Evolution of Gene Structural Complexity: An Alternative-Splicing-Based Model Accounts for Intron-Containing Retrogenes[1][W]

Chengjun Zhang, Andrea R. Gschwend, Yidan Ouyang, and Manyuan Long*

Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637 (C.Z., A.R.G., Y.O., M.L.); and National Key Laboratory of Crop Genetic Improvement, National Center of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China (Y.O.)

The structure of eukaryotic genes evolves extensively by intron loss or gain. Previous studies have revealed two models for gene structure evolution through the loss of introns: RNA-based gene conversion, dubbed the Fink model and retroposition model. However, retrogenes that experienced both intron loss and intron-retaining events have been ignored; evolutionary processes responsible for the variation in complex exon-intron structure were unknown. We detected hundreds of retroduplication-derived genes in human (*Homo sapiens*), fly (*Drosophila melanogaster*), rice (*Oryza sativa*), and Arabidopsis (*Arabidopsis thaliana*) and categorized them either as duplicated genes that have all introns lost or as duplicated genes that have at least lost one and retained one intron compared with the parental copy (intron-retaining [IR] type). Our new model attributes intron retention alternative splicing to the generation of these IR-type gene pairs. We presented 25 parental genes that have an intron retention isoform and have retained introns in the same locations in the IR-type duplicate genes, which directly support our hypothesis. Our alternative-splicing-based model in conjunction with the retroposition and Fink models can explain the IR-type gene observed. We discovered a greater percentage of IR-type genes in plants than in animals, which may be due to the abundance of intron retention cases in plants. Given the prevalence of intron retention in plants, this new model gives a support that plant genomes have very complex gene structures.

Plant and animal genomes are more dynamic than previously thought. Genomes were assumed to hold a finite number of genes, but later it was discovered that new genes can arise through DNA-based duplication, RNA-based duplication, gene fusions, or de novo origination, resulting in a distinct new gene that evolves independently (Long et al., 2003; Shiao et al., 2007; Kaessmann et al., 2009). Gene nucleotide sequences are ever changing, with the natural introduction of various mutations, which can affect the evolutionary trajectory of the gene. Gene structures can also change and evolve over time; changing a gene's coding and noncoding structure can lead to the formation of new genes and neofunctionalization (Chen et al., 2013).

Posttranscriptional gene structure modification commonly occurs through a process called alternative splicing (AS). AS is a regulated process that results in a single gene coding for multiple gene products. There are several types of AS (Fig. 1): exon skipping, intron retention, alternative 3′ splice site, and alternative 5′ splice site selection, to name a few (Blencowe, 2006; Keren et al., 2010). Exon skipping occurs when an interior exon is spliced out of a transcript along with its flanking introns. Intron retention occurs when an intron remains in the mature RNA transcript. Alternative 3′ and 5′ splice site selection occurs when there are multiple splice sites in an exon and part of the exon is spliced out along with the adjacent intron. AS produces various proteins from a single gene and can be important for regulation and tissue-specific gene expression (Blencowe, 2006; Keren et al., 2010).

Changes in the exon-intron structure of a gene can also occur, including the loss and/or gain of introns. Intron loss (IL) has been known to be an important aspect of gene structural variation and plays a vital role in gene evolution (Roy and Gilbert, 2006). Two major models have been proposed to explain the mechanisms behind IL. First is the retroposition model (Brosius, 1991; Kaessmann et al., 2009), where the complementary DNA (cDNA) transcript of a gene is inserted back into the genome, resulting in a gene that lacks introns, has a poly A/T tail, and is flanked by short direct repeats (Fig. 2). Retroposition is widely accepted to create intron-lacking genes, inspiring a great number of studies across species. The three common retrogene signatures, mentioned above, allowed retrogenes to be easily identified in the genomes of *Drosophila* spp. (Betrán et al., 2002), primates (Marques et al., 2005; Vinckenbosch et al., 2006; Pan and Zhang, 2009), and other animals (Pan and Zhang,

[1] This work was supported by the U.S. National Science Foundation (grant no. MCB1026200).

* Address correspondence to mlong@uchicago.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Manyuan Long (mlong@uchicago.edu).

[W] The online version of this article contains Web-only data.
www.plantphysiol.org/cgi/doi/10.1104/pp.113.231696

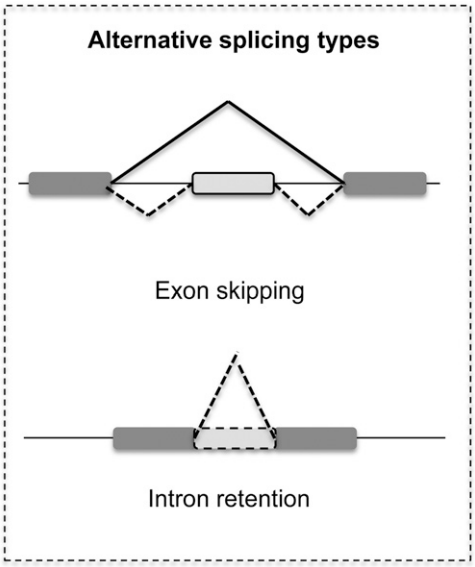

Figure 1. AS model. Two types of AS are depicted. Exon skipping involves splicing out an interior exon. Intron retention retains an intron as a part of the exon.

2009; Fu et al., 2010), as well as in plants, such as, Arabidopsis (*Arabidopsis thaliana*; Zhang et al., 2005), rice (*Oryza sativa*; Wang et al., 2006; Sakai et al., 2011), and *Populus trichocarpa*. (Zhu et al., 2009). In the *Drosophila* spp. genome, 24 recently inserted retrogenes were identified (Betrán et al., 2002). Marques et al. (2005) found 76 new retrogenes were fixed in the primate linage over the past 63 million years, which was subsequently updated to around 120 bona fide retrogenes through in silico analyses (Vinckenbosch et al., 2006). Pan and Zhang (2009) showed that the number of retrogenes range from 95 to 275 in eight mammals and four nonmammal species, and Fu et al. (2010) found 440 intact retrogenes in zebrafish. In plants, researchers have found hundreds (100–380) of retrogenes in rice (Wang et al., 2006; Sakai et al., 2011), 69 retrogenes in Arabidopsis (Zhang et al., 2005), and 108 retrogenes in *Populus* spp. (Zhu et al., 2009), providing further support for the importance of retroposition in genome evolution.

Another model of gene IL was proposed by Gerald R. Fink , which is known as the Fink model (Fink, 1987) or the RNA-based gene conversion theory, and illustrates IL by homologous recombination between a genomic copy of a gene and a cDNA transcript, which lacks introns (Fig. 2). This model was later experimentally demonstrated by Leslie K. Derr (Derr et al., 1991; Derr, 1998) in yeast (*Saccharomyces cerevisiae*). Because reverse transcriptase start at the 3′ end of RNA molecules and can detach prematurely, some cDNA transcripts can be truncated at the 5′ end, resulting in less homologous recombination with the incomplete 5′ end. This, therefore, reduces the IL frequency of the 5′ end, resulting in a 5′-favored distribution of introns (Fink, 1987). This process of gradient IL has been shown to contribute to the intron distribution in genes of the yeast genome (Goffeau et al., 1996), as well as the genes of other species (Mourier and Jeffares, 2003).

