J Mol Evol (2010) 71:364–373
DOI 10.1007/s00239-010-9391-6

Evaluation of Models of the Mechanisms Underlying Intron Loss and Gain in *Aspergillus* Fungi

Lei-Ying Zhang · Yu-Fei Yang · Deng-Ke Niu

Received: 15 July 2010 / Accepted: 8 September 2010 / Published online: 23 September 2010

Abstract Although intron loss and gain have been widely observed, their mechanisms are still to be determined. In four *Aspergillus* genomes, we found 204 cases of intron loss and 84 cases of intron gain. Using this data, we tested common hypotheses of intron loss or gain. Statistical analysis showed that adjacent introns tend to be lost simultaneously and small introns were preferentially lost, supporting the model of mRNA-mediated intron loss. The lost introns reside in internal regions of genes, which is inconsistent with the traditional version of the model (partial length cDNAs are reverse transcribed from 3′ ends of mRNAs), but consistent with an alternate version (partial length cDNAs are produced by self-primed reverse transcription). The latter version was not supported by examination of the abundance of T-rich segments in mRNAs. Preferential loss of internal introns might be explained by highly efficient recombination at internal regions of genes. Among the 84 cases of intron gain, we found a significantly higher frequency of short direct repeats near exon–intron boundary than in conserved introns, supporting the double-strand break repair model. We also found possible source sequences for two cases of intron gain, one by gene conversion and one by insertion of a mitochondrial sequence during double-strand break repair. Source sequences for most gained introns could not be identified and the possible reasons were discussed. In the four *Aspergillus* genomes studied, we did not find evidence of frequent parallel intron gains.

Keywords *Aspergillus* · mRNA-mediated intron loss · Parallel intron gain · Orthologous genes · Double-strand break repair · Self-primed reverse transcription

Introduction

The densities of spliceosomal introns vary greatly among organisms. In humans, most genes contain introns, and some extreme cases have more than 100 introns (Scherer 2008). By contrast, there are only four introns in the genome of the parasite *Giardia lamblia* (Morrison et al. 2007), and no introns at all in the nucleomorph genome of *Hemiselmis andersenii* (Lane et al. 2007). After many years of debate, it is generally accepted that there have been both intron losses and intron gains in the evolution of eukaryotes (de Souza 2003; Jeffares et al. 2006). However, the mechanisms by which old introns were lost and new introns originated have still not been fully elucidated.

On the mechanism of intron loss, there are two main models (Rodriguez-Trelles et al. 2006; Roy and Gilbert 2006). The first is simply genomic deletion, which assumes that introns are lost in unequal crossover recombination between alleles. The second is mRNA-mediated intron loss, which assumes that introns are lost in gene conversion or crossover recombination between genomic DNA and cDNA molecules that are reverse transcribed from the end of fully spliced mRNA (Fink 1987; Derr et al. 1991). These two models differ in their predictions of the patterns of intron loss. The first model proposes an inaccurate mechanism of intron loss; a small number of nucleotides may be added to or deleted from flanking exons during the loss of an intron.

Electronic supplementary material The online version of this article (doi:10.1007/s00239-010-9391-6) contains supplementary material, which is available to authorized users.

L.-Y. Zhang · Y.-F. Yang · D.-K. Niu (✉)
MOE Key Laboratory for Biodiversity Sciences and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing 100875, China
e-mail: dkniu@bnu.edu.cn; dengkeniu@hotmail.com

By contrast, introns are lost precisely in the model of mRNA-mediated intron loss. In addition, the second model predicts that adjacent introns may be lost simultaneously, and introns at the 3′ side of the genes would be preferentially lost, while the first model predicts that introns are lost individually without any bias in position. The model of mRNA-mediated intron loss has received much more support than that of genomic deletion (Llopart et al. 2002; Mourier and Jeffares 2003; Roy and Gilbert 2005a; Stajich and Dietrich 2006; Coulombe-Huntington and Majewski 2007a, b; Roy and Penny 2007; Loh et al. 2008), indicating that reverse transcripts have been involved in most of the intron loss events in evolution. However, there are also patterns of intron losses not predicted by the two models above. In *Cryptococcus* and ascomycetes, intron losses were not biased toward the 3′ end of genes, although adjacent introns tended to be lost simultaneously (Nielsen et al. 2004; Lin et al. 2006; Sharpton et al. 2008). A modified version of the mRNA-mediated intron loss is that the reverse transcription is self-primed by the 3′ end of mRNA molecules (Feiber et al. 2002; Niu et al. 2005). In this modified model, the 3′ end of an mRNA folds back and pairs with an internal segment of the mRNA. The position of the beginning of reverse transcription depends on how long the 3′ end folds back, which is expected to vary greatly among different mRNA molecules. Thus, the position of an intron lost from recombination of cDNA and genomic DNA depends on the secondary structure of the mRNA molecule, and is not necessarily at the 3′ side of the gene.

On the mechanism of intron gain, multiple models have been proposed, ranging from transposable element insertion to reverse splicing and genomic duplication (Rodriguez-Trelles et al. 2006; Roy and Gilbert 2006; Catania and Lynch 2008; Irimia et al. 2008; Roy and Irimia 2009). Unfortunately, there has been little progress in distinguishing models of intron gain because few definitive cases of recent intron gain were observed (Hankeln et al. 1997; Roy 2004). By analyzing the exon/intron structures among *Daphnia pulex* isolates, Li et al. (2009) recently identified 28 cases of recent intron gains, some of which are parallel intron gains. Benefiting from the short history of the intron gain events, they found the evolutionary traces (i.e., short direct repeats flanking the gained introns) and proposed a new model of intron gain: exogenous DNA inserted into an exon during the repair of a double-strand break forms a new intron.

To distinguish different models of intron gain, we require a sufficient number of recently gained introns. Previous studies showed that intron loss greatly outnumbers intron gain in most eukaryotic lineages studied (Roy and Gilbert 2005b; Roy and Hartl 2006; Roy and Penny 2006b; Stajich and Dietrich 2006; Carmel et al. 2007; Coulombe-Huntington and Majewski 2007a, b; Putnam et al. 2007; Stajich et al. 2007; Basu et al. 2008; Loh et al. 2008; Sharpton et al. 2008; Wilkerson et al. 2009). A notable exception on the frequency of intron gain is ascomycetes. Nielsen et al. (2004) showed that there is a rough balance between intron loss and intron gain in ascomycetes. As mentioned above, ascomycetes are also special in the position of lost introns. To obtain further insights into the mechanisms of intron loss and gain, we analyzed the patterns of intron loss and gain in four genomes of *Aspergillus*, a genus of ascomycetes.

