GENOME BIOLOGY AND EVOLUTION

SMBE

Association of Intron Loss with High Mutation Rate in *Arabidopsis*: Implications for Genome Size Evolution

Yu-Fei Yang, Tao Zhu, and Deng-Ke Niu*

MOE Key Laboratory for Biodiversity Science and Ecological Engineering, and Beijing Key Laboratory of Gene Resource and Molecular Development, College of Life Sciences, Beijing Normal University, China

*Corresponding author: E-mail: dkniu@bnu.edu.cn, dengkeniu@hotmail.com.

Accepted: March 16, 2013

Abstract

Despite the prevalence of intron losses during eukaryotic evolution, the selective forces acting on them have not been extensively explored. *Arabidopsis thaliana* lost half of its genome and experienced an elevated rate of intron loss after diverging from *A. lyrata*. The selective force for genome reduction was suggested to have driven the intron loss. However, the evolutionary mechanism of genome reduction is still a matter of debate. In this study, we found that intron-lost genes have high synonymous substitution rates. Assuming that differences in mutability among different introns are conserved among closely related species, we used the nucleotide substitution rate between orthologous introns in other species as the proxy of the mutation rate of *Arabidopsis* introns, either lost or extant. The lost introns were found to have higher mutation rates than extant introns. At the genome-wide level, *A. thaliana* has a higher mutation rate than *A. lyrata*, which correlates with the higher rate of intron loss and rapid genome reduction of *A. thaliana*. Our results indicate that selection to minimize mutational hazards might be the selective force for intron loss, and possibly also for genome reduction, in the evolution of *A. thaliana*. Small genome size and lower genome-wide intron density were widely reported to be correlated with phenotypic features, such as high metabolic rates and rapid growth. We argue that the mutational-hazard hypothesis is compatible with these correlations, by suggesting that selection for rapid growth might indirectly increase mutational hazards.

Key words: intron gain, genome reduction, mutational-hazard hypothesis, rapid growth.

Introduction

Spliceosomal introns are unevenly distributed among different organisms and among different genes of the same organism (Jeffares et al. 2006; Roy and Gilbert 2006; Rogozin et al. 2012). Vertebrates and plants have hundreds of thousands of introns in their genomes, whereas the tiny genome of the nucleomorph *Hemiselmis andersenii* has no introns at all (Lane et al. 2007). Among the genes of intron-rich organisms (such as humans), some genes are intronless, whereas some others each have more than 100 introns. Accumulating evidence indicates that the early eukaryotes were intron rich (Roy and Gilbert 2005a; Csuros et al. 2011). The differences in intron density can be explained mainly by the different rates of intron loss and partially by the different rates of intron gain (Roy 2006; Rogozin et al. 2012). In principle, the different rates of intron loss and gain might be caused either by mutational differences in removing old introns or generating new introns, or by differences in selective pressures for or against intron accumulation. However, only a few studies have attempted to explore the selective forces acting on intron gain and loss (Llopart et al. 2002; Lynch 2002; Lynch and Conery 2003; Lane et al. 2007; Stajich et al. 2007; Niu 2008).

Having introns does confer some benefits. Some introns can expand protein diversity through alternative splicing (Kalsotra and Cooper 2011). Some introns or elements contained in introns regulate gene expression (Le Hir et al. 2003; Wang et al. 2007; Rose et al. 2008; Parenteau et al. 2011; Rearick et al. 2011). The selective pressure on the loss or gain of these introns is obvious. However, there is no direct evidence of the proportions of introns that are actually involved in these functions. Surveys of nucleotides subject to purifying selection indicate that functional sequences do not exceed 12% of the human genome (Lindblad-Toh et al. 2011; Ponting and Hardison 2011). Recently, the majority of the human genome was found to have biochemical activity, which was considered as evidence against the existence of junk DNA in the large genome (Ecker et al. 2012; Pennisi 2012; ENCODE Project Consortium 2012). However, this conclusion has been criticized to be rather farfetched (Eddy 2012;

Niu and Jiang 2013). Other possible benefits of having introns that do not depend on specific sequences have been proposed (Forsdyke 1981; Fedorova and Fedorov 2005; Niu 2007; Niu and Yang 2011). They could, in principle, be applied to all spliceosomal introns. However, no convincing evidence implicates the sequence-independent benefits as active selective forces in intron evolution.

There is also a nonadaptive view of the evolution of genome complexity, including the accumulation of introns (Lynch 2002, 2007a, 2007b; Gray et al. 2010). Most introns are considered as nearly neutral but slightly deleterious. The absence of introns in prokaryotes and the scarcity of introns in some eukaryotes are attributed to efficient removal of introns by purifying selection. In vertebrates, smaller population sizes are thought to have relaxed the purifying selection against introns. An advanced version of the nonadaptive view of noncoding sequence evolution is the mutational-hazard hypothesis (Lynch 2006, 2007b; Lynch et al. 2006). More noncoding DNA is seen as more likely to accumulate deleterious mutations. The length of the DNA and the mutation rate determine the selective burden of carrying the surplus DNA. The evolutionary fate of the surplus DNA is determined by the efficiency of selection, which is mainly determined by effective population size. This hypothesis is more accessible and has attracted a great deal of attention; however, it is still a matter of debate (Lynch and Conery 2003; Whitney and Garland 2010; Whitney et al. 2010, 2011; Boussau et al. 2011; Lynch 2011; Kelkar and Ochman 2012).

In recent years, intron loss has been associated with genome reduction. The highly compacted genome of the nucleomorph *H. andersenii* has lost all its introns (Lane et al. 2007). The plant pathogen *Ustilago maydis* has a rather small genome compared with related, sequenced fungi. Comparative analysis revealed that massive intron loss had contributed to the genome reduction (Kamper et al. 2006). Within 10 Myr, *Arabidopsis thaliana* lost almost half of its genome (Hu et al. 2011; Proost et al. 2011). Consistent with this rapid genome reduction, a six-time higher rate of intron loss in *A. thaliana* compared with its relative *A. lyrata* was reported. Introns make a huge contribution to the genome size; therefore, it is quite likely that the selective force for genome reduction acts as a selective force against intron accumulation (Fawcett et al. 2012).

In both plants and animals, small genomes are associated with many phenotypes, including small nuclei, small cells, short cell cycles, high metabolic or photosynthetic rates, rapid growth, and short generation time (Gregory 2002, 2005; Cavalier-Smith 2005; Knight et al. 2005; Dufresne and Jeffery 2011). However, the molecular mechanism underlying the selective force for genome reduction remains unclear (Knight et al. 2005; Dufresne and Jeffery 2011; Lynch et al. 2011). Both selection for metabolic, temporal, and spatial economy and selection to minimize mutational hazards might have constrained genome sizes. In this study, we found that intron loss and genome reduction of *A. thaliana* are significantly associated with high mutation rates, which is consistent with the mutational-hazard hypothesis. Furthermore, we suggest that the mutational-hazard hypothesis might underlie the reported correlations between genome size and phenotypes.

