Rate and Polarity of Gene Fusion and Fission in *Oryza sativa* and *Arabidopsis thaliana*

Yoji Nakamura,† Takeshi Itoh,‡ and William Martin†*

*Division of Bioengineering and Bioinformatics, Graduate School of Information Science and Technology, Hokkaido University, Kita-ku, Sapporo, Japan; †Institut für Botanik III, Heinrich-Heine Universität Düsseldorf, Düsseldorf, Germany; and ‡Genome Research Department, National Institute of Agrobiological Sciences, Kannondai, Tsukuba, Ibaraki, Japan

Eukaryotic gene fusion and fission events are mechanistically more complicated than in prokaryotes, and their quantitative contributions to genome evolution are still poorly understood. We have identified all differentially composite or split genes in 2 fully sequenced plant genomes, *Oryza sativa* and *Arabidopsis thaliana*. Out of 10,172 orthologous gene pairs, 60 (0.6% of the total) revealed a verified fusion or fission event in either lineage after the divergence of *O. sativa* and *A. thaliana*. Polarizing these events by outgroup comparison revealed differences in the rate of gene fission but not of gene fusion in the rice and *Arabidopsis* lineages. Gene fission occurred at a higher rate than gene fusion in the *O. sativa* lineage and was furthermore more common in rice than in *Arabidopsis*. Nucleotide insertion bias has promoted gene fission in the *O. sativa* lineage, consistent with its generally longer nucleotide sequences than *A. thaliana* in selectively neutral regions, and with the abundance of transposable elements in rice. The divergence time of monocots and dicots (140–200 Myr) indicates that gene fusion/fission events occur at an average rate of 1×10^{-11} to 2×10^{-11} events per gene per year, ~100-fold slower than the average per site nuclear nucleotide substitution rate in these lineages. Gene fusion and fission are thus rare and slow processes in higher plant genomes; they should be of utility to address deeper evolutionary relationships among plants—and the relationship of plants to other eukaryotic lineages—where sequence-based phylogenies provide equivocal or conflicting results.

Introduction

In eukaryotic gene fusion, 2 or more separate transcription units are joined, forming 1 transcription unit. Gene fission is the converse process in which a gene is split into 2 or more separate transcription units. The mutational mechanisms affecting gene fusions and fissions differ in prokaryotes and eukaryotes. In prokaryotes, operons are common (Price et al. 2005), and operon organization can render genes readily predisposed to translational fusions. In eukaryotes, introns are common, such that mutations affecting splicing and recombination within introns can readily lead to novel fusions or fissions. Gene fusion and fission can contribute to the generation of novel eukaryotic gene structures, but both processes are thought to be less common than the other mechanisms that produce novel sequences (gene duplication and nucleotide substitution) because fusion and fission cause drastic changes in the higher-order organization of the encoded proteins. Previous studies on naturally occurring gene fusion events have focused on inferring protein function and protein–protein interaction (Enright et al. 1999; Marcotte et al. 1999; Enright and Ouzounis 2001; Yanai et al. 2001; Suhre and Claverie 2004). From the genome evolutionary perspective, however, the dynamics and specifics of gene fusion or fission events are yet poorly understood. Although several studies using multiple species have reported the tendencies of gene fusion and fission across taxa, such studies have been mostly limited to extremely compact genomes such as in prokaryotes or yeast (Snel et al. 2000; Yanai et al. 2001; Suhre and Claverie 2004). Because functionally related genes tend to be organized as operons in prokaryotic genomes, translational fusion or fission can occur by simple mutational changes. Studies of gene fusion and fission in large eukaryotic genomes are yet rare (Kummerfeld and Teichmann 2005) and are complicated by the circumstances that 1) the number of genes in many sequenced eukaryotic genomes is yet unknown, 2) gene annotation errors exist, and 3) alternative splicing can make it difficult to ascertain correct gene structures for comparison.

Recently, the genome sequence of rice, *Oryza sativa* (*O. sativa* L. ssp. *japonica* cv. Nipponbare), has been determined. Its gene repertoire was quite thoroughly annotated using full-length cDNA libraries (International Rice Genome Sequencing Project 2005; Ohyanagi et al. 2006). This permits a monocot–dicot comparison to the *Arabidopsis thaliana* genome (Arabidopsis Genome Initiative 2000). Here, we addressed the evolutionary dynamics of gene fusion and fission events in these plant genomes. We identify all of the candidates of gene fusion or fission events, which have occurred after the divergence of *O. sativa* and *A. thaliana*. We report the number and rate of the events including all genes and coordinates involved as well as their functional annotations and reconstruct the evolutionary scenario of differential gene fusion and fission in each lineage.

Key words: gene fusion and fission, introns, transposable elements, plant phylogeny.

E-mail: yojnakam@ist.hokudai.ac.jp.

Mol. Biol. Evol. 24(1):110–121. 2007
doi:10.1093/molbev/msl138
Advance Access publication October 11, 2006

Materials and Methods

Protein Sequences in *O. sativa* and *A. thaliana*

We collected a total of 40,041 protein sequences in *O. sativa* genomes annotated in the Rice Annotation Project (RAP) as of 14 June 2005 (Ohyanagi et al. 2006) and 28,860 protein sequences in *A. thaliana* in GenBank. We then checked the locations of protein-coding genes on genomes and whether their overlap was due to alternative splicing or redundant annotation, using longer ones if locations overlapped. This yielded a total of 28,759 protein sequences from *O. sativa* and 26,364 sequences from *A. thaliana*. Among those *O. sativa* sequences, 21,818 are supported by full-length cDNA in RAP. To check the *A. thaliana* sequences, we downloaded 15,295 full-length cDNA records as of 24 May 2005 from RIKEN ftp site (http://rarge.gsc.riken.jp/archives/rafl/sequence/) and confirmed

that 12,923 out of 26,364 sequences correspond to full-length cDNA entries.

Detection of Gene Fusion or Fission Candidates (one-to-many orthologous pairs)

We constructed a database using the protein sequences in *O. sativa* and *A. thaliana* and performed protein similarity search for each sequence using BlastP (Altschul et al. 1997) with the threshold *e* value $< 10^{-10}$. We then collected one-to-one reciprocal best match pairs between *O. sativa* and *A. thaliana* as orthologous pairs. From these, we selected the pairs in which the query in a species has more than 1 hit in the other species, and the query is the best match from these hits in the backward search. We checked these hits in the order of BlastP score and discarded the hits matching to the one of higher score as possible paralogues. The hit sequences obtained are, therefore, the best, second best, … and the *n*-th best hits from the query.

Validation of Gene Fusion or Fission Candidate Pairs

Next, we measured the overlapping length of matched regions on the query sequence in BlastP alignment in each of one-to-many orthologous pairs detected. Here, the overlap ratio of 2 hits A and B is

$$ratio = \text{overlap length between A and B} / \min\{\text{aligned length of A}, \text{aligned length of B}\}.$$

Then, we chose the query-hit pairs with the cutoff ratio < 0.3 following the bimodal distribution (supplementary fig. 1, Supplementary Material online). We excluded the pairs in which split genes are defined in a single locus by the RAP (Ohyanagi et al. 2006) because these might not be reliable annotations.