Materials and Methods

Sequence Data and Ortholog Identification

Eight genomes of *Aspergillus* were retrieved from the Broad Institute (http://www.broadinstitute.org/science/data) on March 19, 2009. *Aspergillus clavatus* was first excluded because its genome sequence annotation was not completed. We also discarded *Aspergillus niger* because it has fewer matches in ortholog searches among *Aspergillus* genomes. For a simple tree with only one representative species in each most recent clade, we retained five species, *Aspergillus flavus*, *Aspergillus oryzae*, *Aspergillus terreus*, *Neosartorya fischeri*, and *Aspergillus nidulans* (Fig. 1). Five other fungal genomes, *Coprinopsis cinerea*, *Cryptococcus neoformans*, *Neurospora crassa*, *Histoplasma capsulatum*, and *Spizellomyces punctatus*, and one choanoflagellate genome *Proterospongia* sp., used as additional outgroups in the verification of intron gains were also downloaded from the Broad Institute.

Using the BLAST reciprocal best hits, orthologous proteins were detected with thresholds of E values $<10^{-9}$ and identities $\geq 25\%$. In total, we found 3,845 orthologous groups of proteins across the five *Aspergillus* species. The genes of 238 orthologous groups entirely lack introns and were not included in further analyses.

Alignment and Filtration of the Orthologous Genes

The proteins in each orthologous group were aligned using ClustalW (version 2.0.10) (Larkin et al. 2007) and intron positions were mapped into the protein alignments using ~ and 0/1/2 as marks for the absence and presence of an intron at a given position, respectively, with 0, 1, or 2 indicating the phase of the intron (Long et al. 1995). According to previous studies (Roy and Penny 2006a, 2007), the alignments were filtered by excluding introns with the following characteristics: (1) introns at two adjacent positions with a distance of ≤ 5 amino acids; (2) the presence of an intron at a position that flanks a gap or is close to a gap (within two amino acid positions), which might be a

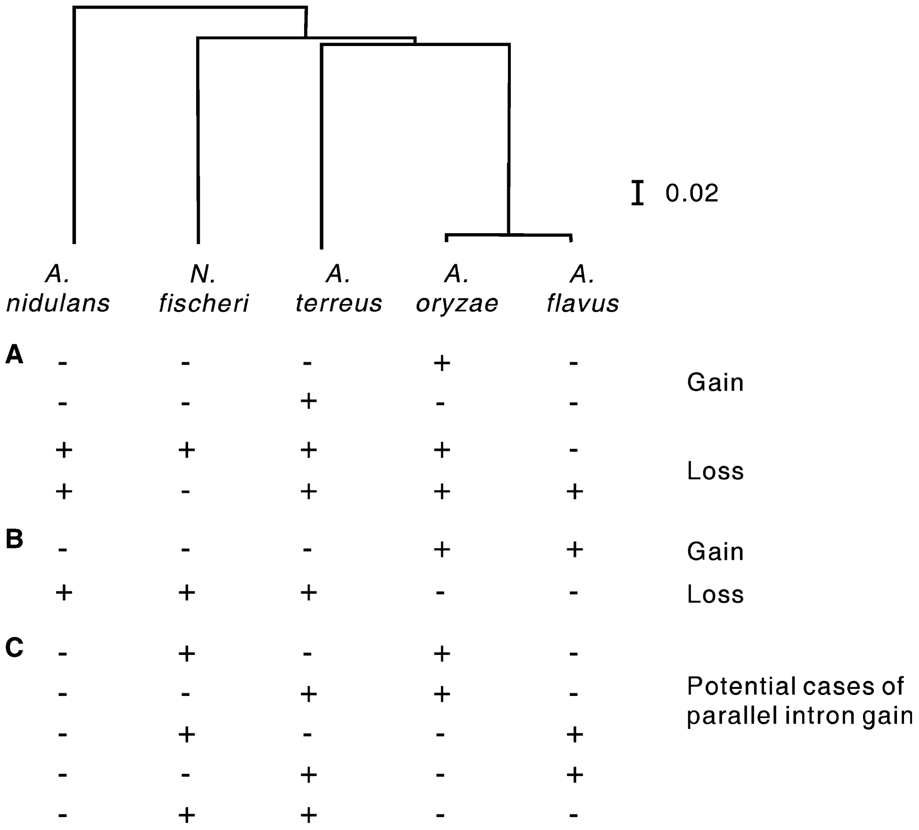


Fig. 1 Parsimony strategies for detecting intron gains and losses. The phylogenetic tree was downloaded from the *Aspergillus* Comparative Database of the Broad Institute (http://www.broadinstitute.org/annotation/genome/aspergillus_group/MultiHome.html) and is scaled according to phylogenetic distances. "+" and "−" represent the presence and the absence of an intron in a given position, respectively. (*A*) In a given intron position of the alignment, one presence and four absences were considered a case of intron gain, and similarly one absence and four presences were considered a case of intron loss. (*B*) Two presences in *A. flavus* and *A. oryzae* and three absences across *A. terreus*, *N. fischeri*, and *A. nidulans* were also considered cases of intron gain in *A. flavus* and *A. oryzae*. Two absences in *A. flavus* and *A. oryzae* and three presences across *A. terreus*, *N. fischeri*, and *A. nidulans* were also considered cases of intron loss in *A. flavus* and *A. oryzae*. (*C*) We also adapted a much less conservative strategy to encompass all potential cases of parallel intron gain

case of intronization (Irimia et al. 2008); (3) a large gap ($\geq$5 continuous amino acid positions) within 15 amino acid positions; (4) the identity (across all five species) of 15 amino acids flanking an intron position on each side <45%; and (5) introns whose positions are too close to the 5′ or 3′ end of a coding sequence (<15 amino acids positions).

After these filtrations, we obtained 2,144 groups of orthologs containing 3,333 conserved intron positions and 602 nonconserved intron positions. The *E* values of these 2,144 orthologous groups in the reciprocal BLAST range from 8×10^{-10} to 0.