Both models for IL were experimentally tested and well supported by observation in yeast (Derr et al., 1991). To detect RNA-mediated recombination, Derr et al. (1991) utilized a yeast strain containing a *HIS3* gene that could be expressed in both the sense and antisense direction. The *HIS3* gene was interrupted by an artificial intron, which could be spliced out in the antisense orientation compared with the *HIS3* promoter. If the gene was expressed from the sense strand, the intron sequences were not removed and the *HIS3* gene was nonfunctional, but if it was driven by the *GAL1*

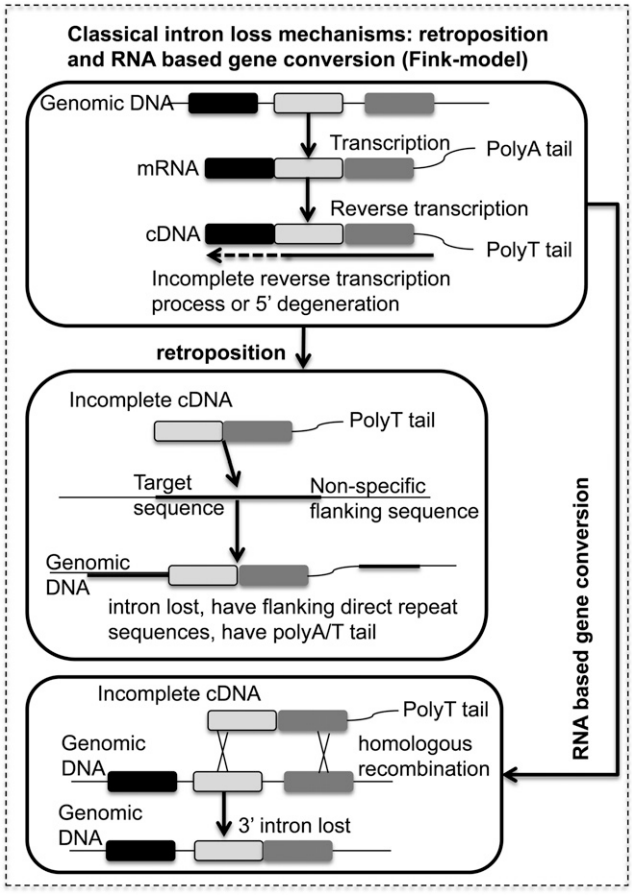

Figure 2. Retroposition and RNA-based gene conversion (Fink model). The top box depicts a gene with three exons (black, light gray, and gray) going through transcription and incomplete reverse transcription, resulting in a partially reverse-transcribed gene (missing the black exon). The second section shows how the incomplete cDNA can become integrated back into the genome through retroposition, adding an additional, intronless partial copy of the gene to the genome. (The process or retroposition can occur with complete cDNA as well). The third section illustrates the partial cDNA engaging in RNA-based gene conversion with its progenitor. Crossing over between the intronless cDNA and the genomic copy leads to the loss of an intron in the genomic copy.

promoter, which led to an antisense RNA transcript, the intron was spliced out and the spliced transcript (*HIS3*$^-$) was the antisense form of the functional *HIS3* (*HIS3*$^+$) gene. If the *HIS3*$^-$ transcript was retroposed back into the chromosome or if it recombined with the plasmid, then the yeast would have a functional *HIS3*$^+$ gene and could survive on the His-lacking media. They discovered that there was an almost 50/50 chance for the yeast plasmid or chromosome to have the intronless, and therefore functional, copy. These data suggest RNA-based gene conversion and retroposition can be responsible for IL.

Although both IL models were reported decades ago, they do not completely explain all cases of IL observed across species. For example, the Fink model was extended to test the intron positions in the genomes of 18 eukaryotic species, but only a few species fit the model by having intron distributions that favored the 5′ end (Mourier and Jeffares, 2003); most of the eukaryotic species, including human (*Homo sapiens*), rice, and Arabidopsis, however, did not fit the model (Mourier and Jeffares, 2003; Roy and Gilbert, 2005). The retroposition model also allows cases of IL to slip through the cracks; genome-wide retrogene scans tend to apply very strict thresholds to identify retroposition events, which bias the analyses to only detect retrogenes that are completely lacking introns compared with the parental gene. These stringent parameters ignore possible genes that have lost the majority of their introns but still have retained one or more introns compared with the paralogous gene copy. Genes that display this structural anomaly are excluded from retrogene analyses, and the structural variations between the parent and daughter gene cannot always be explained by the Fink model.

In this study, we identify duplicated genes with uncommon gene structures in the genomes of human, *Drosophila melanogaster*, rice, and Arabidopsis and propose a new theory of gene structure evolution to complement the Fink model and retroposition model. We discovered duplicated gene pairs that have experienced IL but still retain one or more of the parental introns. Interestingly, we found a higher percentage of these genes in the plant genomes than in the human and *Drosophila* spp. genomes. We attribute these interesting intron-retaining (IR)-type gene structures to the retroposition or gene conversion of an intron retention AS isoform, offering a new model that more effectively explains complex exon-intron structures of eukaryotic organisms.

RESULTS AND DISCUSSION

Identifying Duplicated Genes with Complex Gene Structures

Using a custom pipeline, we aim to identify the duplicated genes present in four diverse species and address gene structure conundrums by applying a novel model of intron retention. Past studies have utilized various pipelines to identify duplicated genes in *Drosophila* spp. (Betrán et al., 2002), human (Marques et al., 2005; Vinckenbosch et al., 2006; Pan and Zhang, 2009), rice (Wang et al., 2006; Sakai et al., 2011), and Arabidopsis (Zhang et al., 2005). Because we aim to elucidate the mechanisms that underlie intron differences between a duplicated gene and its progenitor, we designed a specific pipeline that varies from these previous works (see "Materials and Methods"), which is depicted in Figure 3, to target intron sites and identify the duplicated gene copies that have lost introns compared with their parental genes in the genomes of four species, *Drosophila* spp., human, rice, and Arabidopsis. Briefly, the annotated genes from the aforementioned four genomes were downloaded and formatted. Because we are interested in the intron structures of the duplicate genes and their parental genes, we combined the first and last 25 bp surrounding an intron site for all applicable genes (genes with introns) and aligned those compiled 50-bp sequences to the respective whole genome sequences to identify candidate sequences for IL events (Fig. 3, top right). The sequences that aligned to the 50-bp exon compilations were extracted, as were the surrounding sequences (the length of the parental gene on either side of the aligned region), and then were aligned back to the candidate parental genes using TFASTY (version 35; Pearson, 2000). The intron sites of the parental and daughter genes were scrutinized and compared and then divided into two categories: IR events and IL events. In more detail, if the relative additional sequences at the intron site of the hit sequences was larger than 30 bp, we defined it as an IR event, if it was 1 to 30 bp, we defined it as an intron indel (this type was ambiguous and was ignored in the following analysis), and if it was 0 bp (no sequence at the intron site), we defined it as an IL event (Fig. 3, bottom right). If a duplicated gene is void of introns, these genes are examples of classical retrocopies. Interestingly, duplicate genes that contained at least one IL event and one IR event were identified in the four genomes. These genes are of great interest and evoke questions about how these structures arise in duplicated copies. We explore these unusual structural differences in the next sections.

Starting from 27,538 *Drosophila* spp., 33,855 human, 66,338 rice, and 27,416 Arabidopsis total gene isoforms, our search resulted in 957, 23,417, 10,268, and 9,342 candidate duplicated copies with at least one IL event, respectively (Fig. 3, left). After filtering out ambiguous matches, further investigation revealed 78 *Drosophila* spp., 451 human, 220 rice, and 200 Arabidopsis classical retrocopies, which are completely void of introns. Because we are proposing that some retrocopies can still contain a parental intron, which will be discussed more in the next section, we will refer to the duplicated genes that have experienced an IL event resulting in the loss of all the parental introns in the daughter gene as IL types. The duplicated genes that have experienced at least one IL event and have retained at least one parental intron will be referred to

