# Higher Intron Loss Rate in *Arabidopsis thaliana* Than *A. lyrata* Is Consistent with Stronger Selection for a Smaller Genome

Jeffrey A. Fawcett,[1,2,†] Pierre Rouzé,[1,2] and Yves Van de Peer[1,2,*]

[1]Department of Plant Systems Biology, VIB, Gent, Belgium

[2]Department of Plant Biotechnology and Bioinformatics, Ghent University, Gent, Belgium

†Present address: Department of Evolutionary Studies of Biosystems, Graduate University for Advanced Studies, Hayama, Kanagawa, Japan

*__Corresponding author:__ E-mail: yves.vandepeer@psb.vib-ugent.be.

**Associate editor:** Aoife McLysaght

## Abstract

The number of introns varies considerably among different organisms. This can be explained by the differences in the rates of intron gain and loss. Two factors that are likely to influence these rates are selection for or against introns and the mutation rate that generates the novel intron or the intronless copy. Although it has been speculated that stronger selection for a compact genome might result in a higher rate of intron loss and a lower rate of intron gain, clear evidence is lacking, and the role of selection in determining these rates has not been established. Here, we studied the gain and loss of introns in the two closely related species *Arabidopsis thaliana* and *A. lyrata* as it was recently shown that *A. thaliana* has been undergoing a faster genome reduction driven by selection. We found that *A. thaliana* has lost six times more introns than *A. lyrata* since the divergence of the two species, but gained very few introns. We suggest that stronger selection for genome reduction probably resulted in the much higher intron loss rate in *A. thaliana*, although further analysis is required as we could not find evidence that the loss rate increased in *A. thaliana* as opposed to having decreased in *A. lyrata* compared with the rate in the common ancestor. We also examined the pattern of the intron gains and losses to better understand the mechanisms by which they occur. Microsimilarity was detected between the splice sites of several gained and lost introns, suggesting that nonhomologous end joining repair of double-strand breaks might be a common pathway not only for intron gain but also for intron loss.

**Key words:** *Arabidopsis thaliana*, *Arabidopsis lyrata*, intron, selection, genome reduction.

## Introduction

A central question in molecular evolutionary biology is whether various differences between species are caused by differences in mutation rates or by differences in selective pressure including population genetic processes that result in more or less efficient selection. The uneven phylogenetic distribution of spliceosomal introns, ranging from only two detected in the entire *Trypanosoma brucei* genome to >100,000 in various vertebrate genomes (Roy and Gilbert 2006; Siegel et al. 2010), is one such case. Introns are sequences within genes that are spliced out during transcription and are prevalent throughout eukaryotic genomes, although their origin, role, and function are still under debate. The difference in the intron density across species can be explained by the differences in the rates of intron gain and loss (Jeffares et al. 2006; Roy 2006; Roy and Gilbert 2006). Whether a given intron gain or loss mutation reaches fixation depends on a multitude of factors. Some introns contain regulatory sequences, encode other genes or noncoding RNAs, or are required for various processes such as alternative splicing, the processing and export of the mRNA, and translation efficiency (Le Hir et al. 2003; Nott et al. 2004; Cenik et al. 2011). The loss of such introns is likely to be selected against, whereas the gain of an intron that happens to improve the function of the gene will be more likely to be fixed. On the other hand, it is not clear how many introns are functional, and it has also been suggested that many introns are nonessential and instead impose a cost on genes by prolonging transcription (Lynch 2002). Thus, introns are thought to be disfavored in genes that have to be rapidly regulated (Chen et al. 2005; Jeffares et al. 2008). However, in addition to such intron-specific and gene-specific factors, organism-specific factors that affect the entire genome are likely to be crucial in determining the genome-wide rates of intron gain and loss (Jeffares et al. 2006). One such factor might be selection. For instance, it has been proposed that population genetic settings in which selection is less efficient (e.g., smaller effective population size) should result in a higher intron gain rate and a lower intron loss rate because nonfunctional introns should be more likely to be gained and less likely to be lost (Lynch 2002; Lynch and Conery 2003; Omilian et al. 2008). By contrast, stronger or more efficient selection, such as selection for reduced genomes (e.g., due to a shorter generation time), might result in a higher intron loss rate (and a lower intron gain rate) because an intron loss would have a higher chance of reaching fixation if the intron is nonessential or if the selection is strong enough (Jeffares et al. 2006; Lane et al. 2007). However, although several species with a compact genome are intron

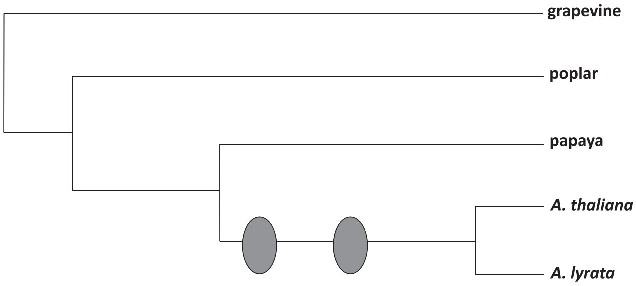


FIG. 1. Relationship of species used in this study. The gray ovals represent WGDs in the *Arabidopsis* lineage. See text for details.

poor, there are also many species that have a small genome but are intron rich and also species with a large genome but very few introns (Jeffares et al. 2006; Carlton et al. 2007). In addition, although the rates of gain and loss are predicted to be inversely correlated if selection were the major determinant, many recent studies have observed positive correlations between the rates of gain and loss (Roy and Hartl 2006; Carmel et al. 2007b; Roy and Penny 2007b; Stajich et al. 2007; Farlow et al. 2011). Therefore, mechanistic factors that influence the rate of mutations generating novel introns or removing introns have been proposed to be key determinants in shaping intron density (Roy and Hartl 2006; Carmel et al. 2007a, 2007b; Roy and Penny 2007b; Stajich et al. 2007; Farlow et al. 2011). As such, the link between genome reduction and intron loss and the role of selection on the rates of intron gain and loss is still unclear.

Here, we utilized the recently sequenced genome of *Arabidopsis lyrata* (Hu et al. 2011), which shares a common ancestry with *A. thaliana* at ~10 Ma (Wright et al. 2002; Ossowski et al. 2010), to study the pattern and rate of intron gain and loss in the two species since their divergence. These two species are characterized by their small genome size, despite sharing two whole-genome duplications (WGDs) since their divergence from papaya (Tang et al. 2008) (fig. 1), with the genome of *A. thaliana* being 1.5–2 times smaller than that of *A. lyrata*. A recent comparison of the two genomes coupled with population studies of *A. thaliana* suggested the presence of a much stronger selection for genome downsizing acting on a genome-wide scale in *A. thaliana* compared with *A. lyrata* (Hu et al. 2011). Thus, the genomes of these two species allow us to test whether the difference in the selection for genome compactness will be reflected in a difference in the rate of intron gain and loss. Furthermore, comparison of these closely related genomes provides an opportunity to study the mechanism of intron gain and loss.