Materials and Methods

Genomes and Transcriptomes

We downloaded the genome sequences and annotation files of *A. thaliana* from the *Arabidopsis* Information Resource (TAIR10 release, http://www.arabidopsis.org/, last accessed May 9, 2011), *Thellungiella parvula* (synonym: *Eutrema parvulum*) from Thellungiella.org (version 2.0, http://Thellungiella.org/, last accessed July 27, 2012), and *Brassica rapa* from the *Brassica* Database (version 1.2, http://brassicadb.org/brad/, last accessed July 27, 2012). The genomes and annotations for the following organisms were obtained from Phytozome (version 7.0, http://www.phytozome.net/, last accessed May 12, 2011): *A. lyrata* (JGI release v1.0), *Carica papaya* (ASGPB release of 2007), *Glycine max* (JGI Glyma1.0 annotation of the chromosome-based Glyma1 assembly), *Medicago truncatula* (Release Mt3.0 from the Medicago Genome Sequence Consortium), *Populus trichocarpa* (JGI assembly release v2.0, annotation v2.2), and *Vitis vinifera* (March 2010 12X assembly and annotation from Genoscope).

Following the recommendations of a previous study (Jeffares et al. 2008), we downloaded the "Stress Series" data related to the AtGenExpress Project from the microarray database of the Nottingham *Arabidopsis* Stock Center (http://arabidopsis.info/, last accessed October 9, 2011). The data set includes genome-wide expression data of *A. thaliana* under cold, osmotic, salt, drought, genotoxic, oxidative, UV-B, wounding, and heat stress conditions, as well as control conditions. The median value of all control plant data points of a gene was used to represent the gene expression level. Genes that could quickly change their rate of transcription were assumed to have a low time cost. The speed of gene expression regulation was defined as the maximum rate of transcriptional change per unit of time (Jeffares et al. 2008).

Identification of Orthologous Genes

Initially, genes were filtrated out for obvious annotation errors, such as coding sequences that were not a multiple of three nucleotides. For the genes with alternatively spliced isoforms, the longest mRNA was used for analysis. Using the best reciprocal Basic Local Alignment Search Tool (BLAST) hits, the homologous proteins between *A. thaliana* and *A. lyrata* were detected with thresholds of E values $< 10^{-10}$. A total of 21,158 groups of homologous proteins were obtained. Among these, 2,187 homologous groups consisting only of intronless genes were excluded, leaving 18,971 intron-containing homologous groups.

Recent genome analysis revealed several rounds of genome duplication and subsequent massive loss of genes in the evolution of *Arabidopsis* (Proost et al. 2011). In cases where asymmetric losses of paralogs have occurred between *A. thaliana* and *A. lyrata*, the best hits of reciprocal best BLAST represent homoeologs, rather than true orthologs. For this reason, the above intron-containing homologous groups were filtered by SynMap (Lyons et al. 2008). We have run our own analysis in SynMap with its recommended or default settings. For two settings that lack recommendations, we used BLASTN as the BLAST algorithm and Quota align algorithm to enforce a 1:1 syntenic depth between genomes. Ultimately, 16,266 homologous groups were found in conserved synteny blocks and regarded as orthologs.

Identification of Intron Loss and Gain

First, the orthologous proteins were aligned using ClustalW (version 2.1) (Larkin et al. 2007). With the aligned protein segments as markers, the full-length DNA sequences of the orthologous genes were aligned using ClustalW (version 2.1). These two steps were repeated using MUSCLE (version 3.8.31) (Edgar 2004). All conflicting results were checked manually. Before determining intron loss or gain, some alignments were manually improved.

Referring to previous studies (Roy and Penny 2006a, 2007; Zhang et al. 2010), the alignments were filtered. An intron position was discarded if 1) it is too close to the gene ends, with less than 45 nucleotides flanking either side, or 2) its flanking exon sequence has an identity lower than 0.68. The identity was calculated by counting 45 bp either side of an intron position. The value 0.68 represents the first quintile of the identities of all the aligned *A. thaliana* and *A. lyrata* orthologous mRNAs. In total, 2,534 positions that differ in the presence/absence of introns were observed between *A. thaliana* and *A. lyrata*. Seven species (*B. rapa*, *C. papaya*, *G. max*, *P. trichocarpa*, *T. parvula*, *M. truncatula*, and *V. vinifera*) were then used as outgroups to distinguish intron losses from intron gains (fig. 1). In *Arabidopsis*, two previous studies gave conflicting results on the relative frequency of intron loss and gain (Knowles and McLysaght 2006; Fawcett et al. 2012). However, in other eukaryotic lineages, most studies consistently indicated that intron loss greatly outnumbers intron gain (Roy and Gilbert 2005b, 2006b, 2007; Coulombe-Huntington and Majewski 2007a, 2007b; Stajich et al. 2007; Csuros et al. 2011). In this study, a more conservative criterion was used to define intron gain than to define intron loss. Dollo parsimony was used to define intron gains, whereas standard parsimony was used to define intron losses. The possible underestimation of the intron gain rate would affect both *A. thaliana* and *A. lyrata* simultaneously; therefore, it would not affect comparisons between *A. thaliana* and *A. lyrata*. In this way, 132 putative intron losses and no intron gains were detected in *A. thaliana*, and 35 putative intron losses and 55 putative intron gains in *A. lyrata*.

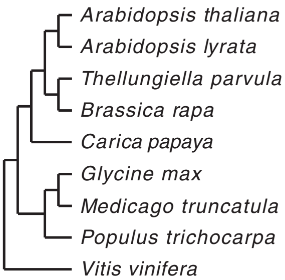

Fig. 1.—Phylogenetic tree used to distinguish intron loss and gain in *Arabidopsis thaliana* and *A. lyrata*. The tree was constructed using the phylogenetic tree from Phytozome (v8.0, http://www.phytozome.net/, last accessed January 20, 2012) and is not scaled according to phylogenetic distances. Dollo parsimony was used to define intron gains. That is, an intron gain in *Arabidopsis* was categorized when there were no introns in the position of the orthologous genes of any outgroup species. Meanwhile, there should be at least two outgroup branches that definitely showed absence of the intron. Standard parsimony was used to define intron losses. That is, an intron should be present more often than it is absent in the outgroup branches to define an intron loss. In cases where two or three species of the same outgroup branch differ in the absence/presence of an intron, the branch was not referred to as defining intron loss. Full lists of the presence, absence, and uncertainty of introns in the orthologous genes of the nine species are available in supplementary tables S5 and S6, Supplementary Material online.