Furthermore, we validated these pairs by the following 2 steps using BlastP and TBlastN: 1) we performed BlastP searches using the protein sequences in each pair to public databases, GenBank/European Molecular Biology Laboratory (EMBL)/DNA Data Bank of Japan (DDBJ) and Swiss-Prot, and compared the gene structures with those of entries in the databases. We did not use the pairs for further analysis if 1–1) the pair in which the query as a composite gene has separate entries as component genes from the same species in the databases or 1–2) the pairs in which the hits nearby located on a chromosome have a composite entry from the same species in the databases. Here, nearby located genes are defined as the ones between which there are 3 or less other genes. 2) We performed TBlastN using each of composite genes as a query, against noncoding regions around split genes with *e* value $< 10^{-3}$. We then concatenated the matched "exon-like" sequences, translated them into amino acid sequences, and compared the Blast alignment and score with those of the split genes. We did not use the pairs in which such an exon-like sequence next to 1 split gene is aligned with a higher Blast score than the other split gene.

Estimation of Gene Fusion or Fission

We performed Blast comparisons using the protein sequences in each pair of a composite gene and split genes to the gene sets from the red algae *Cyanidioschyzon merolae* (Matsuzaki et al. 2004), the green algae *Chlamydomonas reinhardtii* (http://genome.jgi-psf.org/Chlre3/Chlre3.home.html), and entries in GenBank/EMBL/DDBJ and Swiss-Prot by BlastP. Then we performed Blast comparisons using the top hits to the above-mentioned database of *O. sativa* and *A. thaliana* and chose the ones as orthologous outgroup genes with reciprocal best matching. Using these outgroups, we inferred the ancestral state of pair of a composite gene and split genes.

Assignment to Biological Function

We referred the gene function from RAP annotation for *O. sativa* genes. For *A. thaliana*, we used the GenBank annotation. To investigate the function at the domain level, we performed InterProScan (Zdobnov and Apweiler 2001) using the Pfam database (Finn et al. 2006) with *e* value $< 10^{-3}$. In each of the gene fusion/fission candidate pairs, we defined the pairs in which gene splits occur "within" domains by the following criterion: the domain region in a composite gene detected by InterProScan is aligned across the regions aligned to split genes by BlastP.

Repetitive Element Sequences

We downloaded the *Arabidopsis* and rice repeat sequences from The Institute for Genomic Research Plant Repeat Database (http://www.tigr.org/tdb/e2k1/plant.repeats/) and constructed a BlastN database. We then performed sequence homology searches for intergenic regions around the fusion/fission candidate genes examined with the threshold *e* value $< 10^{-5}$.

Graphical Views of Gene Fusion/Fission Candidate Pairs

We developed the Perl program package, "FUFIA viewer (gene FUsion and FIssion Alignment viewer)" for drawing the fusion/fission candidate pairs detected by Blast.

Results

Detection of Gene Fusion and Fission Events

A total of 10,172 one-to-one orthologous gene pairs between *O. sativa* and *A. thaliana* genomes were determined by reciprocal BlastP searches (see Materials and Methods). Out of those, 277 pairs were defined as one-to-many orthologous pairs in which 1 query (a composite gene) in a genome has more than 1 orthologous hit (split genes) in the other genome, and these hits are not paralogous to each other (Enright et al. 1999). After excluding the pairs in which the alignments of hits heavily overlapped (overlap ratio ≥ 0.3), we checked the RAP annotations and excluded the pairs in which *Oryza* split genes are defined by a single locus in RAP. Although these genes might be genuine split genes, here we adopted the RAP annotations. We thus obtained 114 conservative pairs as the preliminary fusion/fission candidates. Then we validated those pairs using BlastP and TBlastN (see Materials and Methods). We first found that in 45 pairs, either the rice or the *Arabidopsis* gene prediction was inconsistent with the public database entry. Next, for each of the remaining

Table 1
Number of Fusion or Fission Events in *Oryza sativa* and *Arabidopsis thaliana*

Candidate Pairs	Fusion	Fission	Unknown	Total
Arabidopsis-composite–*Oryza*-split	3[a]	6[b]	30	39
Oryza-composite–*Arabidopsis*-split	3[b]	2[a]	16	21
Total	6	8	46	60

[a] Events in *Arabidopsis* lineage.
[b] Events in *Oryza* lineage.

69 pairs, we detected 9 pairs in each of which an exon-like structure near 1 split gene is aligned to the composite gene with a higher Blast score than the other split gene. Due to the possibility that these 54 (45 + 9) genes are misannotated in the genome sequence, they were excluded from further analysis. This left a total of 60 candidate pairs encompassing a composite gene in a species and 2 or more split orthologues in the other species (table 1). Of these, 21 were composite in *O. sativa* and split in *A. thaliana* (*Oryza*-composite–*Arabidopsis*-split), whereas 39 pairs, nearly twice as many, were composite in *A. thaliana* and split in *O. sativa* (*Arabidopsis*-composite–*Oryza*-split).

Next, we investigated the locations and orientations of the genes in 60 candidate pairs (figs. 1, 2, and 4; supplementary figs. 2 and 3, Supplementary Material online). Out of the 39 *Arabidopsis*-composite–*Oryza*-split pairs, 21 are termed "distal" pairs because the 2 split genes are distantly located on the same chromosome or dispersed on different chromosomes. In these pairs, recombination or translocation of components might have directly caused fusion or fission or occurred after insertion or deletion had generated fused genes or fissioned genes (fig. 1*A*). Seventeen pairs are termed "proximal" because 2 split genes were separated by ≤3 other genes on the same chromosome (fig. 1*B*). In the majority of these pairs, 2 split genes lie next to each