## Materials and Methods

### Data Set

Whole-genome sequences and the annotation of protein-coding genes for the genomes of *A. thaliana*, *Carica papaya*, *Populus trichocarpa*, and *Vitis vinifera* were obtained from the PLAZA database (http://bioinformatics.psb.ugent.be/plaza/) (Proost et al. 2009), and *A. lyrata* from JGI (http://www.phytozome.net/alyrata.php). The TAIR8 release was used for *A. thaliana* (Swarbreck et al. 2008), although we did confirm that all the intron gain and loss events did not change in the TAIR10 release.

### Detection of Orthologs

One-to-one orthologous gene pairs between *A. thaliana* and *A. lyrata* were identified based on collinearity. First, homologous gene pairs within or between species that have a BlastP hit with a cutoff of 1e-05 and whose alignment covers at least half of both genes were extracted. The synonymous substitution rate ($K_s$) was calculated for each pair using CODEML from the PAML package (Yang 1997), and the $K_s$ value with the lowest log value after ten runs was taken. The *A. thaliana*–*A. lyrata* homologous gene pairs with $K_s$ <0.25 (the peak of the $K_s$ distribution of putative orthologs of *A. thaliana* and *A. lyrata* was 0.12 to 0.13) were used as an input to run i-ADHoRe 2.0 (Simillion et al. 2008), which identifies homologous genomic segments based on gene homology matrices. Only scaffolds one to eight were considered for *A. lyrata*. The gap size was set to seven genes, and the minimum number of homologs (anchors) to define a homologous segment was set to seven genes, and the *P* value cutoff used was 0.001. Next, each collinear gene pair (anchors) where either gene had more than one closely related homolog (BlastP hit of <1e-05, alignment covers at least half of both genes, and $K_s$ <0.25) in the same species or in the other species were removed, and thus, one-to-one orthologs that are collinear and have a $K_s$ of <0.25 were retained. Orthologs between *A. thaliana* or *A. lyrata* and papaya, poplar, and grapevine were identified by running Inparanoid (Remm et al. 2001) between *A. thaliana* or *A. lyrata* against papaya, poplar, or grapevine, using rice as an outgroup.

### Detection of Paralogs

First, paralogs in *A. thaliana* and *A. lyrata* created by the most recent WGD in the *Arabidopsis* lineage were identified based on collinearity. All-against-all BlastP was run with all amino acid sequences of *A. thaliana*, *A. lyrata*, papaya, poplar, and grapevine with a cutoff of 1e-05, and those matching the criteria of Li et al. (2001) were retained. i-ADHoRe 2.0 (Simillion et al. 2008) was run using the retained gene pairs to detect collinear segments as mentioned above, except that the gap size was set to 20 genes. Each collinear gene pair where both genes were from *A. thaliana* or from *A. lyrata* was extracted, and clusters of collinear gene pairs were created by single-linkage clustering. Thus, each cluster would include genes of *A. thaliana* and *A. lyrata* that were created by multiple rounds of WGDs (Proost et al. 2011). The pairwise $K_s$ of each *A. thaliana* and *A. lyrata* pair within a cluster was calculated, and the pair with the lowest $K_s$, if the $K_s$ was between 0.4 and 0.8, was retained as duplicates derived from the most recent WGD. The cutoff of 0.8 might appear to be rather stringent. However, *Arabidopsis* has undergone two WGDs since the *Arabidopsis*–papaya divergence (Tang et al. 2008), and it turned out to be difficult to discriminate the pairs generated by the two different WGDs based on $K_s$, unless all four duplicates remained intact as anchors. We opted for a slightly conservative cutoff as inclusion of pairs from the older WGD will affect the estimate of the intron loss rate. It must be noted that the rate slightly differed depending on the cutoff, and it is likely that there are still pairs from the older WGD present, and thus, the estimated intron loss rates should be treated as an approximate.

### Detection of Unique Intron Positions

For a given gene pair, the amino acid sequences of the two genes were aligned using bl2seq, and the intron positions were mapped onto the alignment. The longest transcript was used when there were multiple splice forms. Each intron position was further examined and shared introns and unique introns were identified according to the following criteria: 1) intron positions within five alignment positions from the alignment borders (including the beginning and end of the genes) were not retained, 2) neighboring intron positions within five alignment positions were not retained, 3) only introns with canonical splice sites, GT-AG or GC-AG, were retained, 4) intron positions not present in the alignment were not retained, 5) ten flanking alignment positions on both sides of the introns were examined, and the intron was retained only if more than half of the alignment positions were identical amino acids for both sides (a gap is treated as a mismatch), thus removing introns in poorly aligned regions and introns neighboring alignment gaps.

### Determining Intron Gain and Loss

Each unique intron position was determined whether it represents an intron gain or loss by using outgroup sequences. If an outgroup sequence also contains an intron in the same alignment position, it can be considered that that intron was lost, whereas if it does not, it can be considered that that intron was gained. As outgroup sequences, orthologous genes from papaya, poplar, and grapevine were considered. When looking for intron gains and losses in orthologs of *A. thaliana* and *A. lyrata*, homologs within the two species with a $K_s$ of 0.3–1.0 were also considered (intended to capture those genes that were duplicated in the ancestral *Arabidopsis* lineage after the *Arabidopsis*-papaya divergence). The amino acid sequence of the gene with the unique intron position was aligned with each outgroup sequence, and it was examined whether an intron is present in the same alignment position of the outgroup sequence or not, according to the same criteria outlined above. Only cases where at least one outgroup sequence could be considered and every outgroup sequence supported the same conclusion (shared or not shared) were retained and classified as gains or losses. The same was done also for the shared intron positions to estimate the rates of gains and losses, and these shared introns that were confirmed by outgroups were used for comparing the different characteristics between shared and lost introns. Each identified intron gain and loss was also checked manually.

### Rate of Intron Loss

The rate of intron loss was calculated based on the number of intron loss ($L$), number of shared introns that could be confirmed by outgroups ($S$), and time ($T$). The rate of loss in *A. thaliana* or *A. lyrata* that took place since the divergence of the two species was calculated as follows: $L/(S + L) \times T$, where $T = 10$ myr. The rate of loss in the ohnologs created by the most recent WGD in the *Arabidopsis* lineage prior to the *A. thaliana*–*A. lyrata* divergence was calculated as follows: $L/(S \times 2 + L \times 2) \times T \times 2$, where $T = 30$ myr. Note that for the ohnologs, the introns can be lost in both lineages. Also, the number of intron losses that were found to be actually lost after the *A. thaliana*–*A. lyrata* divergence was not included in $L$.