In a recent study, Fawcett et al. (2012) found 90 intron losses and two intron gains in *A. thaliana* and 15 intron losses and nine intron gains in *A. lyrata*. The nonoverlapping results were manually checked and filtered by SynMap (Lyons et al. 2008). Two intron losses for *A. thaliana* and one intron loss and one intron gain for *A. lyrata* were integrated from their results. We also examined the intron losses and gains observed by Knowles and McLysaght (2006). No convincing nonoverlapping results were observed. Therefore, 134 putative intron losses in *A. thaliana* and 36 putative intron losses and 56 putative intron gains in *A. lyrata* remained.

A simple insertion in exonic sequence might be misannotated as an intron and consequently misidentified as an intron gain. Similarly, if an exonic segment was misannotated as an intron in orthologous genes, a simple deletion of the segment would result in misidentification of an intron loss. Therefore, we verified the annotations of the introns related to the intron losses and gains using transcriptome data. The expressed sequence tag (EST) data of *A. thaliana*, *A. lyrata* and the outgroup species were retrieved from dbEST (Expressed Sequence Tags database, http://www.ncbi.nlm.nih.gov/dbEST/, last accessed November 13, 2012), and the RNA-Seq reads data of *A. thaliana* (ERP001616) (Manavella et al. 2012), *A. lyrata* (SRP004429) (Hollister et al. 2011), and *B. rapa* (ERR037339) (Harper et al. 2012) were downloaded from the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra, last accessed November 19, 2012). These ESTs and reads were mapped to the genome using BLAT (BLAST-like Alignment Tool) (Kent 2002) and TopHat 2.0.5 (Trapnell et al. 2009). For an intron loss to be confirmed, we must first confirm that the lost sequence is an intron. So the extant orthologous introns must be actively transcribed and spliced correctly in at least one intron-extant species. By consulting the transcriptome data, 14 putative intron losses in *A. thaliana* were discarded due to lack of transcriptional information, and 6 putative intron losses also in *A. thaliana* were discarded because the introns were not spliced out in other species. In addition, we checked whether the intron-lost (IL) genes are still active after losing the introns by examining the transcripts they produced. We could not find any evidence of transcription for only one IL gene in *A. lyrata*. Furthermore, transcriptome data covered 27 of the 56 putative intron gains in *A. lyrata*. And among these 27 cases, 20 introns were found to be retained in transcripts and thus discarded from our data set. For the remaining seven intron gains, the active transcription of their orthologous genes in *A. thaliana* was also confirmed by transcriptome data. In total, 114 intron losses from *A. thaliana*, 35 intron losses from *A. lyrata,* and 7 intron gains in *A. lyrata* were supported by transcriptome data (table 1).

Calculating Nucleotide Substitution Rates

A previous study showed that, except for the two ends, most intron sequences in plants are not constrained (Guo et al. 2007). Nucleotide substitution rates of internal regions of introns could approximately reflect mutation rates. Based on this study, 10 nucleotides were removed from both the 5′- and 3′-ends of each intron before the intronic substitution rates were calculated.

Alignment artifacts could increase the calculated divergence, especially in alignments of noncoding sequences. To avoid these potential errors, Gblocks (Castresana 2000) was used to detect and filter unreliable alignment regions. However, too-strict filtration of alignments by Gblocks would delete regions with high frequencies of mutations, which in turn would make the calculated substitution rate lower than the actual mutational rates. When comparing different genes, the differences in mutation rate would become artificially smaller, even becoming statistically insignificant in some cases. Thus, a compromise was made. The intron sequence alignments were filtered using Gblocks (version 0.91b) (Castresana 2000). Specifically for noncoding sequence alignments, the minimum length of a block was adjusted to 5. The maximum number of contiguous nonconserved positions (*b*) was tested using 1, 2, 3, 4, and 8. The detected unreliable alignment regions were discarded. To seek an appropriate *b* value for the program Gblocks, we compared the nucleotide substitution rate of introns with well-aligned control sequences. Coding sequences are much easier to align because of the conservation of amino acid sequences. Initially, the orthologous coding sequences were aligned and filtered by Gblocks with its default parameters. Then, all the third sites of 4-fold degenerate codons were extracted with their relative positions in the alignments. These extracted synonymous bases were used as control sequences. To reduce the effects of random noise, only alignments longer than 30 bp were retained for calculating the substitution rate. The nucleotide substitution rate of intron sequences (d_i) and the control sequences (d_c) was estimated using the algorithmic method of Tamura and Nei (1993), implemented in PAUP 4.0 beta. In rice, the synonymous base substitution rate is half that of transposable elements (Gaut et al. 1996; Ma and Bennetzen 2004). The nucleotide substitution rates in transposable elements are often used as a neutral standard to represent the mutation rates (Gaffney and Keightley 2006). Therefore, if the d_i calculated in this study is more than twice the d_c, it is likely to have been overestimated because of unreliable alignments. By scrutinizing the ratios of d_i/d_c under different values of *b* (supplementary table S1, Supplementary Material online), $b = 2$ was chosen for Gblocks. When $b = 2$, d_i/d_c is < 0.8 when comparing *A. thaliana*, *A. lyrata*, *B. rapa*, and *T. parvula*. All the analyses of d_i were repeated using $b = 4$ for Gblocks and by PAML 4.2b (Yang 2007). Similar results were obtained (data not shown).

The synonymous substitution rates (d_S) of coding sequences were calculated by PAML 4.2b (Yang 2007). Gblocks with its default parameters was also used to filter the alignments of orthologous coding sequences.

Table 1
Arabidopsis thaliana Has Undergone More Intron Losses and Fewer Intron Gains than *A. lyrata*

	A. thaliana	*A. lyrata*	Pearson χ^2 Test
Lost introns	114	35	$P = 2 \times 10^{-6}$
Gained introns	0	7	
Conserved introns	80,262	80,262	

Detection of Regulatory Elements in Introns

CpG islands were detected by the NEWCPGREPORT program with its recommended settings (http://emboss.open-bio.org/wiki/Appdoc:Newcpgreport, last accessed August 17, 2012). The total number and total length of CpG islands present in each intron were counted. The density of CpG islands within an intron was defined as the number of CpG islands and the length of CpG islands divided by intron length.

The IMEter algorithm, with its default parameters, was used to predict the ability of an intron to enhance gene expression (Rose et al. 2008).

Results

Arabidopsis thaliana Lost More Introns and Gained Fewer Introns Than *A. lyrata*

Similar to Fawcett et al. (2012), the sites that differ in terms of the presence and absence of introns between *A. thaliana* and

A. lyrata were detected by comparing their orthologous genes. In most cases, intron loss and gain could not be distinguished because of the absence of reference genes in the outgroup species. Therefore, more outgroup species were added to form a larger data set of intron loss and gain that helped the statistical analysis. Using the orthologous genes from seven outgroup species (fig. 1), 132 losses and no gains in *A. thaliana* and 35 losses and 55 gains in *A. lyrata* were obtained from 2,534 loss/gain events. After integrating some cases from the data set of Fawcett et al. (2012) and filtering out uncertain cases, a larger data set, comprising 114 intron losses in *A. thaliana* and 35 intron losses and seven intron gains in *A. lyrata* (table 1), was obtained. With more and more closely related genomes being sequenced, which could be used as outgroup species, we expect that more intron losses and gains will be identified in *A. thaliana* and *A. lyrata*, which might make the following results more significant.