Results and Discussion

Detection of Intron Loss and Gain

In this article, the terms intron loss and gain are used in their narrowest sense. Intron loss is the complete loss of the sequence of an intron from a gene and intron gain is insertion of an exogenous sequence into a gene. A new intron might be converted from exonic sequence by mutations that create new splicing signals (termed intronization) and an old intron might be converted into an exonic segment by mutations that inactivate the splicing signals (termed de-intronization) (Catania and Lynch 2008; Irimia et al. 2008; Gao and Lynch 2009; Roy 2009). Our preliminary analysis based on the present annotation of genomes indicates that intronization and de-intronization are frequent in *Aspergillus* (unpublished data). However, the identification of intronization and de-intronization events depends heavily on the quality of genome sequence annotation. Mis-annotation of an internal segment of an exon into an intron would produce a false positive case of intronization, and similarly mis-annotation of an intron into an internal segment of a large exon would produce a false positive case of de-intronization. For a reliable analysis of intronization and de-intronization, we must wait for high coverage transcriptome data of the analyzed species By contrast, the gain and loss of an intron can be inferred by the presence and the absence of a sequence in the gene, which depends heavily on the quality of sequencing and assembly of the genome, but not so much on the annotation of the genome sequence. Some authors have noted that the relatively simple gene structures of fungi compared with plants and animals make their gene prediction more accurate (Nielsen et al. 2004). Therefore, draft assemblies of genome sequences have been widely used in studies on intron loss and gain in fungi (Nielsen et al. 2004; Stajich and Dietrich 2006; Stajich et al. 2007; Sharpton et al. 2008).

We adopted different strategies to distinguish between intron loss and gain. Stringent parsimony strategies were used to identify intron losses and gains with a high confidence level and a relaxed parsimony strategy to include all potential parallel intron gains.

In the first stringent parsimony strategy, the state (intron presence or absence) with four occurrences at each position was defined as ancestral, while that with one occurrence was defined as derived. Because this method does not consider multiple hits, it might underestimate the numbers of intron losses and gains. However, the intron gain and loss events obtained are the most unambiguous. Potential intron gains or losses in the outgroup *A. nidulans* were excluded. By this method, we found 159 cases of intron loss and 77 cases of potential intron gain.

Referring to the phylogenetic tree of the five species, we also defined the state (intron presence or absence) with three consistent occurrences across *A. terreus*, *N. fischeri*, and *A. nidulans* as ancestral and that with two occurrences common between *A. flavus* and *A. oryzae* as derived (Fig. 1B). By this method, we found another 45 cases of intron loss and 18 cases of potential intron gain. In total, we found 204 cases of intron loss and 95 potential cases of intron gain.

Although intron losses (204) outnumbered intron gains (95), the frequency of intron gains was still relatively high when compared with most previous studies (Roy and Hartl 2006; Roy et al. 2006; Roy and Penny 2006b, 2007; Stajich and Dietrich 2006; Coulombe-Huntington and Majewski 2007a, b; Loh et al. 2008; Sharpton et al. 2008). We verified the potential intron gains using six other outgroup species, *H. capsulatum*, *N. crassa, C. cinerea*, *C. neoformans*, *S. punctatus*, and *Proterospongia sp.* Although distantly related, the last four species are intron-rich (>4.5 introns per gene) and are thus suitable for the verification of intron gains. We found at least one orthologous gene in the six outgroup genomes for 58 of the 70 genes (a pair of orthologous genes with the same intron gain between *A. flavus* and *A. oryzae* was counted once) that contain the 95 potential cases of intron gain. 11 introns were also found in at least one outgroup species. After eliminating them, there remained 84 unambiguous cases of intron gain in four *Aspergillus* species (Table 1).

Table 1 Intron losses and gains in *Aspergillus*

	Intron loss	Intron gain
A. flavus	0	1
A. oryzae	0	0
A. flavus–A. oryzae[a]	45	17
A. terreus	123	30
N. fischeri	36	36
Total	204	84

[a] Introns commonly lost from or gained in *A. flavus* and *A. oryzae*

Sequence alignments relevant to intron loss and gain are shown in Supplementary materials 1 and 2.

Evidence for mRNA-Mediated Intron Loss

In our study, intron positions that flank gaps in alignments are excluded; thus, all the observed cases are exact intron losses. Therefore, we could not exclude genomic deletion as one of the mechanisms of intron loss in *Aspergillus*. However, we did find evidence supporting mRNA-mediated intron loss by analyzing whether simultaneous loss of adjacent introns was at a significantly higher frequency than random, independent loss of each intron.

Among the 123 cases of intron loss in *A. terreus*, we found nine groups of two adjacent intron losses and one group of four adjacent intron losses. We performed a re-sampling analysis to check whether the pattern of intron loss was caused by random, independent events (Sharpton et al. 2008). In 10,000 re-samplings, we did not obtain patterns with equal or higher number of adjacent intron losses than the observed pattern (i.e., nine groups of two adjacent intron losses and one group of four adjacent intron losses). That is, $P = 0$; adjacent introns in *A. terreus* tend to be lost simultaneously at a significant level. In addition, there were two groups of two adjacent intron losses among the 36 cases of intron loss in *N. fischeri*, and two groups of two adjacent intron losses among the 45 cases of intron loss at the node between *A. flavus* and *A. oryzae*. Re-sampling analyses also indicated that adjacent introns tend to be lost simultaneously ($P < 0.05$ for both the cases). The observed pattern of intron losses is consistent with the model of mRNA-mediated intron loss.

Recombination requires two sequences to be similar. Conceptually, long introns would disturb the in vivo alignment of cDNA with the genomic DNA more strongly than short introns. Thus, short introns are expected to be lost at a higher frequency in the recombination process than long introns (Coulombe-Huntington and Majewski 2007a, b).

