Genomics 99 (2012) 308–314

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Retrotransposon insertions in rice gene pairs associated with reduced conservation of gene pairs in grass genomes

Nicholas Krom [1], Wusirika Ramakrishna [*]

Department of Biological Sciences, Michigan Technological University, Houghton, MI 49931, USA

ARTICLE INFO

Article history:
Received 13 January 2012
Accepted 24 February 2012
Available online 6 March 2012

Keywords:
Gene pairs
Retrotransposons
Conservation
Rearrangements
Grass genomes

ABSTRACT

Small-scale changes in gene order and orientation are common in plant genomes, even across relatively short evolutionary distances. We investigated the association of retrotransposons in and near rice gene pairs with gene pair conservation, inversion, rearrangement, and deletion in sorghum, maize, and *Brachypodium*. *Copia* and *Gypsy* LTR-retrotransposon insertions were found to be primarily associated with reduced frequency of gene pair conservation and an increase in both gene pair rearrangement and gene deletions. SINEs are associated with gene pair rearrangement, while LINEs are associated with gene deletions. Despite being more frequently associated with retrotransposons than convergent and tandem pairs, divergent gene pairs showed the least effects from that association. In contrast, convergent pairs were least frequently associated with retrotransposons yet showed the greatest effects. Insertions between genes were associated with the greatest effects on gene pair arrangement, while insertions flanking gene pairs had significant effects only on divergent pairs.

Published by Elsevier Inc.

1. Introduction

The theme often discussed in the field of plant comparative genomics is the tremendous amount of variation in genome size, gene order, and retrotransposon content among plant genomes. The variation among grass genomes is caused by a wide range of mechanisms, including gene and genome duplication, gene deletion, transposable element amplification, transposon mediated gene movement, polyploidization, and various types of recombination [1]. The combined action of these mechanisms can result in an astonishing degree of polymorphism in orthologous regions of closely related species [2]. For instance, analysis of the *Adh1* region in nine species within the genus *Oryza* identified deletions and duplications of genes and gene clusters, highly variable retrotransposon content, and segmental inversions and deletions [3]. While there are many different forces that produce such changes in genome content, retrotransposons are one of the most influential, both through direct means, such as transposition into a new genomic locus, and through the various processes they promote, such as chromosome breakage. Differential retrotransposon activity between species is one of the primary contributors to the wide range of genome sizes observed among the grasses [4], with rapid expansion of genome size occurring during bursts of amplification which are typically followed by rapid loss of retrotransposon sequence [5]. Much of this sequence loss is believed to occur through unequal homologous recombination or illegitimate recombination, which can delete sequence from the host genome in addition to retrotransposons [6–8].

Gene pairs can be either convergent (→ ←), divergent (← →), or tandem (→ → or ← ←) based on the orientation of the adjacent genes. Functional interaction between closely spaced neighboring genes has been proposed as a novel mechanism of gene regulation through correlated expression in eukaryotes including plants [9–12]. A comparative analysis of convergent and divergent gene pairs in rice, *Arabidopsis,* and *Populus trichocarpa* found that the arrangement of these gene pairs is conserved significantly more frequently when the paired genes displayed strongly correlated expression levels, and thus the genes' regulation may be dependent on maintaining a specific relative arrangement [11]. Further, we identified frequent rearrangements in rice gene pairs in sorghum, maize, and *Brachypodium*, where coexpressed rice gene pairs showed higher conservation rates than non-coexpressed pairs [13]. We have previously identified retrotransposon insertions inside or within 1-kb upstream of one-sixth of all rice genes with implications on gene regulation [14]. Retrotransposon insertions in 5′-upstream regions could serve as a source of novel promoters as shown in human gene pairs [15]. Due to the importance of gene pair order and orientation for their function and regulation, and the role of retrotransposons in creating and promoting gene rearrangements, we investigated the correlation between the presence of retrotransposons within gene pairs and the frequency of gene pair conservation and rearrangement. The results

[*] Corresponding author. Fax: +1 906 487 3167.
E-mail address: wusirika@mtu.edu (W. Ramakrishna).

[1] Current address: The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA.

0888-7543/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.ygeno.2012.02.006

from this study support our hypothesis that retrotransposons promote several types of small-scale genomic rearrangements.