### Randomization of Intron Loss

A randomization test was performed to infer the probability of observing the pattern of intron loss in *A. thaliana* by chance. A data set including every gene which had at least one shared or lost intron was created to represent the ancestral state. An intron to be deleted was randomly picked, and this was repeated according to the observed number of introns that were lost. This cycle was repeated 10,000 times. As we found that short introns are much more likely to be lost, for each analysis, the resampling was also performed allowing only introns $<150$ bp to be lost. In all analyses, this extra filter did not affect whether the result was significant or not.

### 5′ or 3′ Bias of Intron Loss

The relative position along the coding sequences (CDS) (from 0 to 1) of the lost and shared introns was calculated as (length from 5′)/(total length of CDS), and the median value of all lost introns was calculated. The probability of obtaining a smaller (more 5′) or larger (more 3′) median value than the observed value was derived by randomization as described above. Also, each lost intron was assigned as being either 5′ or 3′ in relation to the number of introns. Suppose that there are five introns, the first two (starting from the most 5′) would be labeled as 5′, the third will not be counted, and the last two would be labeled as 3′. If there are six introns, the first three would be 5′ and the last three would be 3′. The probability of obtaining more 5′ or 3′ introns than observed was derived by randomization.

## Results

### Rates of Intron Gain and Loss

We searched for intron gains and losses that occurred in the genomes of *A. thaliana* and *A. lyrata* since the divergence of the two species. We first identified introns unique to one of the two species by mapping the intron positions onto the amino acid alignment of the orthologous genes. These unique introns were classified as gained in that species or lost in the other species by examining whether an outgroup sequence contained an intron in the same alignment position or not (see Materials and Methods). Orthologous genes of *Carica papaya*, *Populus trichocarpa*, *Vitis vinifera*, or the outparalogs of *A. thaliana* and *A. lyrata* were used as outgroups (fig. 1). Of 18,330 orthologous gene pairs, we identified 90 and 15 intron losses in *A. thaliana* and *A. lyrata*, respectively, compared with 2 and 9 putative

**Table 1.** Number of Shared, Lost, and Gained Introns.

| | Pair of Genes | Shared | Unique | Shared, Confirmed | Loss | Rate (intron per year) | Gain |
|---|---|---|---|---|---|---|---|
| *A. thaliana* (*A. thaliana*–*A. lyrata* orthologs) | 18,330 | 81,239 | 42 | 54,941 | 90 | $1.64 \times 10^{-10}$ | 2 |
| *A. lyrata* (*A. thaliana*–*A. lyrata* orthologs) | 18,330 | 81,239 | 227 | 54,941 | 15 | $2.73 \times 10^{-11}$ | 6[a] |
| *A. thaliana* (*A. thaliana*–*A. thaliana* ohnologs) | 1,426 | 6,597 | 168 | 5,115 | 123 | $1.85 \times 10^{-10}$ | 4 |
| *A. lyrata* (*A. lyrata*–*A. lyrata* ohnologs) | 1,470 | 6,507 | 165 | 5,064 | 122 | $1.93 \times 10^{-10}$ | 5 |

[a] Those that were initially identified as intron gains but that are likely to be artifacts are not included (see table 2).

intron gains (table 1). Each intron gain and loss was manually inspected and we confirmed that all losses were not due to any gene prediction or alignment-related artifact, although we found that three of the gains in *A. lyrata* could also be explained by gene prediction-related artifacts, including one that is probably pseudogenized (see below and table 2). It is important to note that unique intron positions that are methodological artifacts will be constantly classified as intron gains and not losses because such intron positions are unlikely to be present in outgroup sequences. Although parallel gains at the same site (as in Li et al. 2009) or losses of the same introns might affect the results, requiring the support of three outgroups instead of one did not change the overall pattern (larger number of losses than gains and larger number of losses in *A. thaliana* than in *A. lyrata*). No parallel gains or losses were detected between the outparalogs of *A. thaliana* and *A. lyrata* derived from the most recent WGD in the ancestral *Arabidopsis* lineage. We also examined U12 introns in *A. thaliana* that were deposited in U12DB (Alioto 2007) and all candidate U12 introns in *A. lyrata* that begin with (G/A)TATCCT, separately by manual inspection, but did not find any gain or loss of such introns in either species. The rates of intron gain and loss can be calculated considering the number of introns shared between *A. thaliana* and *A. lyrata* that are also present in an outgroup. Assuming the divergence of *A. thaliana* and *A. lyrata* at 10 Ma (Wright et al. 2002; Ossowski et al. 2010), we obtained an intron loss rate of $1.64 \times 10^{-10}$ and $2.73 \times 10^{-11}$ introns per year for *A. thaliana* and *A. lyrata*, respectively (table 1). As the numbers of intron gains are very few, in terms of comparing rates, we will only focus on the rates of intron losses.

We also determined the intron loss rate since the most recent WGD in the ancestral *Arabidopsis* lineage prior to the *A. thaliana*–*A. lyrata* divergence. Following the procedure described above (also see Materials and Methods), we identified 4 intron gains and 123 intron losses out of 1,426 paralogous pairs in *A. thaliana* that were created by the most recent WGD and are retained in duplicated blocks (table 1). Assuming that the WGD took place at 40 Ma (Fawcett et al. 2009) and that *A. thaliana* and *A. lyrata* diverged 10 Ma (Wright et al. 2002; Ossowski et al. 2010) (thus, 30 Ma between the WGD and the *A. thaliana*–*A. lyrata* divergence), we obtained an intron loss rate of $1.85 \times 10^{-10}$ introns per year. To ensure that there was no technical bias between the two species in detecting intron gains and losses, the same procedure was repeated using 1,470 ohnolog (WGD-derived paralog) pairs in *A. lyrata* instead of *A. thaliana.* This resulted in an intron loss rate of $1.93 \times 10^{-10}$ introns per year, comparable to the rate calculated using the *A. thaliana* ohnologs.

### *Pattern of Intron Gain*

We further examined the identified intron gains and losses. In particular, the recent gains and losses that occurred after the *A. thaliana*–*A. lyrata* divergence should be informative in understanding the mechanisms of intron gain and loss. Although the mechanism of intron gain had been largely unknown, recent studies have shown that nonhomologous end-joining (NHEJ) repair of double-stranded breaks (DSBs) might be a common pathway for intron gain (fig. 2) (Li et al.