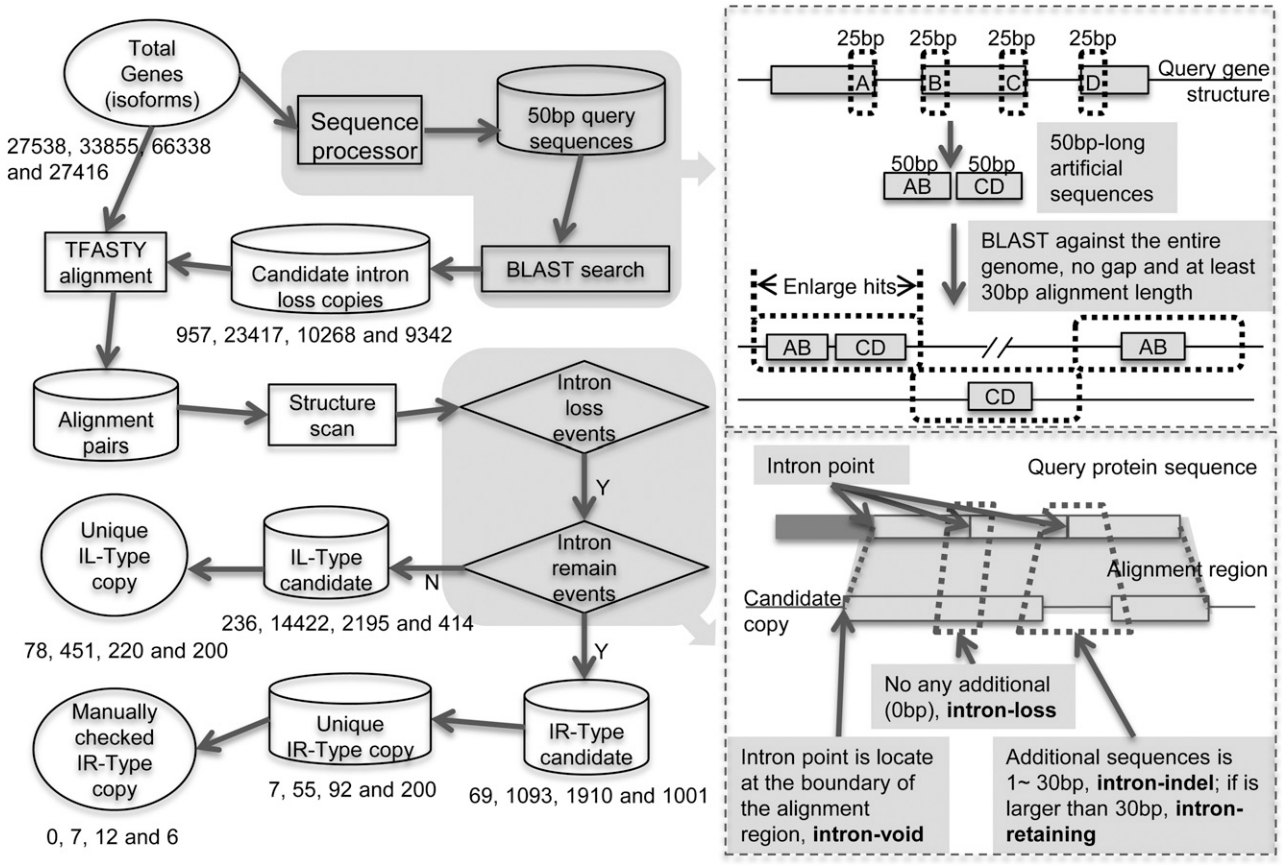

Figure 3. Flowcharts for detecting IR-type pairs. On the left is a flowchart of the pipeline carried out to detect the IL- and IR-type copies used for this study. Starting with the total isoforms for *Drosophila* spp., human, rice, and Arabidopsis (totals follow this order below the steps above), we first extracted the 25 bp of exon sequence from either side of the introns for the entire applicable gene and compiled them into 50-bp artificial sequences (detailed in the top right box). The 50-bp artificial sequences were aligned to their relative genome sequences. Candidate IL sequences were extracted (matched 30 bp/50 bp with no gaps in the sequence) and chained together with nearby IL events to form candidate IL copies. Using TFASTY, the candidate IL copies were aligned back to the genomic DNA sequences of the genes, and a comparison of the sequence structures was performed. Following the intron base pair thresholds described in the bottom right box, the IL copies were classified as either IL- or IR-type.

as IR types (Fig. 3, left). Of the duplicated genes that have experienced at least one IL event, seven *Drosophila* spp., 55 human, 92 rice, and 200 Arabidopsis genes retained at least one intron from the parental gene.

Hypothesis for IR-Type Duplicated Genes

When a duplicated gene has both lost an intron and retained an intron compared with the parental gene structure, two explanations come to mind: either a retrocopy has gained an intron or a DNA-duplicated copy has lost an intron. There are a few ways a retrocopy can gain an intron (Roy and Gilbert, 2006). After a retroposition event, nucleotide sequences, such as transposable elements, can be inserted into the intronless copy and can become an intron (Iwamoto et al., 1998; Lin et al., 2006). In most cases, if a transposable element inserts itself into an exon, the coding sequence will become interrupted, and therefore the gene will not be functional. The introns that do arise through this method are easily identified by their transposable element sequence. Another way a retrocopy can gain an intron is through the introduction of a new splicing signal, which will initialize sequences already present in the retroposed copy to become an intron (Wang et al., 2004). It is also possible for an intron to be reverse spliced back into a RNA transcript and reverse transcribed into cDNA and then undergo recombination with the genomic gene copy, resulting in the presence of the new intron in the genomic DNA (Roy and Irimia, 2009). These three methods of post-duplication intron gain are indicated by structural differences between the parental gene and the duplicated gene at the intron site and in the surrounding exon sequenc.

A duplicated gene can also lose introns; after a DNA duplication event, the parental gene's structure is still present in the daughter copy, but introns can be lost through RNA-based gene conversion (Fink, 1987; Roy

and Gilbert, 2005; Fig. 2), where the duplicated gene copy loses introns through homologous recombination with its cDNA intronless product. The resulting gene would be intronless. In addition, cases of IL due to genomic deletion have also been reported (Llopart et al., 2002; Roy and Gilbert, 2006).

Thus, duplicated gene copies can both gain and lose introns compared with the parental gene structure; however, these methods fall short of explaining all the cases of structural differences we observed. Lin et al. (2006) found that intron gain, as opposed to the IL, is uncommon; they identified five intron gain events versus 49 IL events in rice. Intron gain as an explanation for the structural differences observed in this study is unlikely when we consider the methods we used to identify IL events. The exon sequence surrounding the parental- and duplicate-copy intron site were conserved. The probability that a transposable element would insert into a parental intron site in the retrocopy is extremely low (Sverdlov et al., 2005), and a simple BLAST (Camacho et al., 2009) alignment revealed that the retained intron sequences in the duplicate gene copies were not transposable element sequences.

IL is a more likely explanation for these cases of structural differences, but, again, the methods with which we are familiar do not explain all the cases we observed. According to the RNA-based gene conversion theory, the duplicate gene would be completely intronless or the intron distribution for these genes would favor the 5′ end, but this prediction is not consistent with the intron distributions we observed, which we discuss further in the "Retained Intron Distribution" section.

Here, we propose an AS-based model, which introduces an AS aspect to the structural differences between duplicated genes and their progenitors. The AS-based model proposes that the most recent common ancestor (MRCA) of a gene has two or more isoforms, and one of the isoforms, which has been alternatively spliced to retain an intron as part of the coding sequence, has been retroposed. Therefore, when we compare the duplicated gene's structure to the parental gene, we find that most of the introns have been lost, yet it appears that one intron has been retained, when in actuality, that retained "intron" sequence may be part of the coding sequence of the retroposed isoform. The ancestral isoform from which the duplicate copy was retroposed can either be retained or lost. If the AS isoform is retained, then both isoform structures will be seen as parental coding transcripts. If the ancestral isoform of the retroposed gene is lost, these cases of AS-based retroposition are more difficult to identify, because the parental gene will have only one isoform, appearing as though the retained sequence should be an intron.