2. Results

2.1. Uneven distribution of retrotransposons in rice gene pairs

Analysis of retrotransposons closely associated with the rice gene pairs showed low preference for insertions in convergent gene pairs, with 8.2% of pairs being flanked by a retrotransposon, 11% of pairs having retrotransposons within one or both genes, and 13.6% having insertions between their genes (Fig. 1A). In contrast, retrotransposons were identified in the genes of only 11.4% of divergent pairs, but were found to flank 41.1% of such pairs and in the intergenic regions of 31.6% pairs. The most common positions for retrotransposons in and near tandem pairs also differed greatly, with flanking insertions being least common (6.7% of pairs) and intergenic insertions being most common by a significant margin (26.3% of pairs). The three pair types showed very similar proportion of retrotransposon insertions within genes (11%–11.6%). Pronounced variation was observed among intergenic insertions (13.6% of convergent pairs to 31.6% of divergent pairs) and flanking insertions (6.7% of tandem pairs to 41.1% of divergent pairs).

Significant differences were observed between the four types of retrotransposons in gene pairs. SINE insertions in genes and intergenic regions were the most common with >2-fold difference in frequency compared to *Copia* insertions, which were the least common. *Copia* insertions in flanking and intergenic regions were >2-fold higher than within genes that are part of tandem and divergent but not convergent pairs. *Gypsy* insertions were significantly higher in intergenic regions than in both genes and flanking regions of convergent and tandem pairs. However, *Gypsy* insertions were higher in flanking regions compared to intergenic regions and genes of divergent pairs. SINEs showed similar insertion patterns as *Gypsy* elements. In case of LINEs, insertions were higher in intergenic regions of tandem pairs while flanking regions of divergent pairs showed higher number of insertions than genes and intergenic regions.

2.2. Retrotransposons within, between and flanking rice gene pairs decrease conservation and enhance deletion and rearrangement

Gene pairs that were found to contain retrotransposon insertions were compared as a group with the complete set of gene pairs of that type (convergent, divergent, or tandem) to identify any significant differences in the frequency of gene pair conservation, inversion, rearrangement, or gene deletion in three other grass genomes. Rice gene pairs with retrotransposons in genes or in intergenic regions are less likely to have their orientation conserved in other species, which is statistically significant ($P<0.01$) in 11 out of 18 comparisons (Tables 1 and 2). For instance, 42.4% of all rice convergent pairs are conserved in sorghum, while only 28.5% of convergent pairs with retrotransposons in their intergenic regions are conserved (Table 1). Rice divergent gene pairs with retrotransposons in flanking regions