**Table 2.** Unique Introns in *A. thaliana* and *A. lyrata* Classified as Intron Gains.

| Gained Intron[a] | Ortholog | Similarity to Other Sequences and Various Remarks |
|---|---|---|
| AT1G32450.1.5 | fgenesh1_pm.C_scaffold_1002596 | No similarity |
| AT1G70530.1.1 | scaffold_201941.1 | No similarity |
| fgenesh2_kg.4__1154__AT2G31690.1.1 | AT2G31690.1 | DNA-transposon, covers entire intron |
| fgenesh2_kg.6__1464__AT5G14870.1.5 | AT5G14870.1 | No similarity |
| fgenesh2_kg.7__1979__AT4G23180.1.1 | AT4G23180.1 | Possible artifact related to gene prediction, predicted intron is probably partly exonic and partly intergenic |
| fgenesh2_kg.7__319__AT4G37340.1.5 | AT4G37340.1 | LTR retrotransposon covers entire intron, gene is probably pseudogenized |
| fgenesh2_kg.7__3257__AT5G41650.1.2 | AT5G41650.1 | DNA transposon, does not cover entire intron |
| fgenesh2_kg.8__2511__AT5G64790.1.3 | AT5G64790.1 | No similarity |
| scaffold_403565.1.1 | AT2G46110.1 | No similarity |
| scaffold_700682.1.4 | AT4G34490.1 | ~40 nt tandem duplication of neighboring exonic sequence, no frameshifts and could be part of exon |
| scaffold_701502.1.2 | AT4G27670.1 | Unclassified repetitive element, covers entire intron |

[a] The last digit on the gene identifier represents the intron number that was gained, starting counting from 1 from the 5′ most intron.

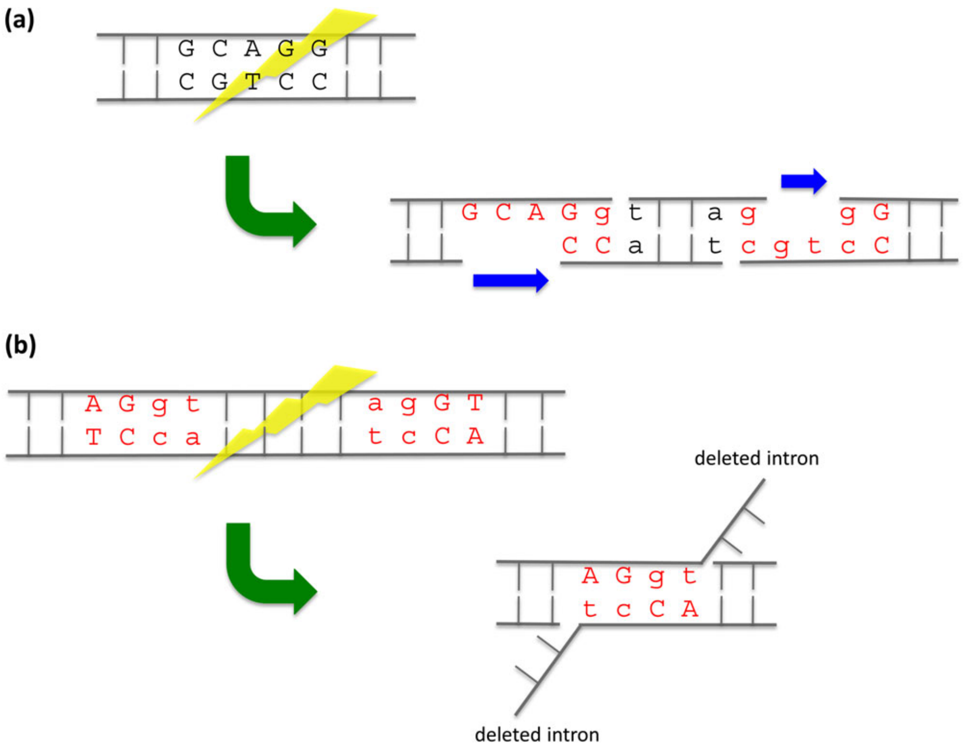
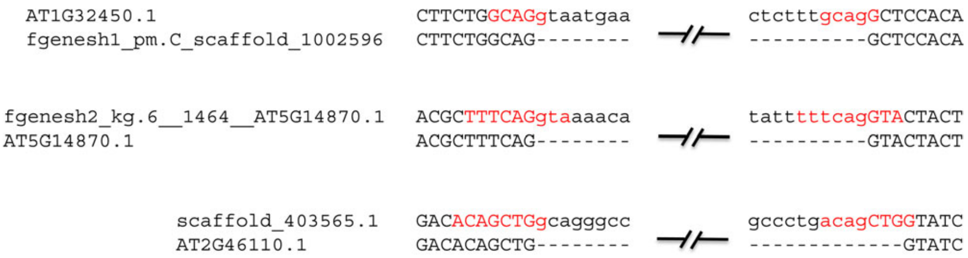


Fig. 2. Gain and loss of introns by NHEJ repair of DSBs. NHEJ is an error-prone pathway of the repair of DSBs that can lead to various insertions and deletions. Single-stranded overhangs are generated by resection, and the repair is often facilitated by the pairing of short stretches of complementary sequences (microsimilarity) in the overhangs (Lieber 2010). (*a*) DNA fragments can be captured during the repair when microsimilarity pairing occurs between the overhangs and exogenous DNA, which might result in the insertion being flanked by short direct repeats. This can result in a novel intron if it occurs in the exon and the inserted sequence happens to satisfy splicing requirements. (*b*) Intronic DSBs might be stabilized by microsimilarity pairing between the 5′ and 3′ splice sites (e.g., AG|GT), resulting in the precise deletion of the intron, which leaves only one of the AGGT motifs. Exonic nucleotides are in upper case and intronic nucleotides are in lower case. Stretches of nucleotides exhibiting microsimilarity flanking the gained or lost introns are in red.

2009; Curtis and Archibald 2010; Farlow et al. 2010, 2011; Zhang et al. 2010). One signature of NHEJ-mediated intron gain is that a short direct repeat is often created at the insertion site of the intron. We examined the novel introns gained in *A. thaliana* or *A. lyrata* since the divergence of the two species and indeed identified three cases where 5, 9, and 8 bp, respectively, spanning the novel intron–exon boundary were repeated (fig. 3), consistent with observations in the previous studies (Li et al. 2009; Farlow et al. 2010; Zhang et al. 2010). Our observation provides further evidence that NHEJ is a widespread pathway for intron gain. Another proposed pathway for intron gain is the insertion of transposable elements (Roy 2004). We noticed that four of the putative novel intron sequences in *A. lyrata* had similarity to repetitive elements. However, based on transcript sequences, we could not confirm that these are actually spliced introns as there is not much transcript data available for *A. lyrata*. Careful inspection suggested that one gene is likely a pseudogene (table 2).