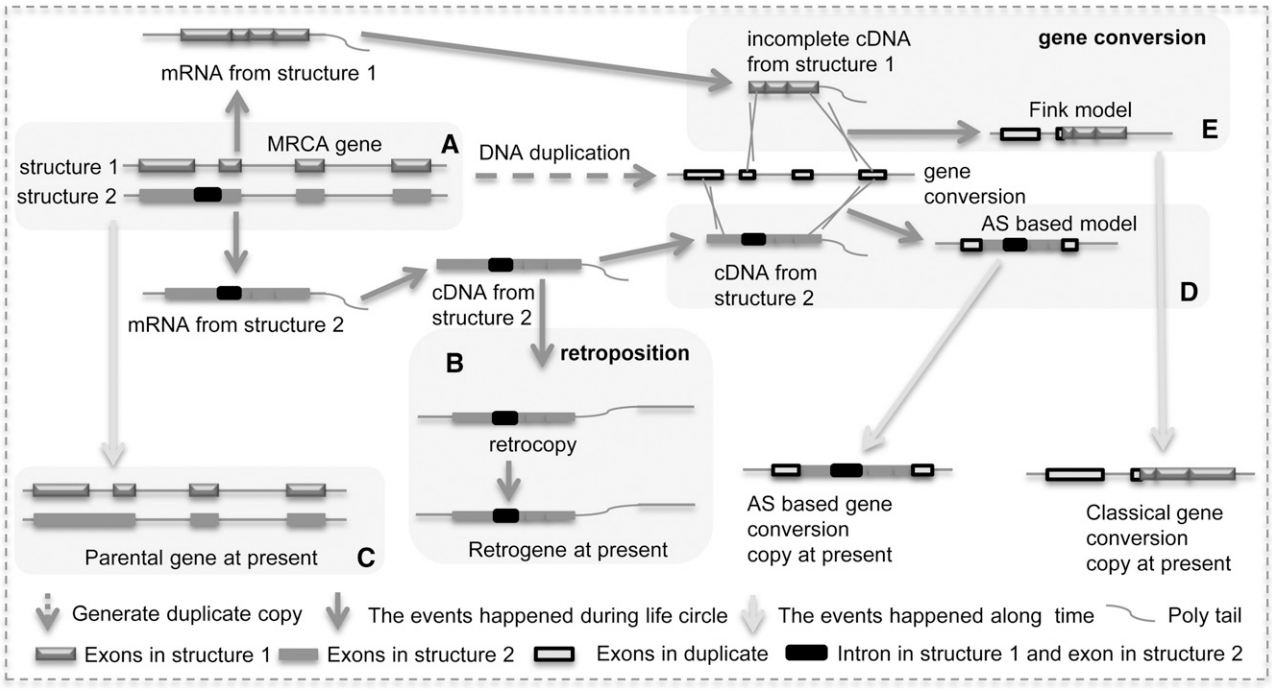

Figure 4. AS-based gene conversion and retroposition. A, Assuming MRCA gene has two isoform structures, in structure 2, the black box denotes the sequence is part of the exon, whereas that same sequence is part of the intron in structure 1. B, Retroposition operated on the AS structure 2 cDNA, leading to an intron-retained retrogene. C, The parental gene at present has two isoforms. D, AS-based gene conversion occurs between the intron retention cDNA and the genomic DNA, leading to an intron-retained novel chimeric DNA copy. E, Gene conversion between the incomplete cDNA from structure 1 and the genomic DNA, leading to an intron-retained novel chimeric DNA copy.

Figure 4 depicts an example of the AS-based model, where the MRCA of the parental and daughter genes has two isoforms, 1 and 2 (Fig. 4A). The first intron of structure 1 is retained as part of exon 1 in structure 2, due to intron retention AS (black box). If structure 2 undergoes RNA-based duplication (Fig. 4B), then the retrocopy will appear to have retained the parental gene's intron, when, in fact, it is a functional part of the exon. On the other hand, if structure 1 is retroposed, then the retrocopy will seem to have lost part of the first exon sequences compared with the parental gene (not pictured). In some cases, both parental gene isoforms may still exist in the transcriptome (Fig. 4C), but there may be cases where one of the MRCA's isoforms is lost, leading to confusion about the gene structure of the parental and new gene.

The AS-based model and the Fink model can be applied together to explain complex cases. For example, a genic region may have been segmentally DNA duplicated, yet the gene of interest may have lost one or more introns compared with the parental gene. The genes surrounding the gene of interest would be syntenic with the parental region that was duplicated, but the gene-of-interest's structure would vary from the parent gene's structure. Using the example from Figure 4, the parental gene has two isoforms, and RNA-based gene conversion may have occurred between the intron retention alternatively spliced isoform and the DNA-duplicated gene, which would appear as though all introns, except for one (Fig. 4D), were removed. This model complements the classical Fink model (Fig. 4E), where the cDNA of the gene-of-interest's intron retention isoform may not have been completely transcribed and is truncated on the 5′ end. Gene conversion may occur between this isoform's incomplete cDNA and the DNA-duplicated gene, making it appear that there have been multiple IL events in different parts of the gene compared with the parent gene structure. Though the AS-based model and Fink model can work together to explain gene structure differences, the introns described by each are different. The AS-based model assumes the "intron" that remains is alternatively spliced to be an intron in one isoform and part of an exon in another. The Fink model views the intron as a survived noncoding genetic element from the gene conversion process.

The AS-based model can be applied in conjunction with the other methods of intron evolution mentioned above, e.g. the introduction of new splicing signals (Wang et al., 2004) or transposon insertion (Roy and Gilbert, 2006), to illustrate how the great diversity of complex exon-intron structures can be generated.

AS Parental Isoforms Support AS-Based Model

As proof of concept, we sought to identify the IR-type duplicated genes with parental genes that had AS isoforms that correspond to the retained intron location. Using the parental genes of the IR-type duplicates that contain one or more of the parental introns (seven

Table I. *Twenty-five IR-type pairs have the same location of intron retention in parental gene and IR in the child copy*

Parental Gene Identifications	Child Copy Identifications or Location	Intron Location[b]	Has Synteny Region	Has Poly A/T
NM_001037738.2	chr15:73454226,73455272	3 (3)		
NM_001037738.2	chr2:198244540,198245617	3 (3)		
NM_000972.2	chr12:39860247,39861148	7 (7)		Y
NM_000972.2	chr15:59699162,59700354	5 (5)		Y
NM_001144012.2	chr9:37885640,37886388	2 (2)		
NM_001145426.1	chr16:31579706,31580819	5 (5)		
NM_004127.4	chr6:90595840,90597575	9 (9)		Y
LOC_Os01g61080.1	LOC_Os05g39720.1	4 (4)		Y
LOC_Os02g14440.1	LOC_Os06g48010.1[a]	1, 2 (2)		Y
LOC_Os03g02920.1	LOC_Os01g73220.1	1, 2 (2)	Y	
LOC_Os03g02920.1	LOC_Os06g48010.1[a]	1, 2 (2)		Y
LOC_Os04g43680.1	LOC_Os02g41510.1	1, 2 (2)	Y	
LOC_Os05g37470.1	LOC_Os10g05690.1	1 (1)		Y
LOC_Os02g44630.1	LOC_Os02g57720.1	2, 3 (3)		Y
LOC_Os02g44630.1	LOC_Os04g47220.1	2, 3 (2)	Y	
LOC_Os02g44630.1	LOC_Os07g26630.1	2, 3 (2, 3)		Y
LOC_Os08g36320.1	LOC_Os03g13300.1	4, 5 (4, 5)		
LOC_Os10g40730.1	LOC_Os03g01270.1/2	2, 3 (2, 3)	Y	
LOC_Os03g53860.4	LOC_Os02g03870.1	7 (6, 7)		Y
AT1G18020.1	AT1G76680.2	2, 3, 4 (2, 4)		
AT2G34560.2	AT5G52882.1	9 (9)		Y
AT2G42590.3	AT1G35160.2	3, 5, 6 (3, 4)		Y
AT2G42590.3	AT4G09000.2	3, 5, 6 (3, 4)		Y
AT3G26300.1	AT5G57260.1	2 (2)		
AT4G30270.1	AT5G57560.1	1 (1)		Y