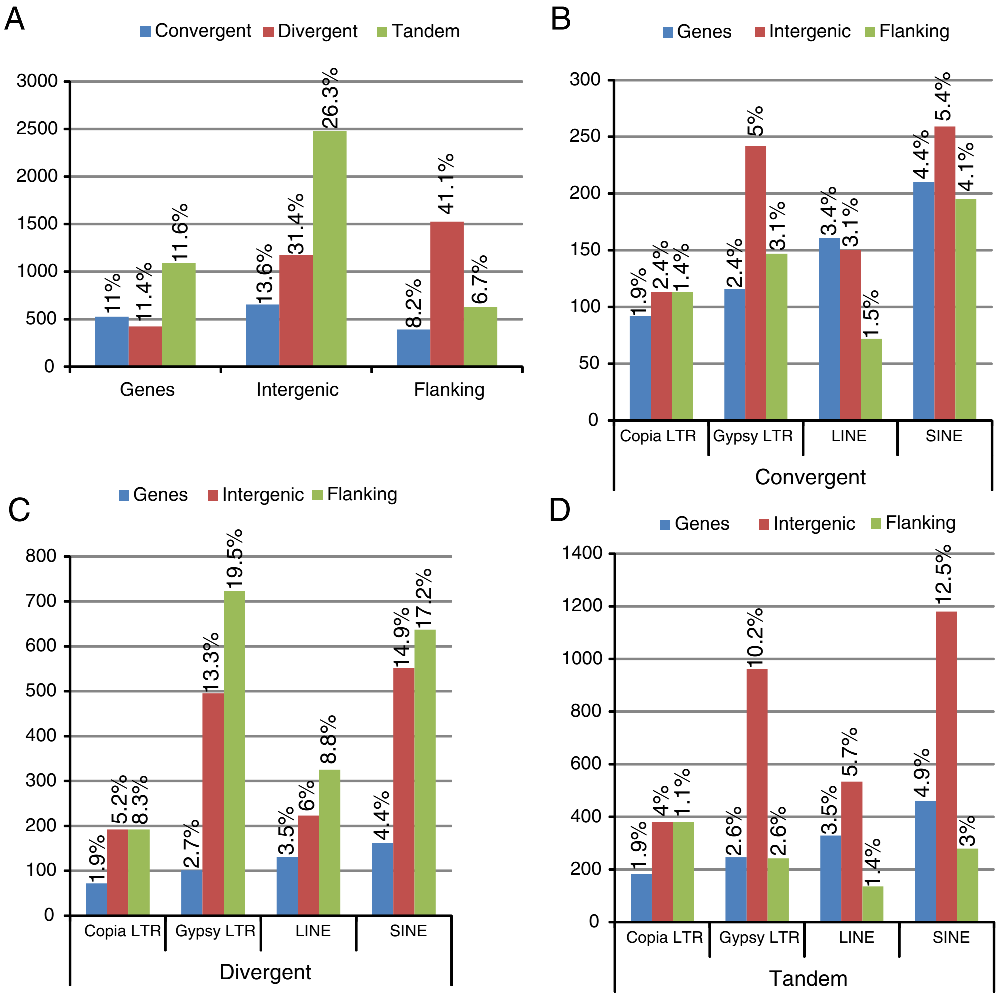

Fig. 1. Distribution of retrotransposon insertions within, between and flanking rice genes that are part of convergent, divergent, and tandem pairs. A. Number of all retrotransposon insertions; number of different types of retrotransposon insertions in or near B. convergent, C. divergent, and D. tandem gene pairs.

Table 1
Conservation and rearrangement of gene pairs with retrotransposon insertions in genes.