## Pattern of Intron Loss

We next studied the intron losses identified in *A. thaliana* that occurred after the *A. thaliana*–*A. lyrata* divergence by comparing these intron losses with the introns that were retained in both species (and also present in one or more outgroup sequences). Intron deletion is likely to occur by


```
AT1G32450.1                    CTTCTGGCAGgtaatgaa  —//—  ctctttgcagGCTCCACA
fgenesh1_pm.C_scaffold_1002596 CTTCTGGCAG--------  —//—  ----------GCTCCACA

fgenesh2_kg.6__1464__AT5G14870.1 ACGCTTTCAGgtaaaaca  —//—  tatttttcagGTACTACT
AT5G14870.1                      ACGCTTTCAG--------  —//—  ----------GTACTACT

scaffold_403565.1              GACACAGCTGgcagggcc  —//—  gccctgacagCTGGTATC
AT2G46110.1                    GACACAGCTG--------  —//—  -------------GTATC
```

Fig. 3. Microsimilarity, indicated in red, between the splice sites of introns that appear to have been gained. The CDS are shown in upper cases and intronic sequences in lower cases.

**Table 3.** Positions of Lost Introns.

| | Median$^{a}$ | Median, <150 bp$^{a}$ | 5′$^{b}$ | 5′, <150 bp$^{b}$ | 3′$^{b}$ | 3′, <150 bp$^{b}$ |
|---|---|---|---|---|---|---|
| *A. thaliana–A. lyrata* orthologs | 0.44 ($P$ = 0.0396$^{c}$) | 0.42 ($P$ = 0.0249$^{c}$) | 46 ($P$ = 0.1488$^{d}$) | 46 ($P$ = 0.1384$^{d}$) | 37 ($P$ = 0.1593$^{c}$) | 34 ($P$ = 0.1468$^{c}$) |
| *A. thaliana–A. thaliana* ohnologs | 0.59 ($P$ = 0.0505$^{d}$) | 0.62 ($P$ = 0.0382$^{d}$) | 42 ($P$ = 0.0073$^{c}$) | 38 ($P$ = 0.0743$^{c}$) | 73 ($P$ = 0.0003$^{d}$) | 63 ($P$ = 0.0668$^{d}$) |

[a] Relative position of lost introns on the CDS.

[b] Number of lost of introns in the 5′ or 3′ in relation to the number of introns.

[c] Probability of obtaining the same or smaller number based on randomization.

[d] Probability of obtaining the same or larger number based on randomization.

homologous recombination (HR) with the intron-less reverse transcript (Mourier and Jeffares 2003; Sverdlov et al. 2004; Roy and Gilbert 2006) or by genomic deletion (Llopart et al. 2002). Although the reverse transcriptase (RT)-mediated model has been considered to be the most prominent mechanism of intron loss (Mourier and Jeffares 2003; Sverdlov et al. 2004; Roy and Gilbert 2005b), it was recently proposed that genomic deletions through NHEJ repair of DSBs might also be a common mechanism of intron loss (Farlow et al. 2011). We first compared the length of the introns in *A. lyrata* whose corresponding introns were lost in *A. thaliana*, with the length of introns in *A. lyrata* that were retained in both species. A strong bias was found toward short introns being more likely to be lost, with a median length of 86 bp ($P < 0.0001$, randomization of intron loss events; see Materials and Methods). The bias toward loss of short introns has also been observed in many other species such as mammals, pufferfish, *Drosophila*, and *Aspergillus* (Roy et al. 2003; Coulombe-Huntington and Majewski 2007a, 2007b; Loh et al. 2008; Zhang et al. 2010). This bias is predicted by the RT model, although loss by the NHEJ model should also be heavily biased toward short introns as deletions that occur during the repair of DSBs by NHEJ are usually very short (Lieber 2010). The RT model predicts a bias toward loss of adjacent introns because recombination with an RT product can sometimes result in multiple neighboring introns being deleted in a single event (Sharpton et al. 2008). We did not observe any clear case of multiple intron loss by a single event; only one adjacent intron loss was observed, and this gene has lost an additional nonadjacent intron within the short time period. Thus, this adjacent intron loss might represent two independent events. We then examined the exon sequences bordering the introns and found that introns flanked by AG and GT on their 5′ and 3′ splice sites, respectively, thus sharing an AG|GT motif (where | indicates the splice site), are more likely to be lost (28/90 vs. 6,844/54,941; $P < 0.0001$, randomization). It has been suggested that the pairing between such identical splice sites (e.g., AG|GT) during the repair of DSBs by the NHEJ pathway might result in the deletion of introns (fig. 2) (Kent and Zahler 2000; Farlow et al. 2011). We therefore investigated whether the lost introns are simply more likely to have identical motifs at their 5′ and 3′ splice sites, regardless of AG|GT. We took 8 bp of the 5′ and 3′ exon/intron border, both centering the splice site—thus typically NNNN|GTNN and NNAG|NNNN and checked each 4-bp combination that would result in the precise deletion of the intron (NNNN| and NNAG|, NNN|G and NAG|N, NN|GT and AG|NN, N|GTN and G|NNN, or |GTNN and |NNNN). Indeed, we found that the lost introns were more likely to contain identical 4-bp motifs in their 5′ and 3′ splice sites that would result in their precise deletion compared with the shared introns (46/90 vs. 11505/54941; $P < 0.0001$). Furthermore, we found that 75 of the lost introns contained identical 3-bp motifs that would result in the precise deletion of the intron, also a significant enrichment (75/90 vs. 25909/54941; $P < 0.0001$). These results suggest a prominent role of NHEJ in the loss of these introns. This pattern is not restricted to *A. thaliana* as 13 of 15 of the lost introns in *A. lyrata* were also flanked by such 3-bp identical motifs. We did not detect any significant differences in GO (gene ontology) category or median or peak expression values measured by MPSS (massively parallel signature sequencing) between *A. thaliana* genes with and without intron losses since the *A. thaliana*–*A. lyrata* divergence.

Another frequently observed pattern is the bias toward loss of 3′ introns. Although selection against loss of 5′ introns due to 5′ introns often containing more regulatory elements could also account for this pattern, the 3′ bias has usually been interpreted as due to the prominent role of RT in intron loss (Mourier and Jeffares 2003; Sverdlov et al. 2004). This is because RT products are generated from the 3′ end and are often truncated, and thus, the 3′ end of a gene would have a higher chance to lose introns if recombination with RT products is the major mechanism of intron deletion. We examined the position across the CDS of the introns that were lost in *A. thaliana* since its divergence from *A. lyrata* and since the most recent WGD using two different measures (table 3). First, we calculated the relative positions of the lost and shared introns across the CDS as (distance from 5′)/(total length of CDS). Second, we assigned each intron as being either 5′ or 3′ in relation to the number of introns (see Materials and Methods). We detected no 3′ bias in either measure in the introns lost in *A. thaliana* but retained in the *A. lyrata* orthologs. In fact, the median of the relative positions was shifted toward 5′ compared with the median of all the shared introns. However, the positions of intron losses in the ohnologs were shifted toward the 3′ end. Although the trends are weak, there appears to be a difference in the position of lost introns between *A. thaliana*–*A. lyrata* orthologs and ohnologs (see below for further discussion).