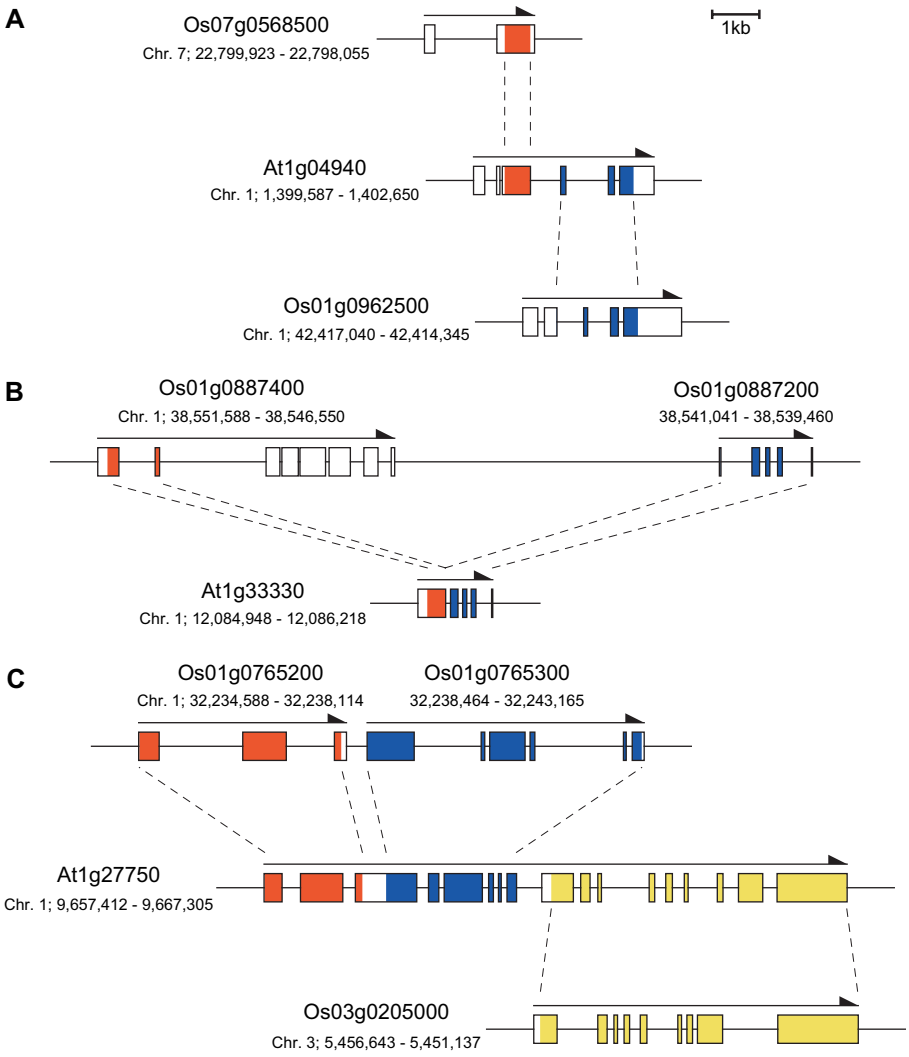

FIG. 1.—Alignments of gene fusion or fission candidates. Three examples of a composite gene in *Arabidopsis* (Atxgxxxxx) and split genes in *Oryza* (Osxxgxxxxxxx) are shown. Each is classified following the locations of split genes: distal (*A*), proximal (*B*), and the hybrid of distal and proximal (*C*). Aligned regions and their correspondences between *Oryza sativa* and *Arabidopsis thaliana* are shown as colors and dash lines. For each gene, its locus name and chromosomal position is shown, and the direction of transcription and range from initiation to termination codons are represented by an arrow. The same scale bar of basepairs is used in (*A*–*C*).

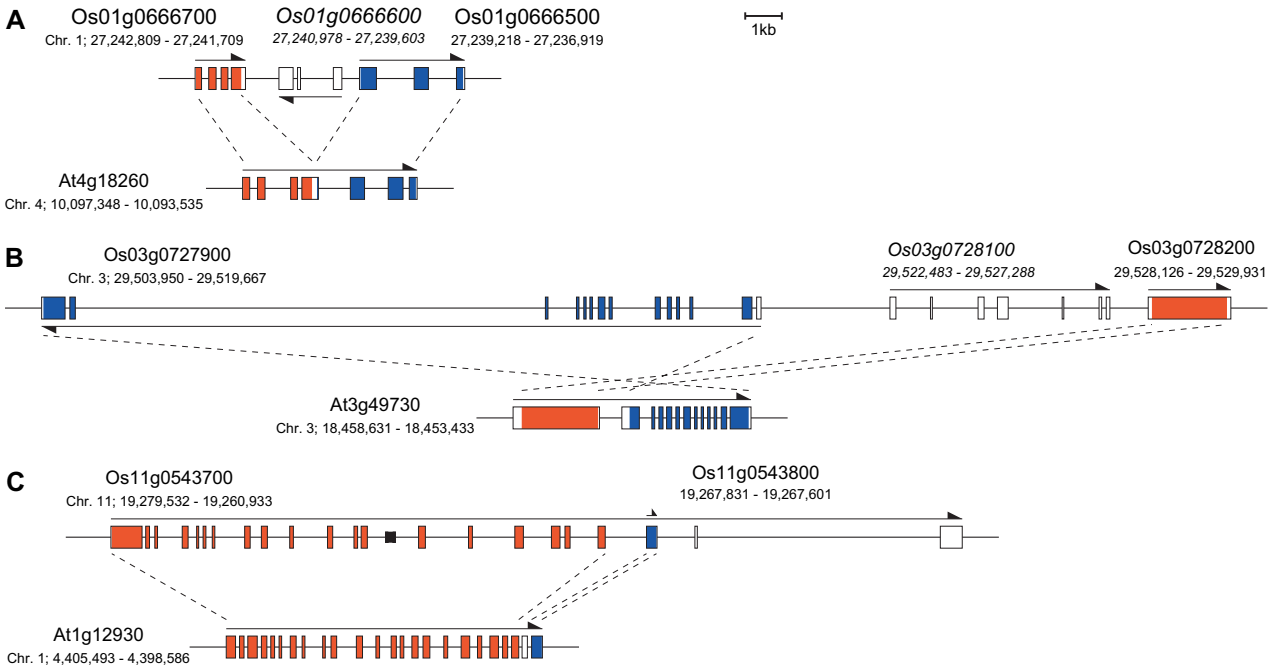

FIG. 2.—Special cases of "proximal" fusion/fission candidates. Each is classified following the locations and orientations of split genes: nearby but split by an unrelated gene (*A*), and nearby inverted (*B*), and one gene is located within another gene (*C*). Denotation of figure is the same as figure 1. In (*A*) and (*B*), unrelated genes are represented in italic. In (*C*), 1 repetitive sequence is shown as a half-size box in black. The same scale bar of basepairs is used in (*A–C*).

other in the same orientation. In this case, insertion or deletion within/between genes probably caused fusion/fission.

We further found 3 special "proximal" subclasses (fig. 2*A–C*). As the first subclass, we detected a pair in which there is an unrelated gene between split genes (fig. 2*A*), involving insertion or recombination. As the other special subclass, we detected a pair in which rice split genes were located nearby in inverted orientation (fig. 2*B*). The second subclass also may involve recombination, as in the case of distal split genes. In this class, however, there is an unrelated gene between split, inverted genes, implying insertion or deletion mechanisms. In the third special subclass, 1 split gene is nested within another split gene (fig. 2*C*). The remaining 1 pair out of 39 *Arabidopsis*-composite-*Oryza*-split pairs was the hybrid of "distal" and "proximal," which involved 3 genes (fig. 1*C*). In this pair, 1 of the split genes was located on a different chromosome, whereas the others are next to each other.

For *Oryza*-composite–*Arabidopsis*-split pairs, we classified the 21 pairs into 7 distal and 14 proximal pairs (table 3 and supplementary fig. 3, Supplementary Material online). Of the proximal pairs, we found a pair of the first special subclass but none of the second or third subclass. In 1 of the proximal pairs (Os01g0388500 vs. At2g48060-40), 3 *Arabidopsis* genes were of the same orientation on chromosome 2 (supplementary fig. 3, Supplementary Material online).

Frequent Gene Fissions in Rice

To determine the evolutionary polarity of gene fusion or fission, we inferred the ancestral states of the candidate pairs by outgroup comparison using BlastP of composite or split translations to National Center for Biotechnology Information and Swiss-Prot and the available translations from the recently sequenced plants *C. merolae* (a red algae) and *C. reinhardtii* (a green algae). We then defined orthologous outgroup genes from these databases by reciprocal BlastP and inferred the ancestral gene structures by parsimony. This defined polarity in 14 cases (6 fusions and 8 fissions) out of 60 pairs examined (table 1). Nine were *Arabidopsis*-composite–*Oryza*-split and 5 were *Oryza*-composite–*Arabidopsis*-split cases. Among the polarized cases, the *Oryza* lineage has undergone 3 fusions and 6 fissions, and the *Arabidopsis* lineage has undergone 3 fusions and 2 fissions. Hence, our result shows that gene fission is more common than gene fusion in the rice genome (6:3), whereas fissions and fusions are equally common in *A. thaliana* (2:3). Moreover, many rice fission genes were nearby located on the chromosome (table 2).

Biological Functions of Fused or Fission Genes

We investigated the functional annotations of fused or fissioned genes (tables 2 and 3). Although many of the candidate genes were hypothetical or unknown proteins, some were assigned to biological functions. In 22 pairs, 1 composite gene and 1 split gene were involved in the same or related function and the other split gene(s) encode different protein(s) or were unknown/hypothetical. In the other pairs, all the genes of *Oryza* and *Arabidopsis* were unknown/hypothetical genes, just expression-confirmed genes found in cDNAs or ESTs or assigned to different functions. These gene pairs are interesting candidates for functional analysis.