[a]Two rice homologs share the same percentage similarity as the duplicated copy, so we cannot distinguish which is the real parental gene. [b]Outside bracketed numbers are intron retention location in the parental gene. Inside bracketed numbers are IR location in the child copy.

in *Drosophila* spp., 55 in human, 92 in rice, and 200 in Arabidopsis), we searched the AS databases (DEDB [Lee et al., 2004] for *Drosophila* spp., ASPicDB [Martelli et al., 2011] for human, and ASIP [Wang and Brendel, 2006] for rice and Arabidopsis), which use RNA sequencing, EST, and cDNA support to identify AS isoforms, following the steps detailed in "Materials and Methods." We identified three, 86, 47, and 32 parental genes with intron retention AS isoforms for *Drosophila* spp., human, rice, and Arabidopsis, respectively. We then manually checked the alternatively spliced intron location of the parental gene against the intron retention site in the child-duplicated copy and finally found zero, seven, 12, and six cases in which the intron affected by AS in the parental gene corresponds to the same location that retained the parental intron in the duplicated daughter gene for *Drosophila* spp., human, rice, and Arabidopsis, respectively (Table I). All 25 gene pair alignment structures are shown in Figure 5, A to F, and continue in Supplemental Figure S1.

An additional line of evidence to support our hypothesis would be to search for orthologs of the ancient gene isoforms retained in closely related outgroup species. We attempted this analysis, but due to the limited availability of data for related species, the results of the analysis were incomplete, and therefore no concrete conclusions could be drawn. We are also hesitant to confidently rely on this line of evidence. Severing et al. (2009) reported that the location and type of AS events did not persist in orthologous genes between Arabidopsis and rice or maize (*Zea mays*) and rice. When a more complete set of resources is available, this would be an interesting question to pursue.

Retained Intron Distribution

To understand the potential mechanisms responsible for the origination of these IR-type copies, we investigated the distribution of the location of retained introns for each gene to see if it supports our AS-based model. The Fink model predicts there will be an abundance of retained introns at the 5′ end of genes (Fink, 1987; Derr et al., 1991; Roy and Gilbert, 2005), due to gene conversion with incomplete cDNA transcripts,

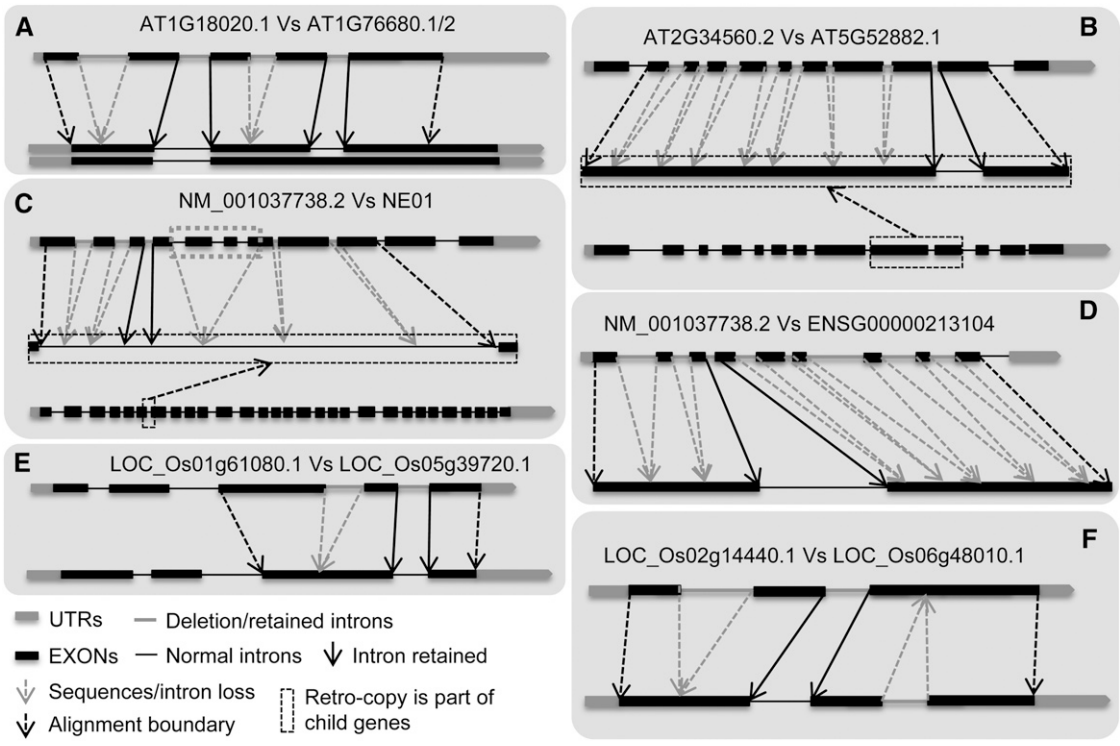

Figure 5. Gene structures of examples of IR-type pairs. Six examples of IR-type gene structures. Parental genes are on top, and the retrocopies are below. A and B are examples from Arabidopsis, C and D are examples from human, and E and F are examples from rice. A depicts a case where IL occurs on both sides of a retained intron in the retrocopy, and the retro-copy has two isoforms. B and C are both examples of the retroposed sequence only being a part of the child gene; the black dashed box shows the relative region of the retroposed sequence in the child gene. C shows an example of a retroposed sequence that lost multiple introns, experienced a deletion of both exons and introns (the sequences in gray dashed box), and makes up an intron of the child gene *neogenin1*. F shows an example in which the retrocopy has both retained and lost and intron compared with the parent gene, as well as gained a new intron. These retained introns provide support for the AS-based model, because the parental genes have retained intron retention in the same genic position as the retrocopy. The remaining 19 gene structures are presented in Supplemental Figure S1.

whereas the AS-based model predicts a more even distribution of retained introns, because AS can occur throughout a gene. The distribution of retained introns for the 25 genes with AS support at the intron retention site (Fig. 6, top), as well as the other 143 copies with support of AS somewhere in the transcript (Fig. 6, bottom), were manually checked (Supplemental Tables S1–S3). The locations of the retained introns were calculated as the length of the coding sequence upstream of the intron divided by the total length of the coding sequence (Mourier and Jeffares, 2003). In general, both data sets suggest that the duplicate-copy intron retention sites are randomly distributed along the gene structures. The random distribution of retained introns supports our AS-based hypothesis.

Furthermore, a combination of the AS-based model and the Fink model can work together to explain the cases that have experienced intron retention events on both sides of an intron that was lost (Fig. 5, A, C, and D; Supplemental Fig. S1, F–I). The Fink model alone is unable to explain these situations. For example, an incomplete cDNA transcript of an intron retention AS isoform may go through gene conversion with a DNA duplicated gene, resulting in a gene with two intron retention events on either side of an IL event.

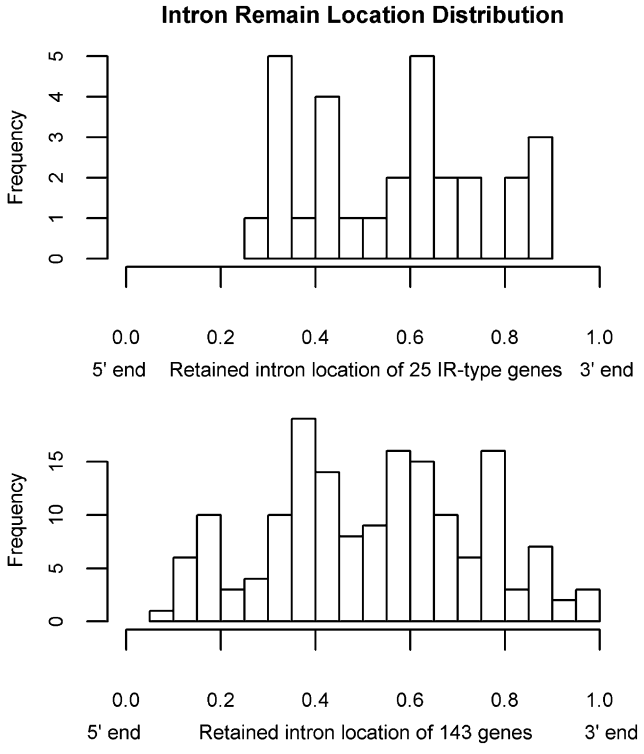

Figure 6. Retained intron location distributions. The locations of the retained introns were calculated as the length of the coding sequence upstream of the intron divided by the total length of the coding sequence (Mourier and Jeffares, 2003).