		Total pairs	Conserved			Inverted			Rearranged			Missing homologs		
			#	%	*Z*	#	%	*Z*	#	%	*Z*	#	%	*Z*
Rice vs. sorghum														
Convergent	All pairs	4800	2036	42.4%		54	1.1%		2021	42.1%		689	14.4%	
	All retros	526	150	28.5%	**−3.77**	6	1.1%	0.00	240	45.6%	1.10	120	22.8%	2.21
	Copia	92	22	23.9%	−2.03	1	1.1%	0.00	44	47.8%	0.76	25	27.2%	1.44
	Gypsy	116	22	19.0%	**−2.81**	3	2.6%	0.16	56	48.3%	0.92	35	30.2%	2.04
	LINE	161	53	32.9%	−1.47	1	0.6%	−0.06	66	41.0%	−0.18	41	25.5%	1.63
	SINE	210	67	31.9%	−1.85	2	1.0%	−0.03	105	50.0%	1.62	36	17.1%	0.44
Divergent	All pairs	3711	1100	29.6%		35	0.9%		1971	53.1%		605	16.3%	
	All retros	423	86	20.3%	−2.15	4	0.9%	0.00	236	55.8%	0.83	92	21.7%	1.27
	Copia	72	11	15.3%	−1.32	0	0.0%	–	42	58.3%	0.69	19	26.4%	1.00
	Gypsy	101	17	16.8%	−1.41	1	1.0%	0.00	51	50.5%	−0.37	32	31.7%	1.87
	LINE	131	27	20.6%	−1.16	2	1.5%	0.07	69	52.7%	−0.07	33	25.2%	1.18
	SINE	162	41	25.3%	−0.64	1	0.6%	−0.04	96	59.3%	1.23	24	14.8%	−0.21
Tandem	All pairs	9428	3503	37.2%		52	0.6%		4359	46.2%		1514	16.1%	
	All retros	1089	323	29.7%	**−2.95**	8	0.7%	0.06	510	46.8%	0.27	225	20.7%	1.71
	Copia	183	51	27.9%	−1.48	0	0.0%	–	90	49.2%	0.56	42	23.0%	1.06
	Gypsy	246	63	25.6%	−2.10	3	1.2%	0.11	119	48.4%	0.47	61	24.8%	1.58
	LINE	329	101	30.7%	−1.41	3	0.9%	0.07	160	48.6%	0.61	65	19.8%	0.75
	SINE	461	147	31.9%	−1.37	3	0.7%	0.02	222	48.2%	0.57	89	19.3%	0.78
Rice vs. maize														
Convergent	All pairs	4800	1369	28.5%		58	1.2%		2381	49.6%		992	20.7%	
	All retros	526	91	17.3%	**−2.83**	7	1.3%	0.03	257	48.9%	−0.24	161	30.6%	**2.74**
	Copia	92	14	15.2%	−1.39	0	0.0%	–	45	48.9%	−0.09	33	35.9%	1.82
	Gypsy	116	13	11.2%	−1.98	2	1.7%	0.06	57	49.1%	−0.07	44	37.9%	**2.36**
	LINE	161	38	23.6%	−0.71	2	1.2%	0.00	72	44.7%	−0.83	49	30.4%	1.49
	SINE	210	34	16.2%	−1.95	4	1.9%	0.10	117	55.7%	1.33	55	26.2%	0.93
Divergent	All pairs	3711	441	11.9%		74	2.0%		2350	63.3%		846	22.8%	
	All retros	423	36	8.5%	−0.73	10	2.4%	0.08	246	58.2%	−1.64	126	29.8%	1.72
	Copia	72	6	8.3%	−0.31	4	5.6%	0.31	42	58.3%	−0.66	20	27.8%	0.50
	Gypsy	101	8	7.9%	−0.42	1	1.0%	−0.10	51	50.5%	−1.83	41	40.6%	2.32
	LINE	131	9	6.9%	−0.59	3	2.3%	0.03	73	55.7%	−1.31	46	35.1%	1.75
	SINE	162	18	11.1%	−0.10	3	1.9%	−0.02	103	63.6%	0.05	38	23.5%	0.10
Tandem	All pairs	9428	2132	22.6%		106	1.1%		4957	52.6%		2233	23.7%	
	All retros	1089	215	19.7%	−1.06	11	1.0%	−0.04	533	48.9%	−1.68	307	28.2%	1.75
	Copia	183	30	16.4%	−0.92	1	0.5%	−0.08	89	48.6%	−0.74	63	34.4%	1.79
	Gypsy	246	42	17.1%	−0.95	2	0.8%	−0.05	120	48.8%	−0.83	82	33.3%	1.85
	LINE	329	71	21.6%	−0.21	3	0.9%	−0.04	167	50.8%	−0.47	88	26.7%	0.65
	SINE	461	102	22.1%	−0.12	6	1.3%	0.04	231	50.1%	−0.75	122	26.5%	0.70
Rice vs. *Brachypodium*														
Convergent	All pairs	4800	2014	42.0%		73	1.5%		1802	37.5%		911	19.0%	
	All retros	526	150	28.5%	**−3.65**	7	1.3%	−0.04	212	40.3%	0.82	147	27.9%	**2.42**
	Copia	92	26	28.3%	−1.55	0	0.0%	–	39	42.4%	0.61	27	29.3%	1.18
	Gypsy	116	18	15.5%	**−3.10**	0	0.0%	–	55	47.4%	1.47	43	37.1%	**2.46**
	LINE	161	55	34.2%	−1.22	3	1.9%	0.04	60	37.3%	−0.04	43	26.7%	1.15
	SINE	210	65	31.0%	−1.92	4	1.9%	0.06	88	41.9%	0.83	53	25.2%	1.05
Divergent	All pairs	3711	1220	32.9%		58	1.6%		1618	43.6%		815	22.0%	
	All retros	423	105	24.8%	−1.91	4	0.9%	−0.13	181	42.8%	−0.22	128	30.3%	2.04
	Copia	72	21	29.2%	−0.37	1	1.4%	−0.01	31	43.1%	−0.06	19	26.4%	0.44
	Gypsy	101	22	21.8%	−1.26	1	1.0%	−0.06	36	35.6%	−1.00	42	41.6%	**2.58**
	LINE	131	30	22.9%	−1.30	1	0.8%	−0.09	55	42.0%	−0.24	45	34.4%	1.75
	SINE	162	44	27.2%	−0.85	1	0.6%	−0.12	78	48.1%	0.80	39	24.1%	0.31
Tandem	All pairs	9428	3483	36.9%		100	1.1%		3964	42.0%		1881	20.0%	
	All retros	1089	294	27.0%	**−3.84**	10	0.9%	−0.05	483	44.4%	1.02	279	25.6%	2.17
	Copia	183	41	22.4%	−2.23	2	1.1%	0.00	87	47.5%	1.03	53	29.0%	1.45
	Gypsy	246	55	22.4%	**−2.60**	3	1.2%	0.03	115	46.7%	1.01	73	29.7%	1.82
	LINE	329	96	29.2%	−1.67	1	0.3%	−0.14	147	44.7%	0.64	85	25.8%	1.24
	SINE	461	128	27.8%	−2.32	4	0.9%	−0.04	219	47.5%	1.62	110	23.9%	0.96