We detected domain regions in 47 out of 60 gene fusion/fission candidate pairs using the Pfam database (Finn et al. 2006). We then found that in 5 pairs, 3 in *Arabidopsis*-composite–*Oryza*-split pairs and 2 in *Oryza*-composite–*Arabidopsis*-split pairs, the split positions are located within

Table 2
Candidates of Gene Fusions in *Arabidopsis thaliana* or Fissions in *Oryza sativa* (*Arabidopsis*-composite–*Oryza*-split)

Location of Split Genes[a]	Composite Gene[b] (*Arabidopsis*)	Split Genes[b] (*Oryza*)	Chr.[c]	Length (aa)	FL-cDNA[d]	Fusion/ Fission[e]	Function[f]
''Distal''	At1g04940 (fig. 1*A*)		1	501			Tic20 family protein
		Os07g0568500	7	272	++		Conserved hypothetical protein
		Os01g0962500	1	551	++		Unknown protein
	At1g11760		1	393			Expressed protein
		Os10g0548400	10	160			Conserved hypothetical protein
		Os03g0197000	3	252	++		Conserved hypothetical protein
	At1g26760		1	967	+	Fusion	SET domain–containing protein
		Os08g0433300	8	381			Sialidase domain–containing protein
		Os03g0168700	3	536			TPR-like domain–containing protein
	At1g32120		1	1206			Expressed protein
		Os03g0565300	3	1030	++		Conserved hypothetical protein
		Os11g0621000	11	309	++		Unknown protein
	At1g49980		1	785	+	Fusion	**UmuC-like DNA repair family protein**
		Os03g0616300	3	617	++		**DNA-directed polymerase kappa**
		Os10g0350800	10	169			Hypothetical protein
	At1g61000		1	974		Fusion	**Nuf2 family protein**
		Os03g0577100	3	482	++		**Nuf2 family protein**
		Os03g0659900	3	664	++		S3 self-incompatibility locus–linked pollen 3.15 protein
	At2g30100		2	897			**Ubiquitin family protein**
		Os05g0353300	5	488			TPR-like domain–containing protein
		Os10g0456200	10	378	++		**Ubiquitin domain–containing protein**
	At3g02650		3	1077	+		PPR repeat–containing protein
		Os06g0179500	6	586			Plant protein of unknown function family protein
		Os01g0897500	1	108	++		Protein prenyltransferase domain–containing protein
	At3g23510 (fig. 4)		3	867	++		**CPA-FA synthase, putative**
		Os07g0474400	7	83		Fission	Adrenodoxin reductase family protein
		Os12g0267200	12	837	++		**Cyclopropane-fatty-acyl-phospholipid synthase family protein**
	At3g49140		3	1229	+		PPR repeat–containing protein
		Os03g0241800	3	760			TPR-like domain–containing protein
		Os11g0544000	11	462	++		Unknown protein
	At3g49640		3	519			Nitrogen regulation family protein
		Os10g0360900	10	270	++		Conserved hypothetical protein
		Os04g0531300	4	319	++		Dihydrouridine synthase, DuS family protein
	At4g14310		4	1087	+		Peroxisomal membrane protein related
		Os02g0809900	2	1030			Quinon protein alcohol dehydrogenase–like domain–containing protein
		Os08g0566900	8	187	++		Mpv17/PMP22 family protein
	At4g19900		4	1302			**Glycosyl transferase related**
		Os07g0567300	7	605			**Alpha 1,4-glycosyl transferase conserved region family protein**
		Os11g0607100	11	671	++		Protein prenyltransferase domain–containing protein
	At4g22760		4	889			PPR repeat–containing protein
		Os02g0448600	2	256	++		Hypothetical protein
		Os08g0162200	8	535	++		TPR-like domain–containing protein
	At4g26450		4	1248			Expressed protein
		Os02g0550000	2	770			Conserved hypothetical protein
		Os08g0497900	8	462	++		Conserved hypothetical protein
	At4g37920		4	673			Expressed protein
		Os04g0539000	4	220	++		Conserved hypothetical protein
		Os01g0306800	1	445	++		Conserved hypothetical protein
	T10O8.20		5	912			Unknown protein
		Os01g0831000	1	215	++		Transcription factor
		Os03g0293400	3	312	++		Aprataxin FHA-HIT
	MQD19.18		5	680			Unknown protein
		Os02g0304800	2	617			Protein prenyltransferase domain–containing protein
		Os10g0566900	10	194	++		Conserved hypothetical protein
	K17N15.9		5	860			Unknown protein
		Os06g0686500	6	707	++		Peptidase M3A and M3B, thimet/ oligopeptidase F family protein
		Os02g0125000	2	187	++		Conserved hypothetical protein
	MTE17.10		5	1332			Unknown protein

Table 2
Continued

Location of Split Genes[a]	Composite Gene[b] (*Arabidopsis*)	Split Genes[b] (*Oryza*)	Chr.[c]	Length (aa)	FL-cDNA[d]	Fusion/ Fission[e]	Function[f]
		Os08g0337300	8	585	++		FYVE/PHD zinc finger domain–containing protein
		Os08g0502000	8	694	++		Conserved hypothetical protein
	MTI20.26		5	1011			Unknown protein
		Os08g0280600	8	182	++		Conserved hypothetical protein
		Os05g0390500	5	536	++		NLI interacting factor domain–containing protein
''Proximal'' next to each other	At1g33330 (fig. 1*B*)		1	257	++		**Peptide chain RF, putative**
		Os01g0887400	1	682			**Peptide chain RF-1**
		Os01g0887200	1	110			Winged helix DNA–binding domain–containing protein
	At1g79280		1	2111	+		Expressed protein
		Os02g0741500	2	501	++		Methionine repressor–like domain–containing protein
		Os02g0741400	2	363	++		Conserved hypothetical protein
	At2g17930		2	3795	+		**FAT domain–containing protein / phosphatidylinositol 3- and 4-kinase family protein**
		Os07g0645200	7	294	++	Fission	Hypothetical protein
		Os07g0645100	7	842	++		**Phosphatidylinositol 3- and 4-kinase domain–containing protein**
	At2g19910		2	992			**RNA-dependent RNA polymerase family protein**
		Os01g0198100	1	456	++		Hypothetical protein
		Os01g0198000	1	582	++		**RNA-dependent RNA polymerase family protein**
	At2g26340		2	230	++		Expressed protein
		Os03g0176700	3	128			Hypothetical protein
		Os03g0176600	3	108			Hypothetical protein
	At2g46560		2	2471	+		**Transducin family protein/WD-40 repeat family protein**
		Os01g0552000	1	500	++		Hypothetical protein
		Os01g0551900	1	629	++		**WD-40–like domain–containing protein**
	At3g42670		3	1256	+		**SNF2 domain–containing protein/helicase domain–containing protein**
		Os07g0692500	7	569	++		Conserved hypothetical protein
		Os07g0692600	7	475	++		**SNF2-related domain–containing protein**
	At3g49410		3	559	+		Transcription factor related
		Os01g0528000	1	150	++	Fission	Hypothetical protein
		Os01g0528100	1	283			Winged helix DNA–binding domain–containing protein
	At3g49600		3	1067	+		**Ubiquitin-specific protease 26**
		Os03g0192800	3	317	++	Fission	**Peptidase C19, ubiquitin carboxyl-terminal hydrolase 2 family protein**
		Os03g0192900	3	89	++		Hypothetical protein
	At3g56330		3	433	+		N2,N2-dimethylguanosine tRNA methyltransferase family protein
		Os05g0324200	5	203	++	Fission	Hypothetical protein
		Os05g0324100	5	98	++		Winged helix DNA–binding domain–containing protein
	At4g02940		4	569	++		**Oxidoreductase, 2OG-Fe(II) oxygenase family protein**
		Os05g0401700	5	252			Conserved hypothetical protein
		Os05g0401500	5	318			**2OG-Fe(II) oxygenase domain–containing protein**
	At4g34100		4	1092	+		**Zinc finger (C3HC4-type RING finger) family protein**
		Os06g0639100	6	129	++	Fission	**Zinc finger, RING domain–containing protein**
		Os06g0639000	6	303	++		Conserved hypothetical protein
	T2L20.8		5	1165			Unknown protein
		Os05g0374500	5	677	++		TPR-like domain–containing protein
		Os05g0374600	5	394	++		Heat shock protein DnaJ, N-terminal domain–containing protein
	K23L20.15		5	2228			Unknown protein
		Os07g0497100	7	306	++		Zinc finger–like, PHD finger domain–containing protein
		Os07g0497000	7	622	++		Chromodomain helicase-DNA-binding protein Mi-2 homolog