Arabidopsis thaliana lost more introns but gained fewer introns than *A. lyrata*. Fawcett et al. (2012) used the selective force for genome reduction of *A. thaliana* to explain these differences. However, the evolutionary mechanisms underlying the genome reduction are in dispute (Knight et al. 2005; Lynch et al. 2011). Both selection for metabolic, temporal, and spatial economy and selection to minimize mutational hazard might have promoted genome reduction (Cavalier-Smith 2005; Knight et al. 2005; Lynch et al. 2011). Using the data sets in *Arabidopsis*, we tested the mutational-hazard hypothesis of intron loss and deduced its influence on the mechanism of genome reduction. It should be noted that the seven gains in *A. lyrata* is too small a sample for statistical analysis.

Synonymous Sites of IL Genes Have Higher Mutation Rates

Within a genome, the mutation rate varies greatly across genes (Gaut et al. 2011). According to the mutational-hazard hypothesis, genes in mutational hot spots have higher hazards and thus experience stronger selective forces to purge surplus sequences. If mutational hazard was the selective force for intron loss, genes with higher mutation rates would be more likely to lose their introns. Using the synonymous substitution rate between *A. thaliana* and *A. lyrata* (d_{Stl}) as a proxy for mutation rate, we found that IL genes have significantly higher mutation rates than other genes ($P = 0.002$, fig. 2A). In this article, only genes that have definitely not lost or gained any introns in *Arabidopsis* were used as control genes and were conveniently described as "other genes."

Exon sequences flanking introns often contain splicing signals, so their synonymous sites are under selective constraints (Parmley and Hurst 2007; Parmley et al. 2007; Warnecke et al. 2008). Loss of introns would eliminate such constraints and in consequence increase the synonymous substitution rate (d_S).

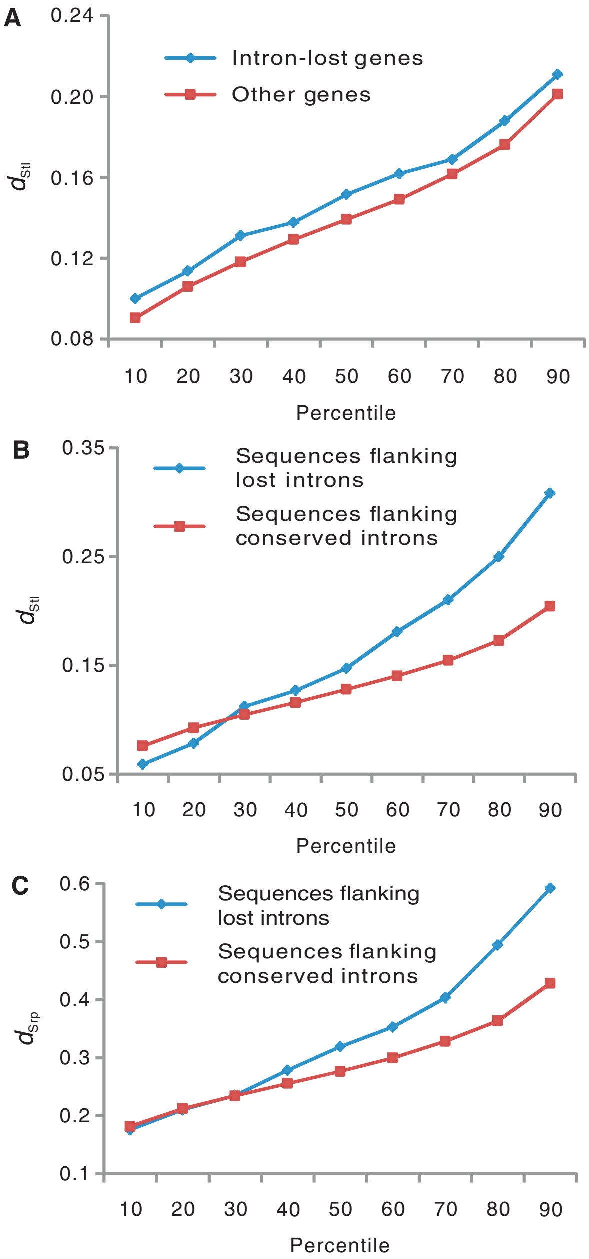

Fig. 2.—*Arabidopsis* IL genes have higher synonymous substitution rates. The 10th to 90th percentiles of the data are presented. (*A*) The IL genes have significantly higher d_{Stl} compared with other genes ($n = 143$ and 4,706, respectively; Mann–Whitney *U* test, $P = 0.002$). (*B*) The coding sequences flanking lost introns have significantly higher d_{Stl} than those flanking conserved introns ($n = 149$ and 14,126, respectively; Mann–Whitney *U* test, $P = 3 \times 10^{-4}$). (*C*) The coding sequences flanking the IL position also have higher d_{Srp} than those flanking conserved introns ($n = 116$ and 12,589, respectively; Mann–Whitney *U* test, $P = 0.002$). Coding sequences within 100 bp of both the 5′ and 3′ sides of an intron were defined as sequences flanking the intron. Using 200 bp and 400 bp to define flanking sequences gave similar results (data not shown). d_{Stl}, the synonymous substitution rate between *Arabidopsis thaliana* and *A. lyrata*. $d_{Srp,}$ the synonymous substitution rate between *Brassica rapa* and *Thellungiella parvula*.

The observed higher d_{Stl} of IL genes may either reflect intrinsic higher mutation rates or result from abandoned splicing signals. If abandon of splicing signals is the main cause, we could expect that coding sequences flanking lost introns in IL species would have elevated d_S, but their orthologous sequences of intron-retained species would not be affected. In contrast, if a higher mutation rate is an intrinsic feature of IL genes, both coding sequences flanking lost introns in IL species and their orthologous sequences of intron-retained species should have a higher d_S. As shown in figure 2*B*, we found that the coding sequences flanking lost introns have significantly higher d_{Stl} than those flanking conserved introns ($P = 3 \times 10^{-4}$). Similarly in *B. rapa* and *T. parvula*, the coding sequences flanking the *Arabidopsis* IL position also have higher d_S ($P = 0.002$, fig. 2*C*). Therefore, the higher d_S of *Arabidopsis* IL genes reflect higher intrinsic mutation rates.

Lost Introns Have Higher Mutation Rates

According to the mutational-hazard hypothesis, introns with higher mutation rates are more likely to be eliminated. Without the exact sequences of the lost introns, it is impossible to directly compare the substitution rate between lost introns and conserved introns. With the assumption that the differences in mutability among different introns are conserved in closely related species, we could examine whether the orthologous introns of the lost introns have higher d_i values than those of conserved introns.