Numbers in the columns labelled *Z* are test statistics from the binomial test, comparing the fraction of the various types of retrotransposon-associated gene pairs in each conservation/rearrangement class with the fraction of all gene pairs in the same class. Bold numbers denote a statistically significant difference ($P<0.01$).

are less likely to have their orientation conserved in sorghum and *Brachypodium* (Table 3). Similarly, retrotransposon association makes gene pairs more likely to be rearranged, with both genes conserved but no longer physically closer to each other. The frequencies with which gene pairs are found to be missing homologs in other species also correlate with increased presence of retrotransposons. Although they are associated with both gene pair rearrangement and gene deletion, the correlation with deletion of genes appears to be higher based on the statistical significance of the comparisons. Retrotransposons do not have significant effect on the likelihood of one or both genes in a pair to be inverted. Furthermore, convergent pairs were most frequently disrupted by retrotransposon insertions in and between genes, displaying the largest decrease in conservation and increase in rearrangements and gene deletions. Divergent pairs were least affected by retrotransposon insertions in and between genes, but most affected by insertions flanking the gene pair.

2.3. Gypsy LTR-retrotransposon insertions have the most effect on gene pair conservation

Analysis of the effect of various types of retrotransposons on gene pair conservation and rearrangements identified that *Copia* LTR-

Table 2
Conservation and rearrangement of gene pairs with retrotransposon insertions between genes.