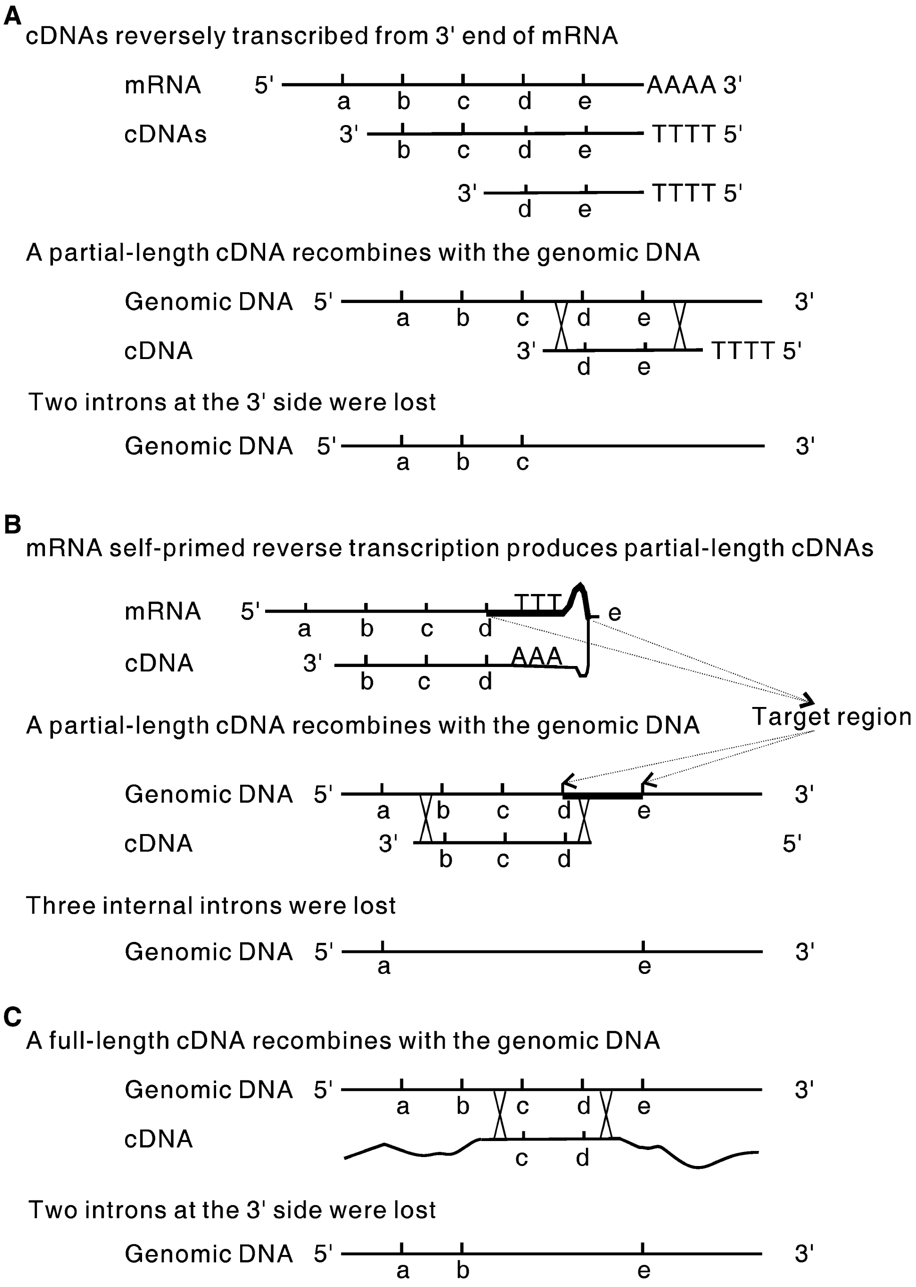

A
cDNAs reversely transcribed from 3' end of mRNA
mRNA 5' a b c d e AAAA 3'
cDNAs 3' b c d e TTTT 5'
3' d e TTTT 5'
A partial-length cDNA recombines with the genomic DNA
Genomic DNA 5' a b c d e 3'
cDNA 3' d e TTTT 5'
Two introns at the 3' side were lost
Genomic DNA 5' a b c 3'
B
mRNA self-primed reverse transcription produces partial-length cDNAs
mRNA 5' a b c d TTT e
cDNA 3' b c d AAA
Target region
A partial-length cDNA recombines with the genomic DNA
Genomic DNA 5' a b c d e 3'
cDNA 3' b c d 5'
Three internal introns were lost
Genomic DNA 5' a e 3'
C
A full-length cDNA recombines with the genomic DNA
Genomic DNA 5' a b c d e 3'
cDNA c d
Two introns at the 3' side were lost
Genomic DNA 5' a b e 3'

Fig. 2 Illustration of the model of mRNA-mediated intron loss and its modified versions. **a** Traditional model of mRNA-mediated intron loss (Fink 1987; Mourier and Jeffares 2003). **b** Reverse transcription is self-primed and so 3′ side introns are not preferentially lost (Feiber et al. 2002; Niu et al. 2005). According to this version, there should be a T-rich segment in the "target region" (highlighted by *bold line*) to pair with the poly-A tail of mRNA. The frequency of T-rich segments in the target regions of genes that lost introns was used to test this version of the model. **c** Most reverse transcripts are full length, and gene conversion tends to occur in interior regions, but not at the two ends of genes

It is difficult to accurately estimate the length of lost introns. We used the average value of the extant introns at the orthologous position as the representative of the length of a lost intron, and the average value of conserved introns at each orthologous position as the control. We found that the lost introns were significantly shorter than the conserved introns ($P = 5 \times 10^{-12}$).

mRNA-Mediated Intron Loss: From Poly-A end, Self-Primed, or…?

The traditional model of mRNA-mediated intron loss and the recently modified version predict different patterns for the position of intron loss (Figs. 2a, b). To distinguish these two models, we studied the positions of the lost introns in these *Aspergillus* species. First, we found that the lost introns are not more abundant at the 3′ side of genes in *A. terreus*, *N. fischeri* or at the node between *A. flavus* and *A. oryzae* ($P > 0.05$ for all three cases). Instead, more than half of them (61% in *A. terreus*, 56% in *N. fischeri*, and 62% at the node between *A. flavus* and *A. oryzae*) occurred at the 5′ side of genes.

However, if the positions of extant introns were highly biased toward 5′ end of genes, preferential loss of introns from 3′ side would also give the appearance of unbiased intron loss when measuring the absolute position of lost

Table 2 Distance from lost introns and conserved introns to the 3′ end of genes (bp)

	Lost introns Mean ± SD Median (*n*)	Conserved introns Mean ± SD Median (*n*)	*P* value
A. terreus	385 ± 284.0	332 ± 286.6	0.010
	328 (123)	265 (3333)	
N. fischeri	369 ± 297.0	332 ± 285.1	0.480
	295 (36)	269 (3333)	
A. flavus[a]	408 ± 388.2	332 ± 188.4	0.267
	376 (45)	266 (3333)	
A. oryzae[a]	396 ± 390.4	328 ± 282.0	0.395
	355 (45)	265 (3333)	

P value was calculated using the Mann–Whitney *U* test

[a] The lost introns at the node between *A. flavus* and *A. oryzae* were calculated according to the gene length of each species separately

introns within genes. Therefore, we compared the position of lost introns and the conserved introns across these five *Aspergillus* species. Still, the lost introns were not more biased toward the 3′ side than the conserved introns ($P > 0.10$ for all the three cases).