We first tested the assumption of mutability conservation by analyzing conserved introns. In 43,894 groups of orthologous introns, the substitution rates between *A. thaliana* and *A. lyrata* (d_{itl}) and between *B. rapa* and *T. parvula* (d_{irp}) were calculated. Spearman's correlation analysis showed that d_{itl} and d_{irp} are significantly correlated (rho = 0.110, $P = 4 \times 10^{-117}$).

Then, we used the d_{irp} to represent the intron mutation rates of *Arabidopsis*. As shown in figure 3, the lost introns have significantly higher mutation rates than the conserved introns of either the same genes or other genes.

Many introns contain regulatory elements (Rose et al. 2008; Rearick et al. 2011), their nucleotide substitute rates are naturally expected to be lower than their mutation rates. Losses of these introns are selected against. The higher d_{irp} of lost introns may result from their paucity of regulatory elements. It is impossible to recognize all the possible regulatory elements, and absolutely confident results on mutation rates are hard to obtain. After surveying two common kinds of regulatory elements, we found it to be unlikely that they affect our conclusion that lost introns have higher mutation rates. CpG islands were found to be enriched in the first introns of rodents and likely to regulate gene expression (Chamary and Hurst 2004). However, we found that the densities of CpG islands are not correlated with d_{irp} in either *B. rapa* or *T. parvula* (supplementary table S2, Supplementary Material online). In plants, promoter-proximal introns contain dispersed signals to enhance gene expression, termed intron-mediated enhancement (IME) signals (Rose et al. 2008). If introns containing these signals are conserved, selection against their loss would result in the pattern that conserved introns have lower nucleotide substitution rates than lost introns. An IMEter score was designed to predict the ability of introns to enhance plant gene expression (Rose et al. 2008). Unexpectedly, we found that the IMEter score is positively correlated with d_{irp} (supplementary table S3, Supplementary Material online). It seems that conserved regulating motifs do not occupy a large percentage of the nucleotides in introns, and thus their existence does not reduce the overall substitution rate of introns. We accept that losses of introns containing regulatory elements are selected against, and only introns with fewer or no regulatory elements are free to be lost in evolution. However, the higher d_{irp} of lost introns we observed could not be explained by paucity of regulatory elements but could be explained by higher mutation rates.

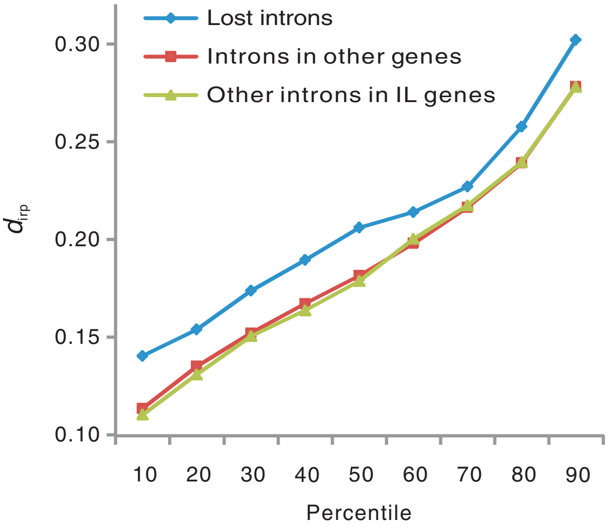

Fig. 3.—Lost introns of *Arabidopsis* have higher substitution rates. The 10th to 90th percentiles of the data are presented. The lost introns ($n = 80$) have significantly higher d_{irp} compared with conserved introns of the same genes ($n = 633$; Mann–Whitney *U* test, $P = 0.004$) and those of other genes ($n = 12{,}182$; Mann–Whitney *U* test, $P = 0.003$). d_{irp} is the nucleotide substitution rate between *Brassica rapa* introns and *Thellungiella parvula* introns.

Higher Mutation Rate Coincides with More Intron Loss Globally

Compared with *A. lyrata*, *A. thaliana* not only compacted its genome globally but also lost more introns and gained fewer introns (table 1). If the mutational-hazard hypothesis accounts for the selective force of intron and genome size evolution, *A. thaliana* should have a globally higher mutation rate than *A. lyrata*. A previous analysis of the internal transcribed spacer sequences of nuclear ribosomal DNA suggested a higher mutation rate in *A. thaliana* than in *A. lyrata* (Soria-Hernanz et al. 2008). In this study, we calculated the *A. thaliana* – *B. rapa* substitution rates (d_{itr} for intron sequences and d_{Str} for synonymous sites) and the *A. lyrata* – *B. rapa* substitution rates (d_{ilr} for intron sequences and d_{Slr} for synonymous sites). The differences between d_{itr} and d_{ilr} and between d_{Str} and d_{Slr} could reflect the differences of mutation rates between *A. thaliana* and *A. lyrata* because the same reference sequences were used. Wilcoxon signed ranks testing showed that *A. thaliana* has a significantly higher substitution rate than *A. lyrata* (table 2). Furthermore, we performed a relative rate test to confirm this difference using RRTree (Robinson-Rechavi and Huchon 2000). For d_S, the evidence for higher mutation rates in *A. thaliana* than in *A. lyrata* is obvious. In 6,917 of the 11,257 comparisons, *A. thaliana* has a higher d_S than *A. lyrata*. Among these 6,917 comparisons, 627 showed significant differences ($P < 0.05$). In contrast, *A. lyrata* has a higher d_S in 4,338 comparisons; among them, 197 are significant ($P < 0.05$). For d_i, *A. thaliana* also appears to have a higher mutation rate than *A. lyrata*. Consistent results were obtained in 33,083 of the 64,133 comparisons, with 1,974 comparisons being significant ($P < 0.05$). In contrast, inconsistent results were obtained less frequently, 27,854 of 64,133 comparisons with 1,269 comparisons being significant ($P < 0.05$).

Negative Correlation between Intron Number and Mutation Rate

If genes with higher mutation rates are more likely to lose introns, and intron loss dominates intron gain, one could expect that genes with higher mutation rates tend to have fewer introns. Using the nucleotide substitution rate between conserved introns as a proxy for mutation rate, we found that the mutation rate was negatively correlated with intron number per gene in both *A. thaliana* and *A. lyrata* (Spearman's rho = −0.170, $P = 2 \times 10^{-92}$, $n = 14{,}139$ in *A. thaliana* and Spearman's rho = −0.171, $P = 3 \times 10^{-93}$, $n = 14{,}139$ in *A. lyrata*). In addition, mutation rate and intron density (i.e., the number of introns per kilobase of mRNA) are also negatively correlated in these two species (Spearman's rho = −0.055, $P = 8 \times 10^{-11}$, $n = 14{,}105$ in *A. thaliana* and Spearman's rho = −0.027, $P = 0.001$, $n = 14{,}139$ in *A. lyrata*). These negative correlations are still significant when expression level and GC content were controlled (supplementary table S4, Supplementary Material online). Similarly, Yang and Gaut (2011) found that d_S is negatively correlated with intron number per gene in *Arabidopsis*.