		Total pairs	Conserved			Inverted			Rearranged			Missing homologs		
			#	%	*Z*	#	%	*Z*	#	%	*Z*	#	%	*Z*
Rice vs. sorghum														
Convergent	All pairs	4800	2036	42.4%		54	1.1%		2021	42.1%		689	14.4%	
	All retros	655	135	20.6%	**−6.26**	10	1.5%	0.10	337	51.5%	**3.43**	143	21.8%	2.16
	Copia	113	26	23.0%	**−2.35**	1	0.9%	−0.03	55	48.7%	0.97	31	27.4%	1.63
	Gypsy	242	30	12.4%	**−4.99**	5	2.1%	0.15	134	55.4%	**3.09**	73	30.2%	**2.94**
	LINE	150	34	22.7%	**−2.75**	1	0.7%	−0.06	78	52.0%	1.75	37	24.7%	1.46
	SINE	259	71	27.4%	**−2.83**	7	2.7%	0.26	137	52.9%	**2.53**	44	17.0%	0.47
Divergent	All pairs	3711	1100	29.6%		35	0.9%		1971	53.1%		605	16.3%	
	All retros	1174	290	24.7%	−1.95	14	1.2%	0.09	624	53.2%	0.02	199	17.0%	0.24
	Copia	192	35	18.2%	−1.75	1	0.5%	−0.06	109	56.8%	0.77	47	24.5%	1.30
	Gypsy	495	107	21.6%	−2.02	5	1.0%	0.01	282	57.0%	1.31	101	20.4%	1.02
	LINE	223	50	22.4%	−1.22	2	0.9%	−0.01	129	57.8%	1.09	42	18.8%	0.42
	SINE	552	153	27.7%	−0.53	9	1.6%	0.16	311	56.3%	1.15	79	14.3%	−0.51
Tandem	All pairs	9428	3503	37.2%		52	0.6%		4359	46.2%		1514	16.1%	
	Any retros	2477	714	28.8%	**−4.91**	10	0.4%	−0.07	1212	48.9%	1.88	470	19.0%	1.61
	Copia	380	92	24.2%	**−2.90**	2	0.5%	0.00	204	53.7%	2.13	82	21.6%	1.22
	Gypsy	961	264	27.5%	**−3.52**	2	0.2%	−0.11	475	49.4%	1.39	220	22.9%	**2.41**
	LINE	534	149	27.9%	**−2.52**	3	0.6%	0.00	271	50.7%	1.49	111	20.8%	1.23
	SINE	1180	376	31.9%	−2.20	4	0.3%	−0.07	597	50.6%	2.13	203	17.2%	0.43
Rice vs. maize														
Convergent	All pairs	4800	1369	28.5%		58	1.2%		2381	49.6%		992	20.7%	
	All retros	655	73	11.1%	**−4.72**	12	1.8%	0.16	347	53.0%	1.26	193	29.5%	**2.68**
	Copia	113	16	14.2%	−1.65	2	1.8%	0.06	57	50.4%	0.13	38	33.6%	1.69
	Gypsy	242	16	6.6%	**−3.53**	4	1.7%	0.07	129	53.3%	0.84	93	38.4%	**3.52**
	LINE	150	24	16.0%	−1.67	5	3.3%	0.26	76	50.7%	0.19	45	30.0%	1.37
	SINE	259	30	11.6%	**−2.90**	7	2.7%	0.24	155	59.8%	**2.60**	67	25.9%	0.97
Divergent	All pairs	3711	441	11.9%		74	2.0%		2350	63.3%		846	22.8%	
	All retros	1174	96	8.2%	−1.33	30	2.6%	0.19	731	62.3%	−0.59	270	23.0%	0.08
	Copia	192	10	5.2%	−0.95	6	3.1%	0.16	116	60.4%	−0.64	60	31.3%	1.41
	Gypsy	495	36	7.3%	−1.07	10	2.0%	0.01	318	64.2%	0.34	131	26.5%	0.95
	LINE	223	16	7.2%	−0.73	7	3.1%	0.17	144	64.6%	0.31	56	25.1%	0.40
	SINE	552	46	8.3%	−0.87	15	2.7%	0.17	380	68.8%	2.32	111	20.1%	−0.71
Tandem	All pairs	9428	2132	22.6%		106	1.1%		4957	52.6%		2233	23.7%	
	All retros	2477	466	18.8%	−2.10	25	1.0%	−0.06	1261	50.9%	−1.19	654	26.4%	1.58
	Copia	380	71	18.7%	−0.85	2	0.5%	−0.12	180	47.4%	−1.40	127	33.4%	2.33
	Gypsy	961	180	18.7%	−1.34	9	0.9%	−0.06	469	48.8%	−1.64	303	31.5%	**2.94**
	LINE	534	93	17.4%	−1.32	8	1.5%	0.09	285	53.4%	0.27	148	27.7%	1.10
	SINE	1180	230	19.5%	−1.20	11	0.9%	−0.07	654	55.4%	1.46	285	24.2%	0.18
Rice vs. *Brachypodium*														
Convergent	All pairs	4800	2014	42.0%		73	1.5%		1802	37.5%		911	19.0%	
	All retros	655	133	20.3%	**−6.21**	10	1.5%	0.00	295	45.0%	**2.59**	187	28.5%	**2.90**
	Copia	113	21	18.6%	**−2.75**	0	0.0%	–	54	47.8%	1.51	38	33.6%	1.91
	Gypsy	242	36	14.9%	**−4.57**	4	1.7%	0.02	108	44.6%	1.48	94	38.8%	**3.95**
	LINE	150	28	18.7%	**−3.16**	2	1.3%	−0.02	76	50.7%	2.29	44	29.3%	1.51
	SINE	259	63	24.3%	**−3.26**	6	2.3%	0.13	127	49.0%	**2.59**	63	24.3%	0.99
Divergent	All pairs	3711	1220	32.9%		58	1.6%		1618	43.6%		815	22.0%	
	All retros	1174	314	26.7%	**−2.45**	19	1.6%	0.02	534	45.5%	0.87	260	22.1%	0.07
	Copia	192	46	24.0%	−1.42	4	2.1%	0.07	87	45.3%	0.32	55	28.6%	1.10
	Gypsy	495	130	26.3%	−1.71	8	1.6%	0.01	239	48.3%	1.45	118	23.8%	0.48
	LINE	223	54	24.2%	−1.49	4	1.8%	0.03	108	48.4%	1.00	57	25.6%	0.62
	SINE	552	162	29.3%	−0.99	9	1.6%	0.02	266	48.2%	1.50	115	20.8%	−0.30
Tandem	All pairs	9428	3483	36.9%		100	1.1%		3964	42.0%		1881	20.0%	
	All retros	2477	702	28.3%	**−5.06**	21	0.8%	−0.11	1091	44.0%	1.33	592	23.9%	2.25
	Copia	380	80	21.1%	**−3.49**	1	0.3%	−0.16	180	47.4%	1.43	119	31.3%	**2.67**
	Gypsy	961	259	27.0%	**−3.62**	4	0.4%	−0.20	419	43.6%	0.64	279	29.0%	**3.34**
	LINE	534	151	28.3%	**−2.36**	4	0.7%	−0.07	248	46.4%	1.39	131	24.5%	1.22
	SINE	1180	362	30.7%	**−2.58**	18	1.5%	0.16	553	46.9%	2.27	247	20.9%	0.38