## Discussion

### Prevalence of Intron Loss over Gain

We utilized the whole-genome sequences of the two closely related species *A. thaliana* and *A. lyrata* to search for intron gains and losses that occurred since the divergence of the two species on a genome-wide scale. In addition, we were able to use the genomes of papaya, poplar, and grapevine as outgroups, in contrast to previous studies on *A. thaliana* that had to rely on rice as the closest outgroup (Knowles and McLysaght 2006; Roy and Penny 2007a), or studies on rice which had to rely on *A. thaliana* as outgroup (Lin et al. 2006). An emerging consensus from many recent studies is that the common ancestor of eukaryotic lineages such as plants, animals, and fungi was probably very intron rich and that intron loss has been much more prevalent than intron gain during the evolution of most eukaryotic lineages (Archibald et al. 2002; Rogozin et al. 2003; Roy and Gilbert 2005a; Roy and Penny 2006; Coulombe-Huntington and Majewski 2007a, 2007b; Stajich et al. 2007; Loh et al. 2008; Sharpton et al. 2008; Csuros et al. 2011). We also found many more losses than gains in both *Arabidopsis* species. Although in earlier studies the rate of intron gains in plants appeared to be similar to the rate of intron losses, and also higher than in many other lineages (Rogozin et al. 2003; Roy and Gilbert 2005b), our results are consistent with the previous studies on *A. thaliana* and rice, which detected much higher numbers of intron losses than gains in ohnologs of both species (Lin et al. 2006; Roy and Penny 2007a). However, we should also point to certain limitations of our study. First, we did not search for intron gain and loss that largely alter the structure of the gene (although the indels of a couple of nucleotides would have been tolerated). For instance, conversion of exonic sequences into intronic sequences might be one pathway to intron gains, but we did not look for such cases (Irimia et al. 2008). We would also not have been able to detect intron-sliding events of up to ~30 nucleotides that might have led to the emergence of unique intron positions (Krauss et al. 2008). In addition, we only looked at introns within the CDS of protein-coding genes because the prediction of untranslated regions or noncoding RNAs is less reliable, especially in *A. lyrata* as transcript data are limited. Finally, the detection of intron gain and loss depends on each intron position being present in outgroup sequences, which is not always the case. For more than half of the introns unique to one of the two species, we could not determine whether they represent gains or losses, mainly due to uncertainty in the alignment with outgroup sequences (table 1).

### Higher Rate of Intron Loss in *A. thaliana*

One striking observation was the six times more intron losses in *A. thaliana* than in *A. lyrata*. Although heterogeneity of intron loss rates has been noted on several occasions (Jeffares et al. 2006; Roy and Gilbert 2006; Coulombe-Huntington and Majewski 2007b; Loh et al. 2008; Farlow et al. 2010), the large difference in such a short period of time (~10 Ma) is surprising. We cannot think of any technical or methodological issue that would lead to the underestimation of the rate in *A. lyrata*. For instance, the intron loss rate since the most recent WGD in the ancestral *Arabidopsis* lineage prior to the divergence of the two species was similar when calculated separately using *A. thaliana* or *A. lyrata*, suggesting that there is no systematic bias hindering the detection of intron losses in *A. lyrata*. Also, the number of unique intron positions before confirming with outgroup sequences was also 5 to 6 times larger in *A. lyrata*, which rules out any differences between the two species in the step validating the unique intron positions with outgroup sequences.

What then might explain the large difference in intron loss rate between both species? One possibility is that the selection favoring intron loss is much stronger in *A. thaliana* than in *A. lyrata*. The genome of *A. thaliana* is 1.5–2 times smaller than that of *A. lyrata*. Comparison of the two genomes showed that *A. thaliana* introns and intergenic intervals are often shorter than their counterparts in *A. lyrata* (Hu et al. 2011). In addition, population analysis of indels in *A. thaliana* showed a much larger number of fixed and segregating deletions than insertions. The bias toward deletions was especially pronounced for indels of >10 bp. If this pattern was due to mutational bias favoring deletions and no selection was involved, deletion and insertion polymorphisms should have similar allele frequencies. However, deletions were segregating at much higher frequencies than insertions, with many of the deletions approaching fixation. These results indicate the presence of a strong genome-wide selection favoring deletions, which has likely contributed to the genome-size reduction of *A. thaliana* (Hu et al. 2011). Although the probability for a given intron loss mutation to reach fixation will depend on the functional importance of that intron and the strength of selection, if the selection favoring deletion is much stronger in *A. thaliana*, even for deletions as small as ~10 bp, it seems highly plausible that an intronless allele will have a much higher chance to reach fixation in *A. thaliana* than in *A. lyrata*, resulting in a much higher rate of intron loss.

Another possibility is that the frequency at which an intronless copy is generated is higher in *A. thaliana*, due to some difference in the efficiency of the intron removal mechanism, such as the number of reverse transcripts generated, or the activity of HR or NHEJ. For instance, the different intron loss rates in apicomplexan species such as *Plasmodium* have been attributed to the difference in retrotransposon activity as retrotransposons can generate RT products (Roy and Hartl 2006; Roy and Penny 2007b). If the rate of RT-mediated intron loss is the major determinant of different loss rates, species with higher retrotransposon activity should experience more intron losses. Contrary to this expectation, *A. thaliana* contains fewer retrotransposons than *A. lyrata*, and the retrotransposon activity appears to be higher in *A. lyrata* (Hu et al. 2011), rendering it unlikely that retrotransposon activity had any influence in the different intron loss rate. In addition, our results point to a major role of NHEJ rather than RT in intron loss. Thus, we find it highly unlikely that any mechanistic difference in the process of RT-mediated intron loss is responsible for the higher intron loss rate in *A. thaliana*. Considering the support for the contribution of NHEJ in gain and loss in both species, some mechanistic difference in the process of DSB repair via NHEJ could result in different loss rates. However, one would expect such a difference to affect the rates of gain and loss in a similar way and result in a positive correlation between the rates of gain and loss. Indeed, positive correlation between the rates of gain and loss has been observed in various lineages (Carmel et al. 2007b; Stajich et al. 2007) and has been interpreted as evidence for mutational process (e.g., HR or NHEJ) being the major determinant of the gain and loss rates (Roy and Hartl 2006; Farlow et al. 2011). We only detected two intron gains in *A. thaliana*, and thus, there is clearly no positive correlation between the gain and loss rates. If some difference in the process of DSB repair by NHEJ or any other mutational mechanism were to be the dominant factor behind the higher rate of intron loss in *A. thaliana*, such difference would have had to emerge within a relatively short period and would have to explain a higher loss rate but not a higher gain rate. Although such a scenario cannot be formally ruled out, our results are more consistent with stronger selection for genome reduction, which predicts a higher loss rate and lower gain rate (Lynch 2002; Lynch and Conery 2003). We therefore argue that selection for genome reduction has played a major role in the higher rate of intron loss in *A. thaliana*.