Duplication Method of 25 IR-Type Genes

Our AS-based hypothesis suggests that both retroposition and gene duplication are possible modes for which these IR-type genes can arise. Of the 25 IR-type genes with AS support, we next wanted to identify the method of duplication from which they arose. To identify the IR-type genes that arose through DNA duplication, we manually checked the synteny of the 10 genes flanking the 25 duplicated genes of interest and their parent genes (five genes on either side). If at least two genes were syntenic between the parental gene region and its duplicate, we concluded that the daughter gene was duplicated through segmental duplication (see "Materials and Methods"). We found four gene pairs from rice that have syntenic regions based on our parameters (Table I), suggesting these genes were formed through RNA-based gene conversion with alternatively spliced, IR isoform transcripts after DNA duplication.

To identify the IR-type genes that arose through retroposition, we tried to find remnants of a poly A/T tail in the daughter copies. We extract 1,000 bp from both sides of the daughter copy and count the poly A/Ts using a 20-bp sliding window; if there were more than 16 A's or 16 T's inside the window, we defined it as a poly A/T tail. Finally, we found three human, six rice, and four Arabidopsis IR-type genes that have poly A/T tails (Table I), suggesting these genes were duplicated through retroposition of the alternatively spliced isoform. The 13 genes that show traces of a poly A/T tail are different genes than the four mentioned above with syntenic support for DNA duplication (Table I).

None of the 25 genes investigated had both syntenic surrounding regions and a poly A tail, which is consistent with our hypothesis. Out of the 25 genes, only seven did not show a trace of DNA duplication or retroposition. Over time, signatures (poly A/T) of retroposition can degenerate and mutations will naturally accumulate, due to the lack of selective pressures (Bacon et al., 2001; Tijsterman et al., 2002). Chromosomal rearrangements and gene deletion can also occur, which can affect the sequence structure and synteny of a genic region (Prince and Pickett, 2002; Juretic et al., 2005; Vinckenbosch et al., 2006). According to the Ks (for the number of synonymous substitutions per synonymous site) value calculated by gKaKs (Zhang et al., 2013; Supplemental Table S4), these seven genes are not ancient, yet traces of the duplication method may have been lost.

IR-Type Genes Are More Abundant in Plants Than Animals

The duplicated gene structure data indicates that the number of IL-type to IR-type genes varied considerably between animals and plants. Out of the total number of unique copies that have experienced an IL event, 8.2% were IR type in *Drosophila* spp. (91.8%

were IL type) and 10.9% were IR type in human (89.1% were IL type), whereas 29.5% were IR type in rice (70.5% were IL type) and 50.0% were IR type in Arabidopsis (50.0% were IL type). The percentages of IR-type genes are much greater in plant species than in animal species. A Pearson's χ^2 test for independence (2×4 tables; Supplemental Table S5) on these data all have very significant P values ($P < 2.2 \times 10^{-16}$). The percentages of IR-type genes are significantly greater in the plant species than in the animal species, indicating more IR events in plant gene duplicates.

Why are there significantly higher incidences of IR duplicates in plants compared with animals? The AS data in plants and animals provide a clue (Table II). In the *Drosophila* spp. and human genomes, previous studies show that more than 60% (Graveley et al., 2011) and 90% (Pan et al., 2008; Wang et al., 2008) of multiple exon genes are involved in AS, whereas recent studies in plants have found AS rates closer to 50% to 60% (Filichkin et al., 2010; Lu et al., 2010; Marquez et al., 2012; Table II). The most common type of AS in *Drosophila* spp. and human is exon skipping, which occurs when an exon is spliced out along with its flanking introns (Fig. 1), accounting for about 40% of the AS events (Keren et al., 2010). Even though the AS rates are lower in plants, they have a higher percentage of intron retention AS, where an intron remains in the mature RNA transcript as part of the exon (Fig. 1); intron retention accounts for about 45.1% to 55% of the AS events in rice and approximately 30% to 64.1% in Arabidopsis (Table II) but occurs at a much lower frequency in *Drosophila* spp. and human, with rates around 5% to 15% (Table II). The AS-based method predicts that the IR-type duplicates arise through retroposition or gene conversion with an intron retention AS transcript. Therefore, because a greater percentage of plant AS events are intron retention, we can imagine there are more intron retention mRNAs in plant cells than in *Drosophila* spp. and human cells, and because this mRNA is the resource used in RNA-based gene conversion and in retroposition, then there is a greater chance of IR-type genes to become incorporated into the genomes of plants than in the animals.

The difference in the type of AS that is most common in plants and animals suggests different splice site recognition mechanisms and different roles of AS in plants versus animals (Barbazuk et al., 2008). About one-half of the intron retention events in Arabidopsis and rice are subject to nonsense-mediated decay (Barbazuk et al., 2008). Nonsense-mediated decay has been linked to the regulation of gene expression, suggesting that AS may be an important regulatory

Table II. *AS and the intron retention rate of four species from previous studies*

The data presented here may not have been directly presented in the previous papers but calculated from the data they provide.

Species	AS Rate	Intron Retention Rate	Exon Skipping Rate	Reference
		%		
Fly	40	—	—	Stolc et al., 2004
	18.6	30.8	13.6	Nagasaki et al., 2005
	—	Approximately 10	Approximately 32	Kim et al., 2007
	60.8	Approximately 11[a]	Approximately 11.4	Graveley et al., 2011
Human	63[b]	36[c]	—	Kan et al., 2002
	—	14.8[d]	52	Galante et al., 2004
	32.1	15.8	28.8	Nagasaki et al., 2005
	—	<10	Approximately 42	Kim et al., 2007
	94	Approximately 1	Approximately 35	Wang et al., 2008
	95	—	—	Pan et al., 2008
	88	41[e]	—	Mollet et al., 2010
Rice	8.1	55	55.0	Nagasaki et al., 2005
	21.2	53.5	13.8	Wang and Brendel, 2006
	32.5	45.1	12.8	Campbell et al., 2006
	48	—	—	Lu et al., 2010
	27.7	51.9	14.9	Severing et al., 2009
Arabidopsis	1.2	—	6.4	Zhu et al., 2003
	Between 7 and 10	30.5	3.2	Ner-Gaon et al., 2004
	11.6	44.8[f]	15.5	Iida et al., 2004
	14.1	42.8	42.8	Nagasaki et al., 2005
	21.8	56.1	8	Wang and Brendel, 2006
	23.5	47.9	6.8	Campbell et al., 2006
	42	64.1[g]	—	Filichkin et al., 2010
	—	Approximately 30	Approximately 5	Kim et al., 2007
	61	40	Approximately 6.8	Marquez et al., 2012
	24.4	51.9	10.0	Severing et al., 2009

[a]Ranges from 6.2% to 22.2% from Table I in the paper. [b]For high coverage rate genes, it is up to 99%. [c]Less than 5% of all genes exhibited intron retention at a 95% confidence interval ($P < 0.05$). [d]Intron retention is 4.6% of the elite group. [e]Unconstrained analysis. [f]790/1,764 = 0.448. [g]Using the novel splice isoforms data in Figure 4B in the paper, 6,000/(3,307 + 775 + 5,273) = 0.641.

mechanism in plants, whereas AS in animals is more commonly linked to expanding protein diversity (Barbazuk et al., 2008; Severing et al., 2009).