Numbers in the columns labelled *Z* are test statistics from the binomial test, comparing the fraction of the various types of retrotransposon-associated gene pairs in each conservation/rearrangement class with the fraction of all gene pairs in the same class. Bold numbers denote a statistically significant difference ($P<0.01$).

retrotransposon insertions between rice genes in convergent and tandem pairs were associated with significant reduction in gene pair conservation in sorghum and *Brachypodium* (Table 2). It is likely that gene pair arrangement was disrupted through the loss of homologous genes primarily affecting tandem pairs when they insert between rice genes (Table 2) and divergent pairs when they flank them in the lineage leading to *Brachypodium* (Table 3).

The most striking feature in the case of *Gypsy* LTR-retrotransposon insertions in rice genes was the marked decrease in conservation rates of convergent pairs in sorghum, maize and *Brachypodium*, largely due to missing homologs (Table 1). This was also observed in tandem pairs in *Brachypodium*. Insertions in one or both genes in rice make gene pairs more likely to be missing one or both homologs in sorghum, maize, and *Brachypodium* although they are significant ($P<0.01$) in 3 out of 9 comparisons (Table 1). Similarly, rice gene pairs with intergenic *Gypsy* insertions showed significant decrease in conservation of convergent and tandem pairs in the other three grass genomes with significant increase of missing homologs (Table 2). Rearranged convergent pairs were also more common in sorghum. Contrary to the observation described above, rice gene pairs with flanking *Gypsy* insertions showed significant decrease in conservation of only divergent pairs in sorghum and *Brachypodium* accompanied

Table 3
Conservation and arrangement of gene pairs with retrotransposon insertions flanking gene pairs.