How then do our results fit with other studies suggesting that intron density is determined by mechanistic factors (Roy and Hartl 2006; Roy and Penny 2007b; Stajich et al. 2007; Farlow et al. 2011)? First, it is possible that mechanistic factors, such as the mutation rate creating novel introns or intronless alleles, are indeed driving the intron density in general and that selection influences the rates of intron gain and loss on rather rare occasions. Second, the role of selection might have been difficult to detect under the schemes of previous studies. For instance, the basis for mutation rates being a major determinant is the positive correlation observed between the rate of gain and loss. With the exception of the recent study on different *Drosophila* species (Farlow et al. 2010; Farlow et al. 2011), such positive correlation has often been inferred over a relatively large evolutionary timescale (Roy and Hartl 2006; Carmel et al. 2007b; Roy and Penny 2007b; Stajich et al. 2007), which may well contain alternating periods of high rates of intron gain and loss driven by the difference in selective pressure. Such roles of selection might be difficult to identify unless there are different species under different selective pressure, as was the case with *A. thaliana* and *A. lyrata* (Hu et al. 2011). Examining organisms that have small genomes but are intron rich, or that have large genomes but are intron poor, might provide further insight into the relationship between selection for reduced genomes and intron density.

Although the intron loss rate is much higher in *A. thaliana* than in *A. lyrata*, we cannot yet determine whether the intron loss rate has accelerated in *A. thaliana*, as opposed to having decreased in *A. lyrata*. In fact, the rate of intron loss we estimated for the time since the most recent WGD in the ancestral *Arabidopsis* lineage prior to the *A. thaliana*–*A. lyrata* divergence was similar to the loss rate in *A. thaliana* and higher than the rate in *A. lyrata* (table 1). This might point to a decrease of losses in *A. lyrata*, although there are a number of issues to be considered. First, the time from the WGD to the *A. thaliana*–*A. lyrata* divergence covers a longer period of time than the time since the *A. thaliana*–*A. lyrata* divergence, and it is conceivable that the rate has not been consistent throughout this period of time. Second, despite having undergone two WGDs in their ancestral lineage, both species including *A. lyrata* have small genomes, indicating that a considerable amount of genome reduction had also taken place prior to the divergence of the two species, although the factors (e.g., the role of selection) behind this process are unclear. Third, the rates are affected by difficulties in distinguishing ohnologs from the younger WGD and those from the older WGD (see Materials and Methods) and also by the uncertainties in the dates of the *A. thaliana*–*A. lyrata* divergence and the most recent WGD (Beilstein et al. 2010). Although it might be argued that the intron loss rate in the ohnologs could be different from the rest of the genome, we did not find ohnologs to be enriched in the intron losses in *A. thaliana* since the *A. thaliana*–*A. lyrata* divergence. Thus, despite the different rates between *A. thaliana* and *A. lyrata* being consistent with selection for genome reduction in *A. thaliana*, we note the importance for further analyses. Comparative analyses involving other Brassicaceae genomes should help us better understand the process of genome reduction and the evolution of intron density in Brassicaceae species. Also, population studies of *A. thaliana* and other species will be necessary to determine the role of selection in the process of genome reduction and how selection affects the rates of intron gain and loss.

## Role of DSB Repair via NHEJ in Intron Loss

Another noteworthy finding of our study is the strong tendency for recently lost introns to be flanked by short stretches of identical motifs. Such microsimilarity at deletion breakpoints is regarded as a signature of NHEJ, which is considered to be the most common nonhomologous recombination mechanism in generating genomic deletions, although other less common mechanisms cannot be ruled out (Hastings et al. 2009; Conrad et al. 2010). A number of recent studies have implicated NHEJ as being responsible for intron gain (Li et al. 2009; Curtis and Archibald 2010; Farlow et al. 2010, 2011; Zhang et al. 2010). Our results suggest that the NHEJ model should also be considered in future studies of intron loss where most studies to date have focused on the RT model. Another recent study examining the positional bias of intron gains and losses (5′ or 3′) in diverse eukaryotic species suggested that RT activity is not sufficient to explain the overall pattern of intron gain and loss (Cohen et al. 2011). It may therefore be worth reexamining past conclusions that have been made assuming that the RT model is the predominant mechanism for intron gain and loss. For instance, despite suggesting retrotransposon activity as the cause of the large increase in the number of intron losses in the ancestral lineage of *Plasmodium* and the ancestral lineage of *Theileria*, it was also noted that very few retrotransposons are present in current *Plasmodium* and *Theileria* species (Roy and Penny 2007b). If RT in general was not involved in intron loss, we need not invoke a hypothetical scenario of retrotransposons once being active in the ancestral lineage and subsequently being lost from both lineages.

We also wish to draw attention to two frequently observed patterns that are often interpreted as evidence for RT-mediated intron loss. First is the 3′ bias of intron loss. Interestingly, we detected a 3′ bias in the ohnologs, but not in the *A. thaliana*–*A. lyrata* orthologs. One possibility would be that the RT-mediated intron loss was more prominent prior to the *A. thaliana*–*A. lyrata* divergence or is more prominent in ohnologs than in other genes. We nevertheless note that we also found enrichment of short identical motifs flanking the lost introns in the *A. thaliana* ohnologs (45/123 vs. 1187/5115; $P = 0.0002$ for 4-bp motifs and 85/123 vs. 2480/5115; $P < 0.0001$ for 3-bp motifs), although the pattern is not as striking and caution is needed as the flanking splice sites of the corresponding retained introns are less likely to represent the ancestral state. Alternatively, differences in the strength of selection against the loss of 5′ introns could be responsible for the difference between the ohnologs and the orthologs, such as due to differences between orthologs and ohnologs, or selection for a compact genome outweighing the functional constraints on 5′ introns. A recent study of the microsporidian *Encephalitozoon cuniculi* genome, which contains very few introns, showed that introns in ribosomal protein-coding genes (RPGs) had a strong 5′ bias, whereas introns in the remaining genes did not. The authors suggested that selection for the retention of the 5′ introns of RPGs due to their regulatory function, and not mutational bias, was responsible for their 5′ bias (Lee et al. 2010). Although it is unlikely that all 5′ introns are functional, the role of selection in shaping the positional bias of intron loss might warrant closer inspection. Furthermore, we argue that the role of RT in generating the 3′ bias of intron loss requires more rigorous testing. Second, despite finding strong support for genomic deletion via NHEJ, we found no cases of inexact deletions that result in a number of nucleotides being added to or removed from the exon. This could be because identical motifs (e.g., AG|GT) would be more likely to occur at positions that would cause exact deletions and also due to selective constraints on the amino acid sequences. We were able to detect one intron gain that resulted in the addition of three nucleotides to the exon (fig. 3), suggesting that the lack of inexact deletions is not due to technical issues. Thus, interpreting the lack of inexact deletions as evidence for the RT model could be misleading.