According to the traditional model of mRNA-mediated intron loss, reverse transcriptase de-associates with mRNA template at a constant rate, so the possibility of an mRNA position being reverse transcribed depends on the absolute distance to the 3′ end of the gene, rather than the relative distance. Thus, we also compared lost introns and conserved introns for their distance to the 3′ end of genes (as UTRs have not been annotated, we have to use the coding sequences as representatives of genes). Still, we did not find that lost introns are closer to the 3′ end of genes than conserved introns (Table 2). Instead, the 123 introns lost in *A. terreus* are significantly further from the 3′ end of genes than the conserved introns ($P = 0.01$).

These results are consistent with previous studies on fungi (Nielsen et al. 2004; Sharpton et al. 2008), and could be explained by the modified hypothesis of mRNA-mediated intron loss (Fig. 2b). In this version, the loss of an intron depends on the particular secondary structure of the mRNA; 3′ side introns are not necessarily lost more frequently than internal introns or 5′ side introns.

According to the modified hypothesis of mRNA-mediated intron loss (Fig. 2b), the poly-A tail was the primer of reverse transcription and so there must be a poly-T or T-rich segment in the mRNA to pair with it (Fig. 2b). The segment between the position of intron loss and the position of the next downstream intron is termed as the target region (as illustrated between intron *d* and intron *e* in Fig. 2b). Considering that evolutionary traces may be erased or partially erased by subsequent mutations after intron loss during evolution, we do not expect to find a T-rich segment in every target region. Using 10 thymines in a window of 15 nucleotides as the criteria of T-rich, we found 11 T-rich segments across the 10,0879 nucleotides of 159 target regions and 31 T-rich segments across 225,419 nucleotides of other regions of the genes that lost introns in *A. terreus* and *N. fischeri*. A Chi-square test showed that the target regions did not have a significantly higher frequency of T-rich segments than other regions of the genes that lost introns ($P > 0.10$). Similar results were obtained for the introns lost at the node between *A. flavus* and *A. oryzae*. Using nine thymines in a window of 15 nucleotides as the criterion of T-rich gave results consistent with 10 thymines as the criterion. We did not obtain sufficient samples for confident statistical analysis using 11 thymines in a window of 15 nucleotides as the criterion of T-rich. The poly-A tail as the primer of the reverse transcription (Feiber et al. 2002; Niu et al. 2005) is thus not supported by this study.

In the reverse transcription of retroviruses, tRNA molecules are often used as primers (Mak and Kleiman 1997). For mRNA-mediated intron loss, there is also the possibility that cDNA molecules were synthesized by the reverse transcriptase of retroviruses using tRNA molecules as primers. If so, the target regions of the mRNA molecules (Fig. 2b) should have a higher frequency of sequences that can pair with the 3′ end of tRNA molecules. However, we did not observe that pattern (data not shown).

We could look at intron loss as a chemical reaction. The cDNA and the genomic DNA are the substrates, recombination is the chemical reaction, and intron loss is the product. We could not explain the products of the reaction by its substrates. This left the possibility that DNA sequences at different positions of genes have different rates of the chemical reaction. It has been observed that different regions of a chromosome vary greatly in their recombination frequencies (Paigen and Petkov 2010; Szekvolgyi and Nicolas 2010). Thus, we propose that introns flanking recombination hot spots are preferentially lost. A recent survey of all published cases of gene conversion showed that gene conversion tends to occur preferentially in internal regions, rather than at the two ends of genes (Lawson et al. 2009). The possible preference of recombination for internal regions of genes might explain the observation that internal introns were more frequently lost than 3′ end introns (Fig. 2c).

Higher Frequency of Short Duplicates Flanking Gained Introns

Recently, Li et al. (2009) found direct repeats (marks of double-strand break repair) near or within recently gained introns. Based on this observation, they put forward a new hypothesis on intron gain: Exogenous sequences (possibly mitochondrial sequences) inserted during double-strand DNA break repair might create new introns. We tested this hypothesis using the recent intron gains in the four *Aspergillus* genomes.

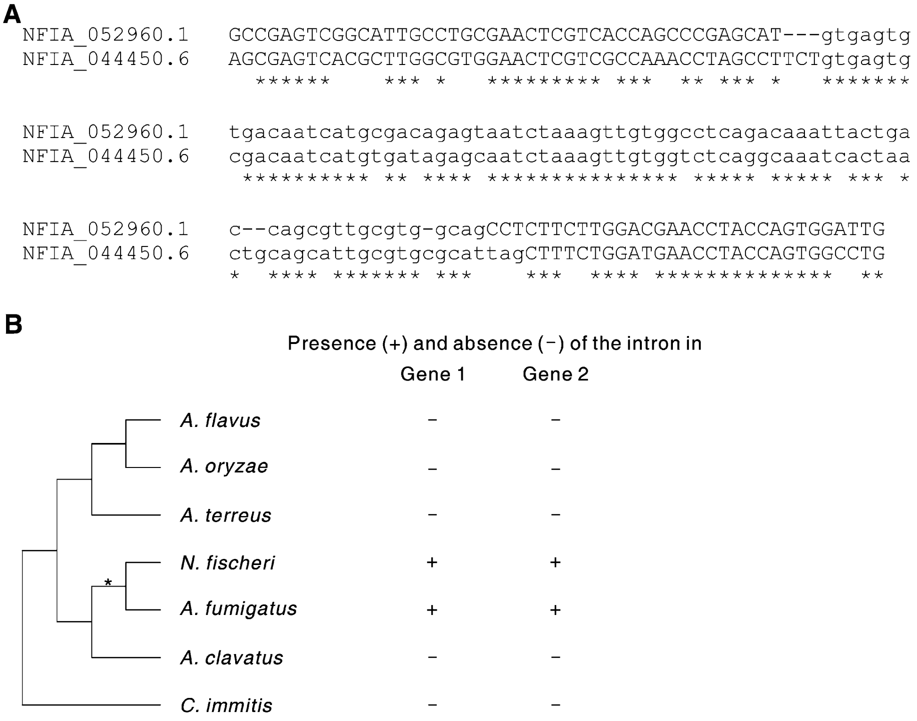
In 17 of the 84 gained introns, we also observed short direct repeats (>5 bp), with each of the repeats near one exon–intron boundary (see Supplementary material 3 for details). To test whether the appearance of these short direct repeats occurred by chance, we surveyed the short direct repeats flanking conserved introns. Among the 13,332 conserved introns across the four nonoutgroup species, we found 1,469 introns with short direct repeats flanking their boundaries. Pearson Chi-square test showed that the appearance of short direct repeats is significantly higher in recently gained introns (20%) than conserved introns (11%) ($P = 0.007$). Thus, our data support the double-strand-break repair model of intron gain (Li et al. 2009).