Table 2
Continued

Location of Split Genes[a]	Composite Gene[b] (*Arabidopsis*)	Split Genes[b] (*Oryza*)	Chr.[c]	Length (aa)	FL-cDNA[d]	Fusion/ Fission[e]	Function[f]
Split by unrelated genes	At4g18260 (fig. 2*A*)		4	545	+		**Cytochrome B561 related**
		Os01g0666700	1	253			**Cytochrome B561/ferric reductase transmembrane domain–containing protein**
		Os01g0666500	1	287	++		Conserved hypothetical protein
Inverted	At3g49730 (fig. 2*B*)		3	1184	+		PPR repeat–containing protein
		Os03g0728200	3	601			Protein prenyltransferase domain–containing protein
		Os03g0727900	3	568	++		GTP1/OBG domain–containing protein
Nested	At1g12930 (fig. 2*C*)		1	1005	++		Importin related
		Os11g0543700	11	1065			ARM repeat fold domain–containing protein
		Os11g0543800	11	76	++		Hypothetical protein
Hybrid of ''distal'' and ''proximal''	At1g27750 (fig. 1*C*)		1	1973	+		**Ubiquitin system component cue domain–containing protein**
		Os01g0765200	1	431	++		Hypothetical protein
		Os01g0765300	1	598	++		RNA-binding region RNP-1 (RNA recognition motif) domain–containing protein
		Os03g0205000	3	927	++		**Ubiquitin system component cue domain–containing protein**

NOTE.—FL, full-length; PPR, pentatricopeptide; FHA-HIT, forkhead-associated domain–histidine triad–like protein; RF, release factor; SET, suvar3-9, enhancer-of-zeste, trithorax; TPR, tetratricopeptide repeat; FYVE, Fab1, YOTB/ZK632.12, Vac1, and EEA1; PHD, plant homeodomain; NLI, nuclear LIM interactor; FAT, FRAP, ATM, and TRRAP (FKBP12-rapamycin complex-associated protein, ataxia telangiectasia mutant, and transformation/transcription domain associated protein); RNP, RNA-binding protein.

[a] ''Proximal'' split genes are the ones separated by 3 or less other genes on the same chromosome. Other genes are classified into ''distal'' genes.

[b] *Arabidopsis* and *Oryza* genes are represented by locus names. Seven pairs viewed in the main text are noted by their figure numbers. Graphical views of other 32 pairs are shown in supplementary figure 2, Supplementary Material online.

[c] Chromosomal number.

[d] ++: Supported by FL cDNAs of RAP (*Oryza*) or RIKEN (*Arabidopsis*); +: not supported by RIKEN entries but described to be supported by cDNAs or massive parallel signature sequencing in GenBank.

[e] The event inferred by outgroup comparison. If not inferable, it is left blank.

[f] Similar functions are shown in bold.

domains (table 4). For 2 pairs of At3g23510 versus Os07g0474400–Os12g0267200 and At3g56330 versus Os05g0324200-100, gene fissions are inferred by outgroup comparison, and for others the directions are unknown.

Discussion

We identified 10,172 orthologous gene pairs, of which 60 confirmed pairs (0.6%) have undergone fusion or fission events after the divergence of *O. sativa* and *A. thaliana* (table 1). Even if we add to that number the 54 pairs excluded because of possible annotation errors, the percentage of differentially composite/split genes would still only rise to 1.1%. This paucity indicates that in these plant genomes, gene fusion or fission events are either mechanistically rare or often counterselected, or both. Of 60 pairs, we found that *Arabidopsis*-composite–*Oryza*-split cases (39) are nearly twice as common as *Oryza*-composite–*Arabidopsis*-split cases (21). This significant difference ($P < 0.05$) strongly indicates 3 possible polarities of gene fusion and fission events in each species: 1) frequent gene fusions in *Arabidopsis*, 2) frequent gene fissions in rice, or 3) both. Gene fission is twice as common as gene fusion in the rice genome, although it is not statistically significant due to the small number of observations (table 1). Because most gene splits detected involve proximal genes on the same chromosome, the issue arises whether these are true fusions/fissions or artifacts of annotation error. Here it is important to note that almost all of the fission genes in rice genome were supported by full-length rice cDNA records (table 2). Therefore, artifactual fissions due to frameshifts by sequencing errors can be excluded in the case of the rice genome. On the other hand, both gene fusion and fission are equally common in *A. thaliana*, implying a relative richness of gene fusion over fission as compared with rice (table 1).

The observed polarity trends are consistent with the length differences between orthologous regions in *O. sativa* and *A. thaliana*, intron lengths in particular. The distribution of intron lengths within orthologous genes showed a clear bimodal distribution: one conservative class and one shifted toward longer rice introns (fig. 3*A*). In the conservative distribution, the genes have few or no introns. The other component of the bimodal distribution indicates an insertion or deletion bias in introns. Moreover, the numbers of introns are not biased toward rice (fig. 3*B*), suggesting that the differences in length are not due to amplification or loss of introns but by nucleotide insertion or deletion within selectively neutral intron regions. Our observations reveal a "genome-wide" nucleotide insertion bias in the *Oryza* lineage and/or deletion bias in the *Arabidopsis* lineage after the divergence of these species.

Table 3
Candidates of Gene Fusions in *Oryza sativa* or Fissions in *Arabidopsis thaliana* (*Oryza*-composite–*Arabidopsis*-split)

