura dioica*. *J. Mol. Evol.* **59:** 448–457.

Escalante, A.A. and Ayala, F.J. 1995. Evolutionary origin of *Plasmodium* and other Apicomplexa based on rRNA genes. *Proc. Natl. Acad. Sci.* **92:** 5793–5797.

Fedorov, A., Merican, A.F., and Gilbert, W. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc. Natl. Acad. Sci.* **99:** 16128–16133.
Fedorov, A., Roy, S., Fedorova, L., and Gilbert, W. 2003. Mystery of intron gain. *Genome Res.* **13:** 2236–2241.
Ferreira, M.U., Lin, Q., Kimura, M., Ndawi, B.T., Tanabe, K., and Kawamoto, F. 1998. Allelic diversity in the merozoite surface protein-1 and epidemiology of multiple-clone *Plasmodium falciparum* infections in northern Tanzania. *J. Parasitol.* **84:** 1286–1289.
Hickey, D.A. 1982. Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* **101:** 519–531.
Jeffares, D.C., Mourier, T., and Penny, D. 2006. The biology of intron gain and loss. *Trends Genet.* **22:** 16–22.
Kent, W.J. and Zahler, A.M. 2000. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae–C. elegans* genomic alignment. *Genome Res.* **10:** 1115–1125.
Kiontke, K., Gavin, N.P., Raynes, Y., Roehrig, C., Piano, F., and Fitch, D.H. 2004. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc. Natl. Acad. Sci.* **101:** 9003–9008.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Moriyama, E.N., Petrov, D.A., and Hartl, D.L. 1998. Genome size and intron size in *Drosophila*. *Mol. Biol. Evol.* **15:** 770–773.
Mourier, T. and Jeffares, D.C. 2003. Eukaryotic intron loss. *Science* **300:** 1393.
Mu, J., Joy, D.A., Duan, J., Huang, Y., Carlton, J., Walker, J., Barnwell, J., Beerli, P., Charleston, M.A., Pybus, O.G., et al. 2005. Host switch leads to emergence of *Plasmodium vivax* malaria in humans. *Mol. Biol. Evol.* **22:** 1686–1693.
Neafsey, D.E., Hartl, D.L., and Berriman, M. 2005. Evolution of noncoding and silent coding sites in the *Plasmodium falciparum* and *Plasmodium reichenowi* genomes. *Mol. Biol. Evol.* **22:** 1621–1626.
Nielsen, C.B., Friedman, B., Birren, B., Burge, C.B., and Galagan, J.E. 2004. Patterns of intron gain and loss in fungi. *PLoS Biol.* **2:** e422.
Qiu, W.G., Schisler, N., and Stoltzfus, A. 2004. The evolutionary gain of spliceosomal introns: Sequence and phase preferences. *Mol. Biol. Evol.* **21:** 1252–1263.
Pain, A., Renauld, H., Berriman, M., Murphy, L., Yeats, C. A., Weir, W., Kerhornou, A., Aslett, M., Bishop, R., Bouchier, C., et al. 2005. Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. *Science* **309:** 131–133.
Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R., and Dodgson, J. 1980. The evolution of genes: the chicken preproinsulin gene. *Cell* **20:** 555–566.
Razakandrainibe, F.G., Durand, P., Koella, J.C., De Meeus, T., Rousset, F., Ayala, F.J., and Renaud, F. 2005. "Clonal" population structure of the malaria agent *Plasmodium falciparum* in high-infection regions. *Proc. Natl. Acad. Sci.* **102:** 17388–17393.
Robertson, H.M. and Lampe, D.J. 1995. Recent horizontal transfer of a mariner transposable element among and between Diptera and Neuroptera. *Mol. Biol. Evol.* **12:** 850–862.
Rogozin, I.B., Wolf, Y.I., Sorokin, A.V., Mirkin, B.G., and Koonin, E.V. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* **13:** 1512–1517.
Roy, S.W. 2004. The origin of recent introns: transposons? *Genome Biol.* **5:** 251.
Roy, S.W. and Gilbert, W. 2005a. The pattern of intron loss. *Proc. Natl. Acad. Sci.* **102:** 713–718.
———. 2005b. Complex early genes. *Proc. Natl. Acad. Sci.* **102:** 1986–1991.
———. 2005c. Rates of intron loss and gain: Implications for early eukaryotic evolution. *Proc. Natl. Acad. Sci.* **102:** 5773–5778.
———. 2006. The evolution of spliceosomal introns: Patterns, puzzles, and progress. *Nat. Rev. Genet.* **7:** 211–221.
Roy, S.W. and Hartl, D.L. 2006. Very little intron loss/gain in *Plasmodium*: Intron loss/gain mutation rates and intron number. *Genome Res.* **16:** 750–756.
Roy, S.W., Fedorov, A., and Gilbert, W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl. Acad. Sci.* **100:** 7158–7162.
Salem, A.H., Ray, D.A., Hedges, D.J., Jurka, J., and Batzer, M.A. 2005. Analysis of the human Alu Ye lineage. *BMC Evol. Biol.* **5:** 18.
Sanchez-Gracia, A., Maside, X., and Charlesworth, B. 2005. High rate of horizontal transfer of transposable elements in *Drosophila*. *Trends Genet.* **21:** 200–203.
Seo, H.C., Kube, M., Edvardsen, R.B., Jensen, M.F., Beck, A., Spriet, E., Gorsky, G., Thompson, E.M., Lehrach, H., Reinhardt, R., et al. 2001. Miniature genome in the marine chordate *Oikopleura dioica*. *Science* **294:** 2506.
Sharp, P.A. 1985. On the origins of RNA splicing and introns. *Cell* **42:** 397–400.
Tanabe, K., Sakihama, N., Hattori, T., Ranford-Cartwright, L., Goldman, I., Escalante, A.A., and Lal, A.A. 2004. Genetic distance in housekeeping genes between *Plasmodium falciparum* and *Plasmodium reichenowi* and within *P. falciparum*. *J. Mol. Evol.* **59:** 687–694.
Vanacova, S., Yan, W., Carlton, J.M., and Johnson, P.J. 2005. Spliceosomal introns in the deep-branching eukaryote *Trichomonas vaginalis*. *Proc. Natl. Acad. Sci.* **102:** 4430–4435.
Waters, A.P., Higgins, D.G., and McCutchan, T.F. 1991. *Plasmodium falciparum* appears to have arisen as a result of lateral transfer between avian and human hosts. *Proc. Natl. Acad. Sci.* **88:** 3140–3144.
Zagulski, M., Nowak, J.K., Le Mouel, A., Nowacki, M., Migdalski, A., Gromadka, R., Noel, B., Blanc, I., Dessen, P., Wincker, P., et al. 2004. High coding density on the largest *Paramecium tetraurelia* somatic chromosome. *Curr. Biol.* **14:** 1397–1404.
Zilversmit, M. and Hartl, D.L. 2005. Evolutionary history and population genetics of human malaria parasites. In *Molecular approaches to malaria* (ed. I.W. Sherman), pp. 95–109. American Society for Microbiology Press, Washington, D.C.

Received April 17, 2006; accepted in revised form July 26, 2006.