## Conclusions

Here, we have identified a much larger number of intron losses in *A. thaliana* than in *A. lyrata* since their divergence and very few gains. Although the rate of mutations that remove introns are no doubt an important factor in determining the intron density, our results suggest that selection can also play a role in the evolutionary dynamics of intron gain and loss. To what extent existing introns are under functional constraint is a heavily debated topic (Lynch and Conery 2003; Roy and Gilbert 2006). Despite suggesting the role of selection in intron loss, we do not mean to say that all the lost introns were simply excessive DNA under no functional constraint. All we can say is that the strength of selection for genome reduction outweighed whatever selective constraint on the lost introns. Population analyses of many different species will be necessary to tackle such questions. Our results also indicate the key role of the repair of DSBs by NHEJ in intron loss, rather than the RT-mediated model that has been the assumption of many previous studies. Future studies considering the role of NHEJ in intron gain and loss should also help us better understand the various factors behind the intriguing uneven phylogenetic distribution of introns.

## Acknowledgments

We thank P. Pattyn and S. Proost for technical help and S. Roy and an anonymous reviewer for helpful comments. This research was supported in part by the European Commission through the 6th Framework Programme (contract number: FOOD-CT-2006-016214). Y.V.d.P. acknowledges support from the Belgian Federal Science Policy Office (IUAP P6/25) (BioMaGNet) and Ghent University (MRP N2N).

## References

Alioto TS. 2007. U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res.* 35:D110–D115.

Archibald JM, O'Kelly CJ, Doolittle WF. 2002. The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution. *Mol Biol Evol.* 19:422–431.

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A.* 107:18724–18728.

Carlton JM, Hirt RP, Silva JC, et al. (63 co-authors). 2007. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* 315:207–212.

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

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

Cenik C, Chua HN, Zhang H, Tarnawsky SP, Akef A, Derti A, Tasan M, Moore MJ, Palazzo AF, Roth FP. 2011. Genome analysis reveals interplay between 5'UTR introns and nuclear mRNA export for secretory and mitochondrial genes. *PLoS Genet.* 7:e1001366.

Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD. 2005. Human antisense genes have unusually short introns: evidence for selection for rapid transcription. *Trends Genet.* 21:203–207.

Cohen NE, Shen R, Carmel L. The role of reverse transcriptase in intron gain and loss mechanisms. *Mol Biol Evol.* 29:179–186.

Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, Lee C, Turner DJ, Hurles ME. 2010. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat Genet.* 42:385–391.

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.

Csuros M, Rogozin IB, Koonin EV. 2011. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol.* 7:e1002150.

Curtis BA, Archibald JM. 2010. A spliceosomal intron of mitochondrial DNA origin. *Curr Biol.* 20:R919–R920.

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

Farlow A, Meduri E, Schlötterer C. 2011. DNA double-strand break repair and the evolution of intron density. *Trends Genet.* 27: 1–6.

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci U S A.* 106:5737–5742.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet.* 10:551–564.

Hu TT, Pattyn P, Bakker EG, et al. (30 co-authors). 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet.* 43:476–481.

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, Bähler J. 2008. Rapidly regulated genes are intron poor. *Trends Genet.* 24:375–378.

Kent WJ, Zahler AM. 2000. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae-C. elegans* genomic alignment. *Genome Res.* 10:1115–1125.

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

Krauss V, Tḫmmler C, Georgi F, Lehmann J, Stadler PF, Eisenhardt C. 2008. Near intron positions are reliable phylogenetic markers: an application to holometabolous insects. *Mol Biol Evol.* 25:821–830.

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 U S A.* 104:19908–19913.

Le Hir H, Nott A, Moore MJ. 2003. How introns influence and enhance eukaryotic gene expression. *Trends Biochem Sci.* 28: 215–220.

Lee RCH, Gill EE, Roy SW, Fast NM. 2010. Constrained intron structures in a microsporidian. *Mol Biol Evol.* 27:1979–1982.

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

Li WH, Gu Z, Wang H, Nekrutenko A. 2001. Evolutionary analyses of the human genome. *Nature* 409:847–849.

Lieber MR. 2010. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu Rev Biochem.* 79:181–211.

Lin H, Zhu W, Silva JC, 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 U S A.* 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.

Lynch M. 2002. Intron evolution as a population-genetic process. *Proc Natl Acad Sci U S A.* 99:6118–6123.

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

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

Nott A, Le Hir H, Moore MJ. 2004. Splicing enhances translation in mammalian cells: an additional function of the exon junction complex. *Genes Dev.* 18:210–222.

Omilian AR, Scofield DG, Lynch M. 2008. Intron presence-absence polymorphisms in *Daphnia*. *Mol Biol Evol.* 25:2129–2139.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327:92–94.

Proost S, Pattyn P, Gerats T, Van de Peer Y. 2011. Journey through the past: 150 million years of plant genome evolution. *Plant J.* 66:58–65.

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K. 2009. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell* 21:3718–3731.

Remm M, Storm CE, Sonnhammer EL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol.* 314:1041–1052.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol.* 13:1512–1517.

Roy SW. 2004. The origin of recent introns: transposons? *Genome Biol.* 5:251.

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

Roy SW, Fedorov A, Gilbert W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci U S A.* 100:7158–7162.

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

Roy SW, Gilbert W. 2005b. The pattern of intron loss. *Proc Natl Acad Sci U S A.* 102:713–718.

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

Roy SW, Penny D. 2007a. 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, Penny D. 2007b. Widespread intron loss suggests retrotransposon activity in ancient apicomplexans. *Mol Biol Evol.* 24: 1926–1933.

Sharpton T, Neafsey D, Galagan J, Taylor J. 2008. Mechanisms of intron gain and loss in *Cryptococcus*. *Genome Biol.* 9:R24.

Siegel TN, Hekstra DR, Wang X, Dewell S, GAM Cross. 2010. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. *Nucleic Acids Res.* 38:4946–4957.

Simillion C, Janssens K, Sterck L, Van de Peer Y. 2008. i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* 24:127–128.

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

Sverdlov AV, Babenko VN, Rogozin IB, Koonin EV. 2004. Preferential loss and gain of introns in 3′ portions of genes suggests a reverse-transcription mechanism of intron insertion. *Gene* 338:85–91.

Swarbreck D, Wilks C, Lamesch P, et al. (16 co-authors). 2008. The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* 36:D1009–D1014.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. *Science* 320:486–488.

Wright SI, Lauga B, Charlesworth D. 2002. Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. *Mol Biol Evol.* 19:1407–1420.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 13:555–556.

Zhang L-Y, Yang Y-F, Niu D-K. 2010. Evaluation of models of the mechanisms underlying intron loss and gain in *Aspergillus* fungi. *J Mol Evol.* 71:364–373.