Although the *P* values of these correlations are very small, we do not think that this is strong evidence for the mutational-hazard hypothesis. The total extant introns are the results of long-term evolution of intron gain and loss, whereas the measured mutation rates are recent evolutionary features. It is unclear whether the differences in mutation rates among different genes are conserved in long-term evolution. In contrast, the association of recent intron loss with recently high mutation rate is more likely to reflect intrinsic causal effects.

Table 2

Arabidopsis thaliana Has a Higher Global Mutation Rate than *A. lyrata*[a]

Percentile	d_{itr}	d_{ilr}	d_{Str}	d_{Slr}
10	0.174	0.170	0.317	0.305
20	0.211	0.205	0.353	0.341
30	0.239	0.233	0.381	0.371
40	0.266	0.260	0.410	0.397
50	**0.293**	**0.288**	**0.439**	**0.426**
60	0.325	0.319	0.471	0.456
70	0.360	0.356	0.511	0.495
80	0.408	0.404	0.571	0.551
90	0.481	0.478	0.672	0.654

NOTE.—d_{itr}, nucleotide substitution rate between *A. thaliana* introns and *B. rapa* introns; d_{ilr}, nucleotide substitution rate between *A. lyrata* introns and *Brassica rapa* introns; d_{Str}, synonymous substitution rate between *A. thaliana* genes and *B. rapa* genes; d_{Slr}, synonymous substitution rate between *A. lyrata* genes and *B. rapa* genes. The median values (i.e., the 50th percentiles) are highlighted in bold.

[a] Wilcoxon signed ranks test showed that d_{itr} is significantly higher than d_{ilr} (57,008 pairs of samples were compared, $P = 7 \times 10^{-69}$) and d_{Str} is significantly higher than d_{Slr} (13,208 pairs of samples were compared, $P = 5 \times 10^{-273}$).

Discussion

Spliceosomal introns are a common feature of eukaryotic genes. Their density varies greatly across genomes, as well as among genes of the same genome. However, the evolutionary mechanisms that control the gain and loss of introns are not clear. The model plant *A. thaliana* provides an opportunity to explore the problem. Within 10 Myr, *A. thaliana* has lost about half of its genome (Proost et al. 2011). It has also lost more introns and gained fewer introns than its close relative *A. lyrata* (Fawcett et al. 2012). The selective force that has driven this genome reduction has been proposed as a force favoring intron losses. Certain ideas can be borrowed directly from another well studied but still highly debated subject, the evolution of genome size. In this study, we mainly tested whether the mutational hazard hypothesis could be used to explain the pattern of intron loss. This hypothesis proposes that noncoding sequences have slightly deleterious effects on fitness because of the hazard of accumulating deleterious mutations (Lynch 2006, 2007b; Lynch et al. 2006). According to this hypothesis, selection to minimize the mutational hazard would preferentially remove surplus DNA from genomes and genes with high mutation rates. Consistently, we found that IL genes have higher mutation rates than other genes, and lost introns have higher mutation rates than conserved introns. Furthermore, we found that *A. thaliana* has a higher genome-wide mutation rate than *A. lyrata*.

According to the principle of population genetics, the efficiency of selection in removing slightly deleterious mutations depends heavily on the effective population size (N_e).

Organisms with a larger N_e are expected to purge noncoding sequences more effectively than those with a smaller N_e (Lynch and Conery 2003). Unfortunately, the difference in N_e between *A. thaliana* and *A. lyrata* is not monotonic. Historically, *A. thaliana* had an N_e about four times of *A. lyrata*; however, its N_e is now much smaller than that of *A. lyrata* (Lundemo et al. 2009; Falahati-Anbaran et al. 2011; Gomaa et al. 2011). The differences in the rates of intron loss and gain between *A. thaliana* and *A. lyrata* are consistent with the difference in historical N_e. If most introns are slightly deleterious, as assumed in the nonadaptive view, there should be a significant decline in the rate of intron loss and an expansion of genome size during the most recent evolution of *A. thaliana*. This prediction may be tested with the numerous natural accessions of *A. thaliana* currently being sequenced.

In addition to the deleterious effects of mutational hazards, metabolic, spatial, and temporal economy might also act as selective forces to remove surplus DNA (Cavalier-Smith 2005; Knight et al. 2005). According to selection for economy of gene expression, highly and quickly expressed genes can be expected to experience stronger selective forces for shorter introns. Rapidly expressed genes were found to be intron poor in organisms from yeasts to humans (Chen et al. 2005; Jeffares et al. 2008). However, the fact that highly expressed genes have short introns was only observed in animals (Castillo-Davis et al. 2002; Li et al. 2007; Carmel and Koonin 2009). In plants and yeasts, most studies revealed a positive correlation between intron size and expression level (Juneau et al. 2006; Ren et al. 2006; Li et al. 2007; Jeffares et al. 2008). Even in animals, the energetic cost of long introns seems to be too small to be efficiently selected against (Huang and Niu 2008). In spite of this, we tested whether selection for temporal and energetic economy of gene expression have driven intron losses in the evolution of *Arabidopsis*. If the time cost of transcription and splicing of introns had driven the intron losses, introns in genes that require rapid changes in their rate of expression could be expected to be preferentially lost. However, we did not find that IL genes were significantly more rapidly expressed than other genes in *A. thaliana* (Mann–Whitney *U* test, $P = 0.82$). The IL genes have significantly higher expression levels than other genes (Mann–Whitney *U* test, $P = 0.016$). However, highly expressed genes tend to have more introns in *A. thaliana* (Jeffares et al. 2008). The highly expressed genes may be more likely to lose introns just by chance. For each IL gene, we randomly selected a gene with the same number of introns from those that have not lost or gained any introns. Then we performed Wilcoxon signed-rank test to examine whether IL genes have higher level of gene expression. A total of 10,000 rounds of random samplings and pairwise tests were carried out. In most cases (8,798 rounds), IL genes did not have a significantly higher level of expression. Therefore, selection for economy of gene expression could not explain the rapid intron losses in *A. thaliana*. There is still no convenient way to explore the nuclear space constraint of introns and the selection for economy in DNA replication. Further studies are required to determine whether nuclear space constraint and selection for economy in DNA replication coexist with mutational hazards in driving intron loss and genome reduction.