CONCLUSION

Gene duplicates and their progenitors that have slightly varied gene structures are often excluded from retrotransposons analyses due to the inclusion of an intron, and excluded from gene duplication analyses due to the loss of introns. These genes are often lost in the pipeline abyss, and very few studies have investigated the structural differences between parent and child duplicates. Here, we propose an AS-based model, which can be used in conjunction with the retroposition and Fink model, to explain the IR-type events observed in gene duplicates. Our results provide that some of the genes that have introns that remain in the daughter copy compared with the parent copy have intron retention AS isoforms, which supports our model. We found a relatively even distribution location where introns have been retained in the duplicated genes, as predicted by the AS-based model, and found cases in which IR-type genes were duplicated through DNA duplication and retroposition, which demonstrates how the AS model can be combined with previously reported models. We found a greater number of IR-type genes in plants compared with animals, which may be due to the abundance of intron retention AS cases that occur in plants. Overall, considering the prevalence of AS in eukaryotes, the AS-based model may provide greater explanatory power than the Fink model and the classic retroposition model to understand the evolution of exon-intron structure complexity.

MATERIALS AND METHODS

Identifying IL events in *Drosophila spp.*, Human, Rice, and Arabidopsis

In our analysis, we used genome data sets from *Drosophila* spp., human (*Homo sapiens*), rice (*Oryza sativa*), and Arabidopsis (*Arabidopsis thaliana*); the data from human and *Drosophila* spp. were formatted using a Perl script, and the detail links for each file are presented below.

We downloaded the fly data from Flybase, obtaining the CDS (Coding DNA Sequence; ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.47_FB2012_05/fasta/dmel-all-CDS-r5.47.fasta.gz), genome sequences (ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.47_FB2012_05/fasta/dmel-all-chromosome-r5.47.fasta.gz), and annotations (ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.47_FB2012_05/gff/dmel-all-r5.47.gff.gz), and using only the annotations generated by Flybase. The CDS file and annotation file is reformatted to use the transcript identification as the unique identification.

For human, we download data from National Center for Biotechnology Information genomes, obtaining the genome sequences (version GRCh37.p9) and annotations (version GRCH37.p9) and using only the annotations generated by RefSeq. The CDS data were extracted from the human genome sequence using the locations denoted in the human gene annotations.

We downloaded the rice genome sequence (all.con), annotation (all.gff3), and CDS (all.cds) data directly from Michigan State University (ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/all.dir/).

We downloaded the Arabidopsis data from the Web site of The Arabidopsis Information Resource, obtaining the whole genome sequences (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_chromosome_files/TAIR10_chr_all.fas), annotations (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_gff3/TAIR10_GFF3_genes.gff), and CDS (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_blastsets/TAIR10_cds_20110103_representative_gene_model_updated).

We first extracted the 25-bp-long exon boundary sequences on both sides of introns from every annotated gene with more than one exon (candidate parental genes), following the annotations in the annotation files downloaded (see above), and then combined the two boundary sequences into one sequence. We aligned these 50-bp-long artificial sequences via BLAST (using blastall command version 2.2.25 with blastn parameter; Camacho et al., 2009) to the entire respective genomes (see above; output file parameters set to –m8); hits with more than a 30-bp alignment length and without gaps were chained together and then were expanded to approximately the length of the query gene on both sides of the chain (Fig. 3, top right). Subsequently, the candidate parental gene CDSs were translated into protein sequences and aligned to these enlarged hit sequences using TFASTY (version 35; Pearson, 2000), with –q –m1 –m9c setting options. The alignment region of the hit sequences was aligned with BLAT (Kent, 2002) back to the genome data. We used the overlap region in the query sequence as the threshold and kept the copy that has the longest length (excluding the query sequence). If the best hit was not the query parental gene, it was removed from the data set. We then used Perl scripts to scan the alignment files to compare the query parental gene structure to the hit copy sequence structure (Fig. 3, bottom right) to identify structural differences at intron sites. If the relative additional sequences at the intron site of the hit sequences was larger than 30 bp, we defined it as an IR event, if it was 1 to 30 bp, we defined it as an intron indel, and if it is 0 bp (no sequence at the intron site), we defined it as an IL event. Because some of the gene duplicates identified may be due to partial DNA duplications (Zhang et al., 2011), we paid close attention to cases where the intron site was located at the alignment boundary and defined these as intron void (Fig. 3, bottom right). If the hit copy only had an IL event, we classified it as an IL type, and if the hit copy had both IL and IR events, and we classified it as an IR type (Fig. 3, left). We removed hit redundancy for both the IL and IR types if they located at the same chromosome region. TFASTY alignment results for 25 manually checked IR-type gene pairs are presented in Supplemental Material S1. Other alignment data and analyzed results are available upon request.

Identifying IR-Type Genes with AS Support

We used Perl script to extract the Gene identifications for the intron retention AS isoforms for rice (4028) and Arabidopsis (2760) from ASIP (http://www.plantgdb.org/ASIP/; 2007 version; Wang and Brendel, 2006), human (2028) from ASPicDB (http://t.caspur.it/ASPicDB/; Martelli et al., 2011), and *Drosophila* spp. (2386) from DEDB (http://proline.bic.nus.edu.sg/dedb; Lee et al., 2004). We then compared the intron retention AS isoform identifications to the parental gene identifications, whose hit copy had at least one but no more than two IR events, to find the parental genes that have AS isoforms. The parental AS sites were manually compared on ASIP, ASPicDB, and DEDB for rice and Arabidopsis, human, and *Drosophila*, respectively, to the intron retention sites in the duplicate copy (Supplemental Tables S1–S3). If the location of the intron retention AS event in the parental gene coincides with the IR location in the duplicate copy (supported by the RNA sequencing, EST, and full-length cDNA data), we considered it as a real IR copy (Fig. 3).

Identifying Duplication Methods

For the 25 IR-type duplicates with AS support, we manually investigated the sequence and synteny of the surrounding genic regions to identify the mode in which these genes arose. For rice and Arabidopsis, we first extracted the protein sequences of five flanking genes on either side of the parental gene and daughter copies and then performed a BLASTP (Camacho et al., 2009) alignment to identify genes shared between the two. If there are two or more genes that are found in the regions surrounding both the parental and daughter copies, then the parental gene and the daughter copy were deemed as belonging to a syntenic region, which suggested a segmental duplication may occur in the evolution. For the human data, we first identified the location of the parental gene and the daughter copy and then expanded the region to an additional 270 kb on both side (approximately 10 times the average human

gene size) to include about five flanking genes on either side of the target gene (we assume the average size of the intergenic region is about the same as the genic region). We extract the nucleotide sequences for the genes annotated in this region from the unedited human .gff file and then preformed a BLASTN (Camacho et al., 2009) alignment with the genes flanking the parental gene and duplicate copy. We set the identity parameter to greater than 50% and the match length to greater than 50 amino acid (>150 bp for human). We manually checked for matches due to repeat sequences (for example, Alu- and long terminal repeat-related repeat sequences).

We also searched for poly A/T tails in the 1,000 bp of sequence upstream and downstream of the duplicate copy. Using a 20-bp sliding window, we scanned the 1,000-bp sequence on either side of the copy, and if there were 16 or more A's or T's inside the window, we denoted it is a poly A/T tail.

Ks Value Calculation

We employed the pipeline gKaKs (Zhang et al., 2013) to calculate the Ks value for the gene pairs from rice and human and the YN00 (Yang and Nielsen, 2000) method to calculate the Arabidopsis gene pairs' Ks values, because the Arabidopsis gene pairs are too diverged to use the gKaKs method. When calculating the Arabidopsis gene pair, the alignment was done by MEGA (Tamura et al., 2011). We first translate the CDS into protein sequences, then after the alignment, it was translated back to nucleotide sequences, and all the gaps and nonconserved regions were deleted before calculation.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Gene structures of remaining 19 IR-type pairs.