Failure in Identifying the Sources of Most Recently Gained Intron: Why?

To cover all possible sources of the gained introns, we searched all the supercontigs of the four nonoutgroup *Aspergillus* species and the full nucleotide collection of NCBI (last update on May 13, 2010) for DNA sequences with significant similarity to any of the gained introns we observed. When the Blast program was optimized for searching "highly similar sequences," we did not obtain any homologous sequence except self-hits. However, when the Blast program was optimized for searching "somewhat similar sequences," we obtained 1,696 hits. Manual inspection of self-hits and short hits (<50% in coverage) eliminated most of the hits.

The first intron and flanking exonic segments of gene *NFIA_052960* were found to be similar to the sixth intron and flanking exonic segments of gene *NFIA_044450* (Fig. 3a), while other exonic regions of the two genes are quite different. Because the two introns are very similar in sequence, it is unlikely that the two genes have gained the introns in parallel. It is very likely that one of the two genes first gained an intron and then the other gene acquired the other intron by a later gene conversion event. Phylogenetic analysis of the orthologous genes of these two genes could not determine the temporal order of the two introns (Fig. 3b). For a conclusive analysis, we await the genome sequences of species between *N. fischeri* and *A. clavatus*.

Although only a short match was found to the third introns of gene *AFL2G_10424* and gene *AO090009000242*, it is notable that the matching sequence is a segment of the mitochondrial genome of *A. oryzae* (Fig. 4a). In addition,



```
A
NFIA_052960.1   GCCGAGTCGGCATTGCCTGCGAACTCGTCACCAGCCCGAGCAT---gtgagtg
NFIA_044450.6   AGCGAGTCACGCTTGGCGTGGAACTCGTCGCCAAACCTAGCCTTCTgtgagtg
                  ******    *** *   ********* ***  ** *** *   *******

NFIA_052960.1   tgacaatcatgcgacagagtaatctaaagttgtggcctcagacaaattactga
NFIA_044450.6   cgacaatcatgtgatagagcaatctaaagttgtggtctcaggcaaatcactaa
                 ********** ** **** *************** ***** ***** *** *

NFIA_052960.1   c--cagcgttgcgtg-gcagCCTCTTCTTGGACGAACCTACCAGTGGATTG
NFIA_044450.6   ctgcagcattgcgtgcgcattagCTTTCTGGATGAACCTACCAGTGGCCTG
                *  **** ******* ***    ***  **** **************  **
```

Fig. 3 A case of intron gain by gene conversion. **a** In *N. fischeri*, the first intron and its flanking exonic segments of the gene *NFIA_052960* is similar to the sixth intron and its flanking exonic segment of the gene *NFIA_044450*. Intronic sequences are presented in lowercase and exonic sequences are in *uppercase*. Identical bases are marked by "*" below. **b** Phylogenetic survey indicating that the two introns were gained at the common ancestor of *N. fischeri* and *A. fumigatus*, as marked by "*" at the node. The presence or the absence of an intron in a given position is represented as "+" or "−", respectively. The phylogenetic tree was from the Fungal Genome Initiative of the Broad Institute (http://www.broadinstitute.org/science/projects/fungal-genome-initiative/fungal-genome-initiative) and is not scaled

A

```
AFL2G_10424      gtaa--gtatatttaattttgtga----gccga----ggtttgtagctaacacgatagaactatag
AO090009000242   gtaa--gtatatttaattttgtga----gccga----ggtttgtagctaacacgatagaactatag
NC_008282        TTAATTATATATTTAATTTATTGATATAGTCAATTTTATTTTATAGACTATAATTTATTTATATAG
                  ***   ************  ***    * * *      *** ***   * *   **    *****
```

B

```
AFL2G_10424      AACgtaagtatatttaattttgtgagccgaggtttgtagctaacacgatagaactatagTATAT
AO090009000242   AACgtaagtatatttaattttgtgagccgaggtttgtagctaacacgatagaactatagTATAT
                 ...   .......                                      ...   .......
```

Fig. 4 A possible case of intron gain by insertion of a mitochondrial sequence. **a** The third introns of both the *AFL2G_10424* gene of *A. flavus* and the *AO090009000242* gene of *A. oryzae* are similar to a short segment of the mitochondrial genome of *A. oryzae*. Identical bases are marked by "*" below. **b** There are short duplicates (marked by *dots* below) near the end of these two gained introns. These data support a model of intron gain proposed by (Li et al. 2009): Insertion of a mitochondrial sequence during double-strand repair

there are short duplicates near the end of these gained introns (Fig. 4b). As the sequence similarity between the introns and the mitochondrial DNA segment is not high and the alignment is short and gapped (Fig. 4a), we suggest that the mitochondrial segment is a possible (but not certain) source of the introns. This possible case is similar to that observed in *Daphnia* by Li et al. (2009), and supports their double-strand repair model. Considering the 100% identity between these two introns and their identical position in a pair of orthologous genes, *AFL2G_10424* and *AO090009000242*, we propose that it is an intron gained before the divergence between *A. flavus* and *A. oryzae*.

The sources of most recently gained introns could not be identified. By contrast, we can easily identify most orthologous introns across the five *Aspergillus* species using the same Blast method. This seems to be a common phenomenon. Among the 56 introns gained in duplicated *Arabidopsis* genes, the origin of only one intron could be successfully identified (Knowles and McLysaght 2006). Similarly, among the 28 introns recently gained in *Daphnia* populations, researchers could find homologous sequence to only one intron (Li et al. 2009).