Genome-wide intron density is positively correlated with the generation time of eukaryotes (Jeffares et al. 2006). That is, organisms with shorter life cycles tend to have fewer introns than slowly growing organisms. More notably, small genomes are correlated with many phenotypic features, such as small nuclei, small cells, short cell cycles, high metabolic/photosynthetic rates, small seed sizes, rapid growth, and short generation time (Cavalier-Smith 2005; Gregory 2005; Knight et al. 2005; Dufresne and Jeffery 2011). We suggest that the mutational-hazard hypothesis does not necessarily conflict with such correlations. Instead, the selection to minimize mutational hazards could be an alternative explanation for such correlations. Small organisms with rapid growth and reproduction generally have high metabolic rates (Glazier 2010; Price et al. 2010). A high rate of metabolism is associated with high rates of oxygen consumption and free radical generation, which in turn causes more DNA damage. In addition, rapid cell division results in higher accumulation of replication errors in genomes. Although *A. thaliana* and *A. lyrata* are closely related, *A. thaliana* is annual, whereas *A. lyrata* is perennial. *Arabidopsis thaliana* has a significantly higher mutation rate than *A. lyrata* (table 2). In both plants and animals, many previous studies have shown that organisms with short generations have higher substitution rates than organisms with longer generations (Gaut et al. 1996, 2011; Li et al. 1996; Nikolaev et al. 2007; Soria-Hernanz et al. 2008; Muller and Albach 2010). Thus, selection for rapid growth indirectly increased mutational hazards, which in turn might act as a selective force to remove surplus DNA.

In addition to selection to minimize mutational hazards, there is another possible, but less likely, explanation for the association between mutation rate and intron loss: mutation bias. Similar to the pattern of intron loss we observed in *Arabidopsis*, Magee et al. (2010) found that genes in the hypermutable regions of legume chloroplast genomes are preferentially lost and relocated to the nucleus. The authors did not invoke selective pressure but mutation bias. Could our observation of intron loss be explained simply by mutation bias? Under certain conditions, genes experience higher frequencies of both double-strand breaks (DSBs) and point mutations (Shee et al. 2011; Kim and Jinks-Robertson 2012). If DSB repair was the predominant mechanism of both intron loss and intron gain, as recently suggested (Li et al. 2009; Farlow et al. 2011; Ragg 2011; Fawcett et al. 2012), intron loss and gain would tend to occur simultaneously in hypermutated genes. This hypothesis and the mutational-hazard hypothesis both predict that introns with higher mutation rates would be preferentially lost. However, these two hypotheses differ entirely in their predictions of the rate of intron gain. The mutation bias hypothesis predicts that genes and genomes with higher mutation rates are more likely to gain introns, whereas the mutational-hazard hypothesis predicts that they are less likely to gain introns. Limited by the small number of intron gains observed, we were unable to assess the mutation rates of intron-gained genes with any statistical confidence. However, these two hypotheses can be distinguished by comparisons at the genome-wide level. *Arabidopsis thaliana* has a higher genome-wide mutation rate than *A. lyrata*; therefore, the mutation bias hypothesis predicts that *A. thaliana* would have gained more introns, whereas the mutational-hazard hypothesis predicts that *A. thaliana* is less likely to have gained introns. Fawcett et al. (2012) reported two intron gains in *A. thaliana* and six intron gains in *A. lyrata*. Using more outgroup genomes to distinguish intron loss and gain, we found seven cases of intron gain in *A. lyrata* but no cases of intron gain in *A. thaliana*. If putative gained introns that lack support from transcriptome data are also considered, the difference in the number of intron gains between *A. thaliana* and *A. lyrata* becomes much larger: none versus 56. In conclusion, the pattern of intron loss and gain observed in *A. thaliana* and *A. lyrata* is more consistent with the mutational-hazard hypothesis.

Supplementary Material

Supplementary tables S1–S6 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank the anonymous referees and associate editor Michael Lynch for their useful comments, Jeffrey A. Fawcett and Y. Van de Peer for sharing records of intron gains and losses detected in *A. thaliana* and *A. lyrata*, and Hans K. Stenøien and Mohsen Falahati-Anbaran for discussions on the N_e of *A. thaliana* and *A. lyrata*. This work was supported by the National Natural Science Foundation of China (grant numbers 31121003 and 31071112) and the Fundamental Research Funds for the Central Universities.

Literature Cited

Boussau B, Brown JM, Fujita MK. 2011. Nonadaptive evolution of mitochondrial genome size. Evolution 65:2706–2711.

Carmel L, Koonin EV. 2009. A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes. Genome Biol Evol. 1:382–390.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. Nat Genet. 31:415–418.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17: 540–552.

Cavalier-Smith T. 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann Bot. 95:147–175.

Chamary J-V, Hurst LD. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. Mol Biol Evol. 21: 1014–1023.

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.

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.

Dufresne F, Jeffery N. 2011. A guided tour of large genome size in animals: what we know and where we are heading. Chromosome Res. 19: 925–938.

Ecker JR, et al. 2012. Genomics: ENCODE explained. Nature 489:52–55.

Eddy SR. 2012. The C-value paradox, junk DNA and ENCODE. Curr Biol. 22:R898–R899.

Edgar R. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.

Falahati-Anbaran M, Lundemo S, Ågren J, Stenøien HK. 2011. Genetic consequences of seed banks in the perennial herb *Arabidopsis lyrata* subsp. *petraea* (Brassicaceae). Am J Bot. 98:1475–1485.

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

Fawcett JA, Rouzé P, Van de Peer Y. 2012. Higher intron loss rate in *Arabidopsis thaliana* than *A. lyrata* is consistent with stronger selection for a smaller genome. Mol Biol Evol. 29:849–859.

Fedorova L, Fedorov A. 2005. Puzzles of the human genome: why do we need our introns? Curr Genomics 6:589–595.

Forsdyke DR. 1981. Are introns in-series error-detecting sequences? J Theor Biol. 93:861–866.

Gaffney DJ, Keightley PD. 2006. Genomic selective constraints in murid noncoding DNA. PLoS Genet. 2:e204.

Gaut B, Yang L, Takuno S, Eguiarte LE. 2011. The patterns and causes of variation in plant nucleotide substitution rates. Annu Rev Ecol Evol Syst. 42:245–266.

Gaut BS, Morton BR, McCaig BC, Clegg MT. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. Proc Natl Acad Sci U S A. 93:10274–10279.

Glazier DS. 2010. A unifying explanation for diverse metabolic scaling in animals and plants. Biol Rev. 85:111–138.

Gomaa NH, Montesinos-Navarro A, Alonso-Blanco C, Pico FX. 2011. Temporal variation in genetic diversity and effective population size of Mediterranean and subalpine *Arabidopsis thaliana* populations. Mol Ecol. 20:3540–3554.

Gray MW, Lukes J, Archibald JM, Keeling PJ, Doolittle WF. 2010. Irremediable complexity? Science 330:920–921.

Gregory TR. 2002. A bird's-eye view of the C-value enigma: genome size, cell size, and metabolic rate in the class aves. Evolution 56:121–130.

Gregory TR. 2005. The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. Ann Bot. 95:133–146.

Guo XY, Wang Y, Keightley PD, Fan LJ. 2007. Patterns of selective constraints in noncoding DNA of rice. BMC Evol Biol. 7:208.

Harper AL, et al. 2012. Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. Nat Biotechnol. 30:798–802.