Location of Split Genes[a]	Composite Gene[b] (*Oryza*)	Split Genes[b] (*Arabidopsis*)	Chr.[c]	Length (aa)	FL-cDNA[d]	Fusion/ Fission[e]	Function[f]
''Distal''	Os01g0884500		1	892	++		SWIB/MDM2 domain–containing protein
		MBK5.18	5	571			Unknown protein
		At2g16470	2	659	+		Zinc finger (CCCH-type) family protein/GYF domain–containing protein
	Os03g0432900		3	1837			MscS mechanosensitive ion channel family protein
		F12B17.160	5	519	++		Unknown protein
		MBK23.25	5	557			Unknown protein
	Os05g0497600		5	823			Ribosomal L11 methyltransferase family protein
		K9L2.3	5	486			Unknown protein
		K19P17.9	5	371			Unknown protein
	Os06g0228900		6	925			**Plant regulator RWP-RK domain–containing protein**
		At1g62260	1	656			PPR repeat–containing protein
		At1g18790	1	269			**RWP-RK domain–containing protein**
	Os06g0237300		6	1303	++	Fusion	**Zn-binding protein, LIM domain–containing protein**
		At1g10200	1	190	+		**Transcription factor LIM, putative**
		At3g05900	3	673	+		Neurofilament protein related
	Os10g0422300		10	679			TPR-like domain–containing protein
		MFB13.15	5	487	++		Unknown protein
		T1N24.12	5	226	++		Unknown protein
	Os11g0537300		11	324			RmlC-like cupin family protein
		At1g44960	1	261	++		Expressed protein
		MAC9.10	5	210			Unknown protein
''Proximal'' next to each other	Os01g0388500		1	2196			Conserved hypothetical protein
		At2g48060	2	621			Hypothetical protein
		At2g48050	2	1500	+		Expressed protein
		At2g48040	2	294	+		Expressed protein
	Os02g0281000		2	1086	++	Fusion	**Protein phosphatase 2C family protein**
		At2g20050	2	514			**Protein phosphatase 2C, putative/PP2C, putative**
		At2g20040	2	261	+		Protein kinase, putative
	Os02g0708600		2	563	++		**Nuclear protein SET domain–containing protein**
		At2g23740	2	907			Zinc finger (C2H2-type) family protein
		At2g23750	2	203			**SET domain–containing protein**
	Os02g0709800		2	679	++		RabGAP/TBC domain–containing protein
		F6N7.6	5	327			Unknown protein
		F6N7.7	5	338	++		Unknown protein
	Os03g0159200		3	467	++		**Protein of unknown function XS domain–containing protein**
		At3g22430	3	342		Fission	Expressed protein
		At3g22435	3	183	++		**XS domain–containing protein**
	Os03g0243800		3	331			Conserved hypothetical protein
		At4g35987	4	130	+		Expressed protein
		At4g35990	4	129			Hypothetical protein
	Os04g0442900		4	1376		Fusion	Zn-finger, CCHC-type domain–containing protein
		T15F17.6	5	341			Unknown protein
		T15F17.4	5	1158			Unknown protein
	Os07g0693400		7	957			ARM repeat fold domain–containing protein
		At3g08960	3	754			Importin beta-2 subunit family protein
		At3g08955	3	108	+		Expressed protein
	Os08g0101600		8	641	++		**Single-strand DNA endonuclease-1**
		At3g48900	3	337			**Single-strand DNA endonuclease, putative**
		At3g48910	3	224	++		Expressed protein
	Os08g0245400		8	821	++		Amino transferase class-III family protein
		MUA2.18	5	287	++	Fission	Unknown protein
		MUA2.17	5	523			Unknown protein
	Os09g0566100		9	1069	++		Protein of unknown function DUF618 domain–containing protein
		At2g36485	2	158	++		Expressed protein
		At2g36480	2	828	+		Zinc finger (C2H2-type) family protein
	Os10g0181200		10	1021	++		TPR-like domain–containing protein
		At4g34830	4	749	+		PPR repeat–containing protein

Table 3
Continued

Location of Split Genes[a]	Composite Gene[b] (*Oryza*)	Split Genes[b] (*Arabidopsis*)	Chr.[c]	Length (aa)	FL-cDNA[d]	Fusion/ Fission[e]	Function[f]
		At4g34820	4	321	+		Expressed protein
	Os12g0209700		12	1432			**Zinc finger–like, PHD finger domain–containing protein**
		At4g10940	4	192			**PHD finger family protein**
		At4g10930	4	984	+		Expressed protein
Split by unrelated genes	Os11g0706600		11	517			Thaumatin, pathogenesis-related family protein
		T7H20.160	5	341			Unknown protein
		T7H20.190	5	294			Unknown protein

NOTE.—FL, full-length; PPR, pentatricopeptide; GYF, glycine-tyrosine-phenylalanine; LIM, Lin-11 Isl-1 Mec-3; TBC, Tre-2, BUB2p, and Cdc16p; XS, rice gene X and SGS3.

[a] ''Proximal'' split genes are the ones separated by 3 or less other genes on the same chromosome. Other genes are classified into ''distal'' genes.

[b] *Arabidopsis* and *Oryza* genes are represented by locus names. All of the graphical views of gene pairs are shown in supplementary figure 3, Supplementary Material online.

[c] Chromosomal number.

[d] ++: Supported by FL cDNAs of RAP (*Oryza*) or RIKEN (*Arabidopsis*); +: not supported by RIKEN entries, but described to be supported by cDNAs or massive parallel signature sequencing in GenBank.

[e] The event inferred by outgroup comparison. If not inferable, it is left blank.

[f] Similar functions are shown in bold.

It has been reported that transposable elements are abundant in rice, occupying more than one-third of the genome (International Rice Genome Sequencing Project 2005), whereas the corresponding value is 10% in *Arabidopsis* (Arabidopsis Genome Initiative 2000). The rice genome is thus apparently prone to nucleotide insertion bias, and it is expected that the remnants of transposable element–related sequences exist in very recent fission cases. In this study, we found 3 "proximal" pairs, At1g79280 versus Os02g0741500-400, At2g17930 versus Os07g0645200-100, and At2g19910 versus Os01g0198100-000, in which repetitive sequences exist between split genes (supplementary fig. 2, Supplementary Material online). Of these, 1 (At2g17930 vs. Os07g0645200-100) is inferred as a rice gene fission by outgroup comparison. Although it is still unresolved for the other 2 cases because of lacking outgroups, they are also probably rice gene fission pairs. Because conserved exon-like sequences are still observed between these fission genes, the fission events appear to have occurred recently (supplementary fig. 2, Supplementary Material online). Furthermore, figure 4 reveals a composite gene, At3g23510 in *Arabidopsis*, that is fissioned into 2 genes, Os07g0474400 and Os12g0267200, on different chromosomes in the rice genome. A repetitive sequence is inserted downstream of Os07g0474400, implying that it might have disrupted the expression as a composite gene. Although we observed an exon-like structure homologous to At3g23510 further downstream of Os07g0474400, it is partial, and most of the counterpart exons are encoded in Os12g0267200. Therefore, this case indicates a gene fission mediated by transposable elements in which a gene is split by transposable elements after gene duplication and a part of the gene is inactivated.

We found that the points of gene splits are located within domains in only 5 out of 47 gene fusion/fission candidate pairs in which domains are detected. This suggests that gene fusion or fission events can be fixed more readily if they occur in such a manner as preserves domain structures and gene functions, in turn, to some extent. From this, it would appear that most of the observed gene fusion/fission events are not deleterious. Because it is less likely that fusion of nondomain or partial domain sequences results in the innovation of novel domain sequences, all of the 5 domain-splitting cases might be due to gene fission events. Consistent with that view, 2 cases of those, At3g23510 versus Os07g0474400–Os12g0267200 and At3g56330 versus Os05g0324200-100 were inferred as gene fission by outgroup comparison (table 2). Regarding the pair At3g23510 versus Os07g0474400–Os12g0267200, whose alignment is shown in figure 4, it has been reported that a Java olive, *Sterculia foetida* has an intact and functional homolog to At3g23510 encoding cyclopropane fatty acid (CPA-FA)

Table 4
Gene Fusion/Fission Candidate Pairs in Which Splits Probably Occur within Domains

Composite Gene[a]	Split Genes[a]		Matched Entry[b]
Arabidopsis-composite–*Oryza*-split			
At1g33330	Os01g0887400	Os01g0887200	RF-1
At3g23510	**Os07g0474400**	**Os12g0267200**	Amino oxidase
At3g56330	**Os05g0324200**	**Os05g0324100**	TRM
Oryza-composite–*Arabidopsis*-split			
Os02g0708600	At2g23740	At2g23750	Pre-SET
Os09g0566100	At2g36485	At2g36480	DUF618

Note.—TRM, N2, N2-dimethylguanosine tRNA methyltransferase.

[a] The pairs in which gene fissions are inferred by outgroup comparison are shown in bold.

[b] Pfam domains within which gene splits occur.