Supplemental Table S1. Forty IR-type pairs in rice have intron retention in parental copy according to the ASIP database.

Supplemental Table S2. Thirty-two IR-type pairs in Arabidopsis have intron retention in parental copy according to the ASIP database.

Supplemental Table S3. Seventy-three IR-type pairs in human have intron retention in parental copy according to the ASPicDB database.

Supplemental Table S4. dN and dS value among gene pairs using gKaKs and YN00 methods.

Supplemental Table S5. χ^2 test for independence (2 × 4 tables), redundancy, and unique data.

Supplemental Material S1. Twenty-five IR-type pairs using TFASTY (version 35) alignment file (the red arrow indicates the intron site).

ACKNOWLEDGMENTS

We thank Clause Kemkemer, Ben Krinsky, Li Zhang, Yuan Huang, Muhua Wang, and Walter Gilbert for the fruitful discussion and our editor and two reviewers for their comments.

Received October 31, 2013; accepted January 27, 2014; published February 11, 2014.

LITERATURE CITED

Bacon AL, Dunlop MG, Farrington SM (2001) Hypermutability at a poly (A/T) tract in the human germline. Nucleic Acids Res **29:** 4405–4413

Barbazuk WB, Fu Y, McGinnis KM (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. Genome Res **18:** 1381–1392

Betrán E, Thornton K, Long M (2002) Retroposed new genes out of the X in *Drosophila*. Genome Res **12:** 1854–1859

Blencowe BJ (2006) Alternative splicing: new insights from global analyses. Cell **126:** 37–47

Brosius J (1991) Retroposons: seeds of evolution. Science **251:** 753

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics **10:** 421

Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics **7:** 327

Chen S, Krinsky BH, Long M (2013) New genes as drivers of phenotypic evolution. Nat Rev Genet **14:** 645–660

Derr LK (1998) The involvement of cellular recombination and repair genes in RNA-mediated recombination in *Saccharomyces cerevisiae*. Genetics **148:** 937–945

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

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome Res **20:** 45–58

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

Fu B, Chen M, Zou M, Long M, He S (2010) The rapid generation of chimerical genes expanding protein diversity in zebrafish. BMC Genomics **11:** 657

Galante PAF, Sakabe NJ, Kirschbaum-Slager N, de Souza SJ (2004) Detection and evaluation of intron retention events in the human transcriptome. RNA **10:** 757–765

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al (1996) Life with 6000 genes. Science **274:** 563–567

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al (2011) The developmental transcriptome of *Drosophila melanogaster*. Nature **471:** 473–479

Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, Shinozaki K (2004) Genome-wide analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-length cDNA sequences. Nucleic Acids Res **32:** 5096–5103

Iwamoto M, Maekawa M, Saito A, Higo H, Higo K (1998) Evolutionary relationship of plant catalase genes inferred from exon-intron structures: isozyme divergence after the separation of monocots and dicots. TAG Theor Appl Genet **97:** 9–19

Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE (2005) The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. Genome Res **15:** 1292–1297

Kaessmann H, Vinckenbosch N, Long M (2009) RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet **10:** 19–31

Kan Z, States D, Gish W (2002) Selecting for functional alternative splices in ESTs. Genome Res **12:** 1837–1845

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

Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet **11:** 345–355

Kim E, Magen A, Ast G (2007) Different levels of alternative splicing among eukaryotes. Nucleic Acids Res **35:** 125–131

Lee BTK, Tan TW, Ranganathan S (2004) DEDB: a database of *Drosophila melanogaster* exons in splicing graph form. BMC Bioinformatics **5:** 189

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 USA **99:** 8121–8126

Long M, Betrán E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet **4:** 865–875

Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Li W, et al (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res **20:** 1238–1249

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H (2005) Emergence of young human genes after a burst of retroposition in primates. PLoS Biol **3:** e357

Marquez Y, Brown JWS, Simpson C, Barta A, Kalyna M (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res **22:** 1184–1195

Martelli PL, D'Antonio M, Bonizzoni P, Castrignanò T, D'Erchia AM, D'Onorio De Meo P, Fariselli P, Finelli M, Licciulli F, Mangiulli M, et al (2011) ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing. Nucleic Acids Res **39:** D80–D85

Mollet IG, Ben-Dov C, Felício-Silva D, Grosso AR, Eleutério P, Alves R, Staller R, Silva TS, Carmo-Fonseca M (2010) Unconstrained mining of

transcript data reveals increased alternative splicing complexity in the human transcriptome. Nucleic Acids Res **38:** 4740–4754

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

Nagasaki H, Arita M, Nishizawa T, Suwa M, Gotoh O (2005) Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. Gene **364:** 53–62

Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, Rubin E, Ophir R, Fluhr R (2004) Intron retention is a major phenomenon in alternative splicing in Arabidopsis. Plant J **39:** 877–885

Pan D, Zhang L (2009) Burst of young retrogenes and independent retrogene formation in mammals. PLoS ONE **4:** e5040

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet **40:** 1413–1415

Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol **132:** 185–219

Prince VE, Pickett FB (2002) Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet **3:** 827–837

Roy SW, Gilbert W (2005) The pattern of intron loss. Proc Natl Acad Sci USA **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, Irimia M (2009) Mystery of intron gain: new data and new models. Trends Genet **25:** 67–73

Sakai H, Mizuno H, Kawahara Y, Wakimoto H, Ikawa H, Kawahigashi H, Kanamori H, Matsumoto T, Itoh T, Gaut BS (2011) Retrogenes in rice (*Oryza sativa* L. ssp. *japonica*) exhibit correlated expression with their source genes. Genome Biol Evol **3:** 1357–1368

Severing EI, van Dijk ADJ, Stiekema WJ, van Ham RCHJ (2009) Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. BMC Genomics **10:** 154

Shiao MS, Khil P, Camerini-Otero RD, Shiroishi T, Moriwaki K, Yu HT, Long M (2007) Origins of new male germ-line functions from X-derived autosomal retrogenes in the mouse. Mol Biol Evol **24:** 2242–2253

Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE, et al (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science **306:** 655–660

Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV (2005) Conservation versus parallel gains in intron evolution. Nucleic Acids Res **33:** 1741–1748

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol **28:** 2731–2739

Tijsterman M, Pothof J, Plasterk RHA (2002) Frequent germline mutations and somatic repeat instability in DNA mismatch-repair-deficient *Caenorhabditis elegans*. Genetics **161:** 651–660

Vinckenbosch N, Dupanloup I, Kaessmann H (2006) Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA **103:** 3220–3225

Wang BB, Brendel V (2006) Genomewide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci USA **103:** 7175–7180

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. Nature **456:** 470–476

Wang W, Yu H, Long M (2004) Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. Nat Genet **36:** 523–527

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, et al (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell **18:** 1791–1802

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol **17:** 32–43

Zhang C, Wang J, Long M, Fan C (2013) gKaKs: the pipeline for genome-level Ka/Ks calculation. Bioinformatics **29:** 645–646

Zhang Y, Wu Y, Liu Y, Han B (2005) Computational identification of 69 retroposons in Arabidopsis. Plant Physiol **138:** 935–948

Zhang YE, Vibranovski MD, Krinsky BH, Long M (2011) A cautionary note for retrocopy identification: DNA-based duplication of intron-containing genes significantly contributes to the origination of single exon genes. Bioinformatics **27:** 1749–1753

Zhu W, Schlueter SD, Brendel V (2003) Refined annotation of the Arabidopsis genome by complete expressed sequence tag mapping. Plant Physiol **132:** 469–484

Zhu Z, Zhang Y, Long M (2009) Extensive structural renovation of retrogenes in the evolution of the *Populus* genome. Plant Physiol **151:** 1943–1951