Hollister JD, et al. 2011. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proc Natl Acad Sci U S A. 108:2322–2327.

Hu TT, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 43:476–481.

Huang Y-F, Niu D-K. 2008. Evidence against the energetic cost hypothesis for the short introns in highly expressed genes. BMC Evol Biol. 8:154.

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.

Juneau K, Miranda M, Hillenmeyer ME, Nislow C, Davis RW. 2006. Introns regulate RNA and protein abundance in yeast. Genetics 174:511–518.

Kalsotra A, Cooper TA. 2011. Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet. 12:715–729.

Kamper J, et al. 2006. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. Nature 444:97–101.

Kelkar YD, Ochman H. 2012. Causes and consequences of genome expansion in fungi. Genome Biol Evol. 4:13–23.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12:656–664.

Kim N, Jinks-Robertson S. 2012. Transcription as a source of genome instability. Nat Rev Genet. 13:204–214.

Knight CA, Molinari NA, Petrov DA. 2005. The large genome constraint hypothesis: evolution, ecology and phenotype. Ann Bot. 95:177–190.

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, et al. 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.

Larkin MA, et al. 2007. Clustal W and clustal X version 2.0. Bioinformatics 23:2947–2948.

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

Li SW, Feng L, Niu DK. 2007. Selection for the miniaturization of highly expressed genes. Biochem Biophys Res Commun. 360:586–592.

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

Li WH, Ellsworth DL, Krushkal J, Chang BHJ, Hewett-Emmett D. 1996. Rates of nucleotide substitution in primates and rodents and the generation time effect hypothesis. Mol Phylogenet Evol. 5:182–187.

Lindblad-Toh K, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478:476–482.

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.

Lundemo S, Falahati-Anbaran M, Stenøien HK. 2009. Seed banks cause elevated generation times and effective population sizes of *Arabidopsis thaliana* in northern Europe. Mol Ecol. 18:2798–2811.

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

Lynch M. 2006. The origins of eukaryotic gene structure. Mol Biol Evol. 23:450–468.

Lynch M. 2007a. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A. 104:8597–8604.

Lynch M. 2007b. The origins of genome architecture. Sunderland (MA): Sinauer Associates, Inc.

Lynch M. 2011. Statistical inference on the mechanisms of genome evolution. PLoS Genet. 7:e1001389.

Lynch M, Bobay L-M, Catania F, Gout J-F, Rho M. 2011. The repatterning of eukaryotic genomes by random genetic drift. Annu Rev Genomics Hum Genet. 12:347–366.

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

Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. Science 311:1727–1730.

Lyons E, Pedersen B, Kane J, Freeling M. 2008. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. Tropical Plant Biol. 1:181–190.

Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci U S A. 101:12404–12410.

Magee AM, et al. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res. 20:1700–1710.

Manavella PA, et al. 2012. Fast-forward genetics identifies plant CPL phosphatases as regulators of miRNA processing factor HYL1. Cell 151:859–870.

Muller K, Albach DC. 2010. Evolutionary rates in *Veronica* L. (Plantaginaceae): disentangling the influence of life history and breeding system. J Mol Evol. 70:44–56.

Nikolaev SI, et al. 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. Proc Natl Acad Sci U S A. 104:20443–20448.

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

Niu DK. 2008. Exon definition as a potential negative force against intron losses in evolution. Biol Direct. 3:46.

Niu DK, Jiang L. 2013. Can ENCODE tell us how much junk DNA we carry in our genome? Biochem Biophys Res Commun. 430:1340–1343.

Niu DK, Yang YF. 2011. Why eukaryotic cells use introns to enhance gene expression: splicing reduces transcription-associated mutagenesis by inhibiting topoisomerase I cutting activity. Biol Direct. 6:24.

Parenteau J, et al. 2011. Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. Cell 147:320–331.

Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. Mol Biol Evol. 24:1600–1603.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. PLoS Biol. 5:e14.

Pennisi E. 2012. ENCODE project writes eulogy for junk DNA. Science 337:1159–1161.

Ponting CP, Hardison RC. 2011. What fraction of the human genome is functional? Genome Res. 21:1769–1776.

Price CA, Gilooly JF, Allen AP, Weitz JS, Niklas KJ. 2010. The metabolic theory of ecology: prospects and challenges for plant biology. New Phytol. 188:696–710.

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.

Ragg H. 2011. Intron creation and DNA repair. Cell Mol Life Sci. 68:235–242.

Rearick D, et al. 2011. Critical association of ncRNA with introns. Nucleic Acids Res. 39:2357–2366.

Ren X-Y, Vorst O, Fiers MWEJ, Stiekema WJ, Nap J-P. 2006. In plants, highly expressed genes are the least compact. Trends Genet. 22:528–532.

Robinson-Rechavi M, Huchon D. 2000. RRTree: relative-rate tests between groups of sequences on a phylogenetic tree. Bioinformatics 16:296–297.

Rogozin I, Carmel L, Csuros M, Koonin E. 2012. Origin and evolution of spliceosomal introns. Biol Direct. 7:11.

Rose AB, Elfersi T, Parra G, Korf I. 2008. Promoter-proximal introns in *Arabidopsis thaliana* are enriched in dispersed signals that elevate gene expression. Plant Cell 20:543–551.

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

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

Roy SW, Gilbert W. 2005b. Rates of intron loss and gain: implications for early eukaryotic evolution. Proc Natl Acad Sci U S A. 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, 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.

Shee C, Gibson JL, Darrow MC, Gonzalez C, Rosenberg SM. 2011. Impact of a stress-inducible switch to mutagenic repair of DNA breaks on mutation in *Escherichia coli*. Proc Natl Acad Sci U S A. 108: 13659–13664.

Soria-Hernanz DF, Fiz-Palacios O, Braverman JM, Hamilton MB. 2008. Reconsidering the generation time hypothesis based on nuclear ribosomal ITS sequence comparisons in annual and perennial angiosperms. BMC Evol Biol. 8:344.

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

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 10:512–526.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111.

Wang HF, Feng L, Niu DK. 2007. Relationship between mRNA stability and intron presence. Biochem Biophys Res Commun. 354: 203–208.

Warnecke T, Parmley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. Genome Biol. 9:R29.

Whitney KD, Boussau B, Baack EJ, Garland T Jr. 2011. Drift and genome complexity revisited. PLoS Genet. 7:e1002092.

Whitney KD, Garland T Jr. 2010. Did genetic drift drive increases in genome complexity? PLoS Genet. 6:e1001080.

Whitney KD, et al. 2010. A role for nonadaptive processes in plant genome size evolution? Evolution 64:2097–2109.

Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. Mol Biol Evol. 28: 2359–2369.

Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Zhang LY, Yang YF, Niu DK. 2010. Evaluation of models of the mechanisms underlying intron loss and gain in *Aspergillus* fungi. J Mol Evol. 71:364–373.

Associate editor: Michael Lynch