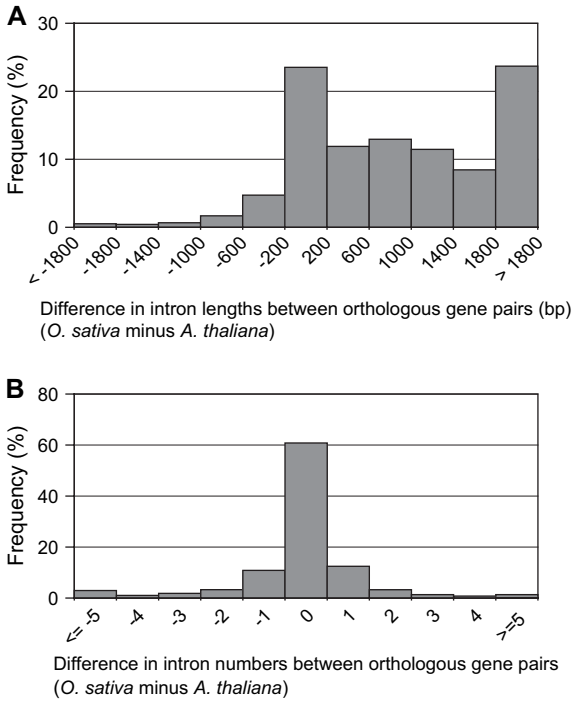

FIG. 3.—Distribution of the difference in intron lengths and numbers between *Oryza sativa* and *Arabidopsis thaliana*. In each of orthologous pairs between *O. sativa* and *A. thaliana*: (*A*) a concatenated length of *A. thaliana* introns is subtracted from the counterpart in *O. sativa*; (*B*) the number of *A. thaliana* introns is subtracted from the counterpart in *O. sativa*.

synthase (Bao et al. 2002, 2003). In that study, the N terminus of these genes was annotated as flavin adenine dinucleotide (FAD) containing oxidase related to "amino oxidase" by Pfam (table 4). Because the significance of FAD-containing oxidase domain of *Arabidopsis* and *Sterculia* composite genes in CPA-FA biosynthesis is poorly understood (Bao et al. 2002, 2003), it may be of interest to investigate the function of Os07g0474400 and Os12g0267200, where the oxidase domain appears to be inactivated.

Newly generated fissions may be deleterious, neutral, or advantageous. But in the latter two cases, they entail the spontaneous origin of novel promoter sequences to afford transcription. These newly arisen promoters in the case of gene fissions may be of interest for further study because they might provide insights into de novo promoter origins. From the comparative standpoint, the maize genome is known to be rich in transposable elements (SanMiguel and Bennetzen 1998) and may thus harbor even more gene fissions than rice. The polarity of gene fusion/fission in *O. sativa* might conceivably relate to rice domestication and breeding, with relaxed constraints during prolonged cultivation, consistent with the richness of transposable elements and the relatively recent occurrence of gene fissions by transposable element insertions in the rice genome (fig. 4).

Previous genome-wide investigations of fusion/fission frequencies have reported that gene fusion may be more common than fission (Snel et al. 2000; Yanai et al. 2001; Suhre and Claverie 2004; Kummerfeld and Teichmann 2005). However, we observe precisely the opposite in the heavily cDNA-supported rice annotations. Previous studies concerned mainly prokaryotic genomes (Snel et al. 2000; Yanai et al. 2001). We emphasize that the frequencies of gene fusion and fission may differ fundamentally for prokaryotic genomes and eukaryotic genomes because there is a much stronger correlation between the functions and locations of genes in prokaryotic genomes—operons (Price et al. 2005)—than in eukaryotic genomes and because translational fusion within operons can involve simple micromutational events, which is not the case in eukaryotes. For example, the *trp* operon has undergone many independent gene fusion and fission events (Xie et al. 2003). In the case of higher plant genomes, the earlier prokaryotic estimates clearly do not apply.

Another earlier investigation of fusion and fission concerned not only prokaryotic but also many eukaryotic genome sequences (Kummerfeld and Teichmann 2005) and reported a 4-fold predominance of gene fusions over fissions. That estimate is inconsistent with our results, where frequent gene fissions have occurred in rice. However, the observations from that earlier study carry 2 caveats. First, there is the possibility of annotation errors, particularly in the genes predicted by the ab initio method in eukaryotic genomes. In that regard, we found that the gene structures of more than 40% of the preliminary fusion/fission candidates are equivocal by database comparison and noncoding region check; they likely represent false positives, and hence we excluded them from our analysis, unlike the previous study (Kummerfeld and Teichmann 2005). Second, the earlier quantitative estimation of fusion and fission rates was

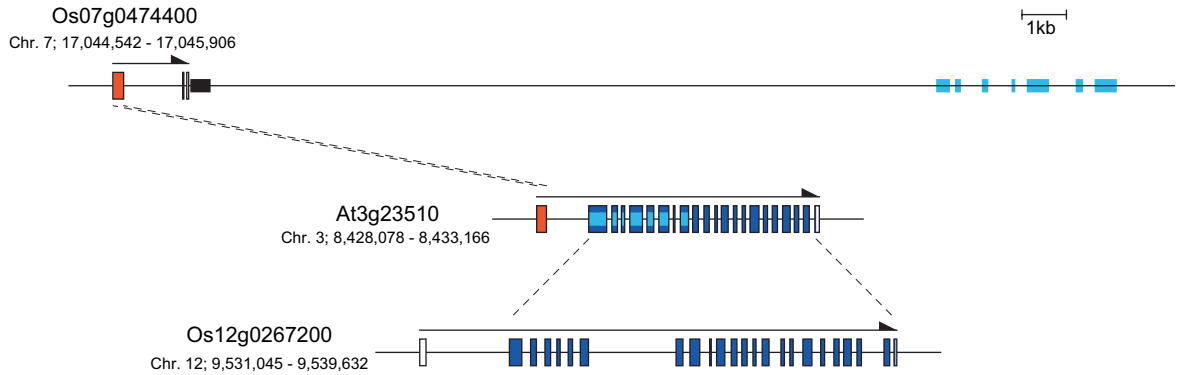

FIG. 4.—Transposon-mediated gene fission. Denotation of figure is the same as in figure 1. Exon-like regions matched to composite genes by TBlastN and repetitive sequences are shown as half-size boxes in cyan and black, respectively.

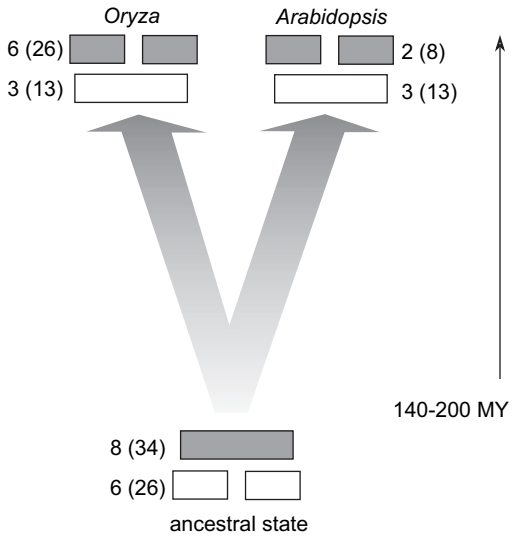

FIG. 5.—Summary of polarized gene fusions and fissions in rice and *Arabidopsis* genomes. Numbers indicate the number of observations. Numbers in parentheses indicate the extrapolation from observed polarized cases (14) to the whole (60).

contingent upon a particular phylogenetic tree linking all genomes considered. If either fusion or fission events had occurred anciently, the ancestral state so inferred will be heavily topology dependent. Furthermore, if any of the composite or split genes were subject to lateral gene transfer among prokaryotes, which does exist (Nakamura et al. 2004; Kunin et al. 2005) and which can also include transfer of operons (Lawrence 1997) and might bear upon the variability of operon structures (Itoh et al. 1999), the rates inferred will also be heavily affected. In particular, the earlier study (Kummerfeld and Teichmann 2005) treated the occurrence of fusion and fission on a much longer timescale (prokaryotes–eukaryotes) as compared with our study (monocot–dicot). Thus, the influence of a guide topology and horizontal gene transfer, as well as the frequency of gene fusions/fissions in operons, will be much larger in the more ancient comparison. In this study, we focused on the events after the divergence of a monocot and dicot and used relatively close outgroups like *C. merolae* and *C. reinhardtii*, or closer where available. The phylogenetic relationships in this estimation are therefore clear and the polarity rather certain, given the rare nature of fusion and fission events in general.