There are three possible reasons for the failure in identifying the source sequences of most recently gained introns. First, a gained intron and its source sequence diverged rapidly and became indiscernible in a very short period of time. In each genome, there are indeed some sequences (e.g., satellite repeats) that are evolving rapidly (Plohl et al. 2008). If the rapidly evolving sequences served as donors in most intron gain events, we would have no hope of finding the source sequences by examining most recently gained introns. However, the intron gains observed by Li et al. (2009) occurred in populations of the same species *Daphnia*, so it is unimaginable for the gained introns and their source sequences to have diverged so quickly to be indiscernible. Second, some cases of intron gain observed may in fact be intron losses in other species. The maximum parsimony method commonly used might overestimate the frequency of intron gain. Most cases of intron gain observed by Knowles and McLysaght (2006) were later questioned by Roy and Penny (2007). In our study, using six distant outgroup species, we eliminated 11 cases from 95 potential cases of intron gain. Extrapolating from this, filtration by more outgroup species might eliminated more cases of intron gain. However, because the choanoflagellate *Proterospongia* sp. is an intron-rich species with about 7 introns per gene, it is unlikely to eliminate most cases of intron gain we observed by examining more outgroup genomes. Last, the source sequences are not present in the subject sequences that we searched. Genome sequences for most eukaryotes are incomplete because of difficulties in sequencing repeated DNA in the heterochromatin (Hoskins et al. 2007). The *Daphnia* genome sequences Li and colleagues used and the *Aspergillus* genome sequences we used are all draft assemblies. There are many small gaps to be filled. In addition, genomes are not constant in evolution, some sequences may be lost and others be inserted (Eickbush and Furano 2002; Volff et al. 2003).

Parallel Intron Gains: Not Frequent in *Aspergillus*

Parallel intron gain was believed to be very rare (Roy and Gilbert 2006; Roy and Penny 2007), so the cases illustrated in Fig. 1C were often discarded as ambiguous cases. However, recent studies on intron gains in *Daphnia* populations (Li et al. 2009) and in proto-mitochondrial derived nuclear genes (Ahmadinejad et al. 2010), indicate that parallel intron gain might be frequent. If we assume that parallel intron gain occurs at a high frequency, the cases illustrated in Fig. 1C should be considered as potential cases of parallel intron gain. By analyzing the 2,144 groups of orthologs, we found 15 pairs of potentially parallel gained introns between *A. terreus* and *N. fischeri*. In addition, the 18 cases illustrated in Fig. 1B might also be parallel intron gains after the divergence between *A. flavus* and *A. oryzae*.

A pair of parallel gained introns originated from independent events, so they are unlikely to have the same source sequence. The identity between introns gained in parallel should be very low as compared with that of truly orthologous introns descended from one common ancestral intron. However, the identities between potential parallel gained introns we observed are not significantly lower than those between conserved introns (Mann–Whitney *U* test, $P > 0.10$ for both the cases), indicating that most of the shared introns could not be attributed to parallel gain. Among the 18 pairs of commonly gained introns between *A. flavus* and *A. oryzae*, the minimal value of identity is 0.94, so all these pairs of introns are likely to be orthologous rather than gained in parallel. Among the 15 pairs of gained introns between *A. terreus* and *N. fischeri*, the minimal value of identity is 0.33. We could not exclude the possibility that a few pairs of introns between *A. terreus* and *N. fischeri* were gained in parallel. Compared with the 84 cases of nonparallel intron gains (Table 1), parallel intron gain, if it exists, is much less frequent in *Aspergillus*. This result is distinct from the recent observation of frequent parallel intron gains in *Daphnia* (Li et al. 2009), which may reflect lineage specificity.

Others Concerns on Intron Gain

We have also attempted to test two other recent hypotheses on intron gain using our data in *Aspergillus*. Catania et al. proposed that in-frame stop codon in a new intron might facilitate fixation of the intron because unspliced transcripts containing the intron would be degraded by nonsense-mediated mRNA decay (Catania and Lynch 2008; Catania et al. 2009). This hypothesis is supported by a recent study in *Drosophila* (Farlow et al. 2010). However, we did not find significantly higher numbers of in-frame stop codons in recently gained introns compared to those in conserved introns (data not shown). The role of nonsense-mediated mRNA decay on intron gain may have lineage specificity.

Niu (2007) proposed that having introns may be advantageous because the splicing of introns could reduce the frequency of R-loop formation, a threat to genome stability. The R-loop forming potential of a gene could be estimated by comparing the free energies of an R-loop (RNA–DNA hybrids and free DNA) and its alternate (DNA duplexes and free RNA) (Huppert 2008). By this method, we analyzed the recently gained introns in *Aspergillus* and *Daphnia* (Li et al. 2009) and found that introns were not preferentially gained at exonic regions with high R-loop formation potentials (data not shown).

Intron density has been found to be correlated with the logarithm of generation time (Jeffares et al. 2006). Jeffares et al. (2006, 2008) suggested that intron gain/loss may be under selection for quick response to external stimuli. The generation time (1–2 h) (Williams 1975) and intron density (~2 introns per gene) of *Aspergillus* fit well with the correlation shown in Fig. 3 of Jeffares et al. (2006).

Acknowledgments We thank the anonymous reviewers for helpful comments and Jian-Li Cao for his kind help. This work was supported by the Fundamental Research Funds for the Central Universities (2009SAP-4) and the Program for New Century Excellent Talents in University (NCET-07-0094).

References

Ahmadinejad N, Dagan T, Gruenheit N, Martin W, Gabaldon T (2010) Evolution of spliceosomal introns following endosymbiotic gene transfer. BMC Evol Biol 10:57

Basu MK, Rogozin IB, Deusch O, Dagan T, Martin W, Koonin EV (2008) Evolutionary dynamics of introns in plastid-derived genes in plants: saturation nearly reached but slow intron gain continues. Mol Biol Evol 25:111–119

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

Catania F, Lynch M (2008) Where do introns come from? PLoS Biol 6:e283

Catania F, Gao X, Scofield DG (2009) Endogenous mechanisms for the origins of spliceosomal introns. J Hered 100:591–596

Coulombe-Huntington J, Majewski J (2007a) Characterization of intron loss events in mammals. Genome Res 17:23–32

Coulombe-Huntington J, Majewski J (2007b) Intron loss and gain in *Drosophila*. Mol Biol Evol 24:2842–2850

de Souza SJ (2003) The emergence of a synthetic theory of intron evolution. Genetica 118:117–121