In particular, the previous study estimated that about 30% of the genes examined have undergone multiple fusions or fissions (Kummerfeld and Teichmann 2005), but that might not be a good estimate due to the aforementioned reasons (frequent gene fusions/fissions in operons, annotation errors, horizontal gene transfer, and also of operons). Also, the previous estimate might include lineage-specific amplified genes, many of which may be subject to frequent structural changes by mutation and affect the estimate of gene fusion and fission events. Here it should be noted that we defined gene fusion/fission candidates from one-to-one orthologous pairs between rice and *Arabidopsis*. Our results thus present an estimate on a conserved gene set that is unaffected by lineage-specific gene gain by duplication, suggesting that our estimate is comparable to the ones in other species pairs and applicable to the extrapolation of gene fusion and fission events in number (Enright and Ouzounis 2001). In general, gene fusion or fission events may be very rare among conserved genes (Conant and Wagner 2005).

The presence or absence of a gene fusion or fission itself can, in principle, be useful for investigating the phylogenetic relationships among taxa (Enright and Ouzounis 2001; Stechmann and Cavalier-Smith 2002). Because only $\sim$1% of orthologous gene pairs in the present genome comparison showed differential fusion or fission and because the divergence time of monocots and dicots is roughly 140–200 Myr (Wolfe et al. 1989; Chaw et al. 2004), the possibility of multiple fusions or fissions in each gene can virtually be neglected at this timescale. Treating each of the orthologous gene pairs examined as a gene "site" in computation, the average rate of fusion and fission events is approximately 1×10^{-11} to 2×10^{-11} per gene per year, $\sim$100-fold slower than the average rate of nucleotide substitution ($\sim 5 \times 10^{-9}$ per nucleotide site per year). If we take the 54 unverified cases into account, the rate increases to 3×10^{-11} to 4×10^{-11} per gene per year. If we assume that an average gene has about 1,000-nt sites, it is clear that gene fusions and fissions in these 2 angiosperms occur roughly 10^5 times more slowly than nucleotide substitutions do.

With this slow rate, gene fusion and fission data should provide a means to address deeper evolutionary relationships among plants or other eukaryotes, where the information contained in sequence-based phylogenies is equivocal. As a prominent example, it was reported that dihydrofolate reductase and thymidylate synthase are encoded as a composite gene in protists and plants and as 2 split genes in fungi and metazoa, indicating a lineage-specific distribution (Stechmann and Cavalier-Smith 2002). In our present study, 46 out of 60 candidates remain to be resolved regarding polarity. But we can extrapolate the numbers of gene fusions and fissions and estimate the total number of events during the evolution of *O. sativa* and *A. thaliana* (fig. 5). Although domestication might have affected the rate of fusion and fission events in *O. sativa*, the complete set of fusions and fissions for this pairwise genome comparison nonetheless provides a first benchmark for the plant rate. Determining the state of fusion or fission of the gene pairs identified here in the suspectedly basal angiosperm *Amborella*, for example, where a raging debate exists regarding its evolutionary position because large sequence data sets give conflicting results with strong support (Goremykin et al. 2004; Lockhart and Penny 2005), may shed further light on this and other currently difficult phylogenetic issues.

Supplementary Material

Supplementary figures 1–3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Tal Dagan for helpful comments. This study was supported by a grant from the Japan Society for the Promotion of Science.

Funding to pay the Open Access publication charges for this article was provided by Oxford Journals for the editor in chief.

Literature Cited

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature. 408:796–815.

Bao X, Katz S, Pollard M, Ohlrogge J. 2002. Carbocyclic fatty acids in plants: biochemical and molecular genetic characterization of cyclopropane fatty acid synthesis of *Sterculia foetida*. Proc Natl Acad Sci USA. 99:7172–7177.

Bao X, Thelen JJ, Bonaventure G, Ohlrogge JB. 2003. Characterization of cyclopropane fatty-acid synthase from *Sterculia foetida*. J Biol Chem. 278:12846–12853.

Chaw SM, Chang CC, Chen HL, Li W-H. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol. 58:424–441.

Conant GC, Wagner A. 2005. The rarity of gene shuffling in conserved genes. Genome Biol. 6:R50.

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature. 402:86–90.

Enright AJ, Ouzounis CA. 2001. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol. 2:RESEARCH0034.

Finn RD, Mistry J, Schuster-Böckler B, et al. (13 co-authors). 2006. Pfam: clans, web tools and services. Nucleic Acids Res. 34:D247–D251.

Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH. 2004. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol Biol Evol. 22:1813–1822.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. Nature. 436:793–800.

Itoh T, Takemoto K, Mori H, Gojobori T. 1999. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol Biol Evol. 16: 332–346.

Kummerfeld SK, Teichmann SA. 2005. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet. 21:25–30.

Kunin V, Goldovsky L, Darzentas N, Ouzounis CA. 2005. The net of life: reconstructing the microbial phylogenetic network. Genome Res. 15:954–959.

Lawrence JG. 1997. Selfish operons and speciation by gene transfer. Trends Microbiol. 5:355–359.

Lockhart PJ, Penny D. 2005. The place of *Amborella* within the radiation of angiosperms. Trends Plant Sci. 10:201–202.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science. 285:751–753.

Matsuzaki M, Misumi O, Shin IT, et al. (42 co-authors). 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature. 428:653–657.

Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet. 36:760–766.

Ohyanagi H, Tanaka T, Sakai H, et al. (14 co-authors). 2006. The Rice Annotation Project Database (RAP-DB): hub for *Oryza sativa* ssp. *japonica* genome information. Nucleic Acids Res. 34:D741–D744.

Price MN, Huang KH, Arkin AP, Alm EJ. 2005. Operon formation is driven by co-regulation and not by horizontal gene transfer. Genome Res. 15:809–819.

SanMiguel P, Bennetzen JL. 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot. 82:37–44.

Snel B, Bork P, Huynen M. 2000. Genome evolution. Gene fusion versus gene fission. Trends Genet. 16:9–11.

Stechmann A, Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. Science. 297:89–91.

Suhre K, Claverie JM. 2004. FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. Nucleic Acids Res. 32:D273–D276.

Wolfe KH, Gouy M, Yang YW, Sharp PM, Li W-H. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA. 86:6201–6205.

Xie G, Keyhani NO, Bonner CA, Jenesen RA. 2003. Ancient origin of the tryptophan operon and the dynamics of evolutionary change. Microbiol Mol Biol Rev. 67:303–342.

Yanai I, Derti A, DeLisi C. 2001. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc Natl Acad Sci USA. 98:7940–7945.

Zdobnov EM, Apweiler R. 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 17:847–848.

Manolo Gouy, Associate Editor

Accepted September 25, 2006