Derr LK, Strathern JN, Garfinkel DJ (1991) RNA-mediated recombination in *S. cerevisiae*. Cell 67:355–364

Eickbush TH, Furano AV (2002) Fruit flies and humans respond differently to retrotransposons. Curr Opin Genet Dev 12:669

Farlow A, Meduri E, Dolezal M, Hua L, Schlotterer C (2010) Nonsense-mediated decay enables intron gain in *Drosophila*. PLoS Genet 6:e1000819

Feiber AL, Rangarajan J, Vaughn JC (2002) The evolution of single-copy *Drosophila* nuclear *4f-rnp* genes: Spliceosomal intron losses create polymorphic alleles. J Mol Evol 55:401–413

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

Gao X, Lynch M (2009) Ubiquitous internal gene duplication and intron creation in eukaryotes. Proc Natl Acad Sci USA 49: 20818–20823

Hankeln T, Friedl H, Ebersberger I, Martin J, Schmidt ER (1997) A variable intron distribution in globin genes of *Chironomus*: evidence for recent intron gain. Gene 205:151–160

Hoskins RA, Carlson JW, Kennedy C, Acevedo D, Evans-Holm M, Frise E, Wan KH, Park S, Mendez-Lago M, Rossi F, Villasante A, Dimitri P, Karpen GH, Celniker SE (2007) Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. Science 316:1625–1628

Huppert JL (2008) Thermodynamic prediction of RNA-DNA duplex-forming regions in the human genome. Mol Biosyst 4:686–691

Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, Roy SW (2008) Origin of introns by 'intronization' of exonic sequences. Trends Genet 24:378–381

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

Jeffares DC, Penkett CJ, Bahler J (2008) Rapidly regulated genes are intron poor. Trends Genet 24:375–378

Knowles DG, McLysaght A (2006) High rate of recent intron gain and loss in simultaneously duplicated *Arabidopsis* genes. Mol Biol Evol 23:1548–1557

Lane CE, van den Heuvel K, Kozera C, Curtis BA, Parsons BJ, Bowman S, Archibald JM (2007) Nucleomorph genome of

Hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function. Proc Natl Acad Sci USA 104:19908–19913

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and clustal X version 2.0. Bioinformatics 23:2947–2948

Lawson MJ, Jiao J, Fan WG, Zhang LQ (2009) A pattern analysis of gene conversion literature. Comp Funct Genom 2009:761512

Li W, Tucker AE, Sung W, Thomas WK, Lynch M (2009) Extensive, recent intron gains in *Daphnia* populations. Science 326: 1260–1262

Lin H, Zhu W, Silva J, Gu X, Buell CR (2006) Intron gain and loss in segmentally duplicated genes in rice. Genome Biol 7:R41

Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M (2002) Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. Proc Natl Acad Sci USA 99:8121–8126

Loh Y-H, Brenner S, Venkatesh B (2008) Investigation of loss and gain of introns in the compact genomes of Pufferfishes (Fugu and *Tetraodon*). Mol Biol Evol 25:526–535

Long M, Rosenberg C, Gilbert W (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. Proc Natl Acad Sci USA 92:12495–12499

Mak J, Kleiman L (1997) Primer tRNAs for reverse transcription. J Virol 71:8087–8095

Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, Best AA, Cande WZ, Chen F, Cipriano MJ, Davids BJ, Dawson SC, Elmendorf HG, Hehl AB, Holder ME, Huse SM, Kim UU, Lasek-Nesselquist E, Manning G, Nigam A, Nixon JEJ, Palm D, Passamaneck NE, Prabhu A, Reich CI, Reiner DS, Samuelson J, Svard SG, Sogin ML (2007) Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. Science 317:1921–1926

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

Nielsen CB, Friedman B, Birren B, Burge CB, Galagan JE (2004) Patterns of intron gain and loss in fungi. PLoS Biol 2:e422

Niu D-K (2007) Protecting exons from deleterious R-loops: a potential advantage of having introns. Biol Direct 2:11

Niu D-K, Hou W-R, Li S-W (2005) mRNA-mediated intron losses: evidence from extraordinarily large exons. Mol Biol Evol 22:1475–1481

Paigen K, Petkov P (2010) Mammalian recombination hot spots: properties, control and evolution. Nat Rev Genet 11:221–233

Plohl M, Luchetti A, Mestrovic N, Mantovani B (2008) Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene 409:72–82

Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94

Rodriguez-Trelles F, Tarro R, Ayala FJ (2006) Origins and evolution of spliceosomal introns. Annu Rev Genet 40:47–76

Roy SW (2004) The origin of recent introns: transposons? Genome Biol 5:251

Roy SW (2009) Intronization, de-intronization and intron sliding are rare in *Cryptococcus*. BMC Evol Biol 9:192

Roy SW, Gilbert W (2005a) The pattern of intron loss. Proc Natl Acad Sci USA 102:713–718

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

Roy SW, Gilbert W (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet 7:211–221

Roy SW, Hartl DL (2006) Very little intron loss/gain in *Plasmodium*: intron loss/gain mutation rates and intron number. Genome Res 16:750–756

Roy SW, Irimia M (2009) Mystery of intron gain: new data and new models. Trends Genet 25:67–73

Roy SW, Penny D (2006a) Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution. Genome Res 16:1270–1275

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

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

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

Scherer S (2008) A short guide to the human genome. Cold Spring Harbor Laboratory Press, New York

Sharpton TJ, Neafsey DE, Galagan JE, Taylor JW (2008) Mechanisms of intron gain and loss in *Cryptococcus*. Genome Biol 9:R24

Stajich JE, Dietrich FS (2006) Evidence of mRNA-mediated intron loss in the human-pathogenic fungus *Cryptococcus neoformans*. Eukaryot Cell 5:789–793

Stajich JE, Dietrich FS, Roy SW (2007) Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. Genome Biol 8:R223

Szekvolgyi L, Nicolas A (2010) From meiosis to postmeiotic events: homologous recombination is obligatory but flexible. FEBS J 277:571–589

Volff JN, Bouneau L, Ozouf-Costaz C, Fischer C (2003) Diversity of retrotransposable elements in compact pufferfish genomes. Trends Genet 19:674–678

Wilkerson MD, Ru Y, Brendel VP (2009) Common introns within orthologous genes: software and application to plants. Brief. Bioinform 10:631–644

Williams RL (1975) The shoot apex and leaf growth. Cambridge University Press, London