		Total pairs	Conserved			Inverted			Rearranged			Missing homologs		
			#	%	*Z*	#	%	*Z*	#	%	*Z*	#	%	*Z*
Rice vs. sorghum														
Convergent	All pairs	4800	2036	42.4%		54	1.1%		2021	42.1%		689	14.4%	
	All retros	393	165	42.0%	−0.11	1	0.3%	−0.17	161	41.0%	−0.29	53	13.5%	−0.19
	Copia	65	22	33.8%	−0.85	0	0.0%	–	27	41.5%	−0.06	16	24.6%	0.95
	Gypsy	147	60	40.8%	−0.25	1	0.7%	−0.05	62	42.2%	0.01	24	16.3%	0.26
	LINE	72	24	33.3%	−0.94	0	0.0%	–	36	50.0%	0.95	12	16.7%	0.21
	SINE	195	96	49.2%	1.34	0	0.0%	–	80	41.0%	−0.20	19	9.7%	−0.68
Divergent	All pairs	3711	1100	29.6%		35	0.9%		1971	53.1%		605	16.3%	
	All retros	1527	339	22.2%	**−3.30**	14	0.9%	−0.01	840	55.0%	1.11	297	19.4%	1.37
	Copia	309	55	17.8%	−2.30	3	1.0%	0.00	177	57.3%	1.12	74	23.9%	1.54
	Gypsy	723	140	19.4%	**−3.08**	8	1.1%	0.04	399	55.2%	0.83	176	24.3%	**2.49**
	LINE	325	72	22.2%	−1.53	2	0.6%	−0.06	179	55.1%	0.53	72	22.2%	1.20
	SINE	637	164	25.7%	−1.14	6	0.9%	0.00	372	58.4%	2.07	95	14.9%	−0.38
Tandem	All pairs	9428	3503	37.2%		52	0.6%		4359	46.2%		1514	16.1%	
	All retros	627	230	36.7%	−0.15	4	0.6%	0.02	278	44.3%	−0.64	103	16.4%	0.10
	Copia	100	31	31.0%	−0.74	0	0.0%	–	48	48.0%	0.24	21	21.0%	0.56
	Gypsy	242	83	34.3%	−0.55	0	0.0%	–	114	47.1%	0.19	45	18.6%	0.44
	LINE	135	49	36.3%	−0.13	1	0.7%	0.02	62	45.9%	−0.05	23	17.0%	0.12
	SINE	279	108	38.7%	0.33	3	1.1%	0.09	125	44.8%	−0.32	43	15.4%	−0.12
Rice vs. maize														
Convergent	All pairs	4800	1369	28.5%		58	1.2%		2381	49.6%		992	20.7%	
	All retros	393	115	29.3%	0.17	3	0.8%	−0.09	197	50.1%	0.15	65	16.5%	−0.90
	Copia	65	18	27.7%	−0.08	1	1.5%	0.03	32	49.2%	−0.04	14	21.5%	0.08
	Gypsy	147	38	25.9%	−0.38	0	0.0%	–	79	53.7%	0.74	30	20.4%	−0.04
	LINE	72	23	31.9%	0.35	1	1.4%	0.02	33	45.8%	−0.43	15	20.8%	0.02
	SINE	195	65	33.3%	0.82	2	1.0%	−0.03	103	52.8%	0.65	25	12.8%	−1.17
Divergent	All pairs	3711	441	11.9%		74	2.0%		2350	63.3%		846	22.8%	
	All retros	1527	147	9.6%	−0.93	29	1.9%	−0.04	914	59.9%	−2.14	400	26.2%	1.55
	Copia	309	20	6.5%	−0.98	6	1.9%	−0.01	180	58.3%	−1.38	103	33.3%	2.27
	Gypsy	723	75	10.4%	−0.43	13	1.8%	−0.05	404	55.9%	**−3.01**	231	32.0%	**2.98**
	LINE	325	37	11.4%	−0.10	3	0.9%	−0.19	203	62.5%	−0.25	82	25.2%	0.51
	SINE	637	64	10.0%	−0.49	19	3.0%	0.25	416	65.3%	0.85	138	21.7%	−0.32
Tandem	All pairs	9428	2132	22.6%		106	1.1%		4957	52.6%		2233	23.7%	
	All retros	627	136	21.7%	−0.26	10	1.6%	0.12	327	52.2%	−0.15	142	22.6%	−0.30
	Copia	100	20	20.0%	−0.29	1	1.0%	−0.01	53	53.0%	0.06	26	26.0%	0.27
	Gypsy	242	59	24.4%	0.32	4	1.7%	0.08	123	50.8%	−0.39	56	23.1%	−0.10
	LINE	135	34	25.2%	0.35	3	2.2%	0.13	68	50.4%	−0.36	30	22.2%	−0.19
	SINE	279	56	20.1%	−0.47	5	1.8%	0.11	150	53.8%	0.29	68	24.4%	0.13
Rice vs. *Brachypodium*														
Convergent	All pairs	4800	2014	42.0%		73	1.5%		1802	37.5%		911	19.0%	
	All retros	393	160	40.7%	−0.32	9	2.3%	0.15	142	36.1%	−0.35	69	17.6%	−0.31
	Copia	65	27	41.5%	−0.04	1	1.5%	0.00	22	33.8%	−0.37	15	23.1%	0.38
	Gypsy	147	49	33.3%	−1.28	4	2.7%	0.15	60	40.8%	0.52	34	23.1%	0.57
	LINE	72	26	36.1%	−0.62	1	1.4%	−0.01	28	38.9%	0.15	17	23.6%	0.45
	SINE	195	97	49.7%	1.53	5	2.6%	0.15	66	33.8%	−0.63	27	13.8%	−0.77
Divergent	All pairs	3711	1220	32.9%		58	1.6%		1618	43.6%		815	22.0%	
	All retros	1527	400	26.2%	**−3.04**	22	1.4%	−0.05	669	43.8%	0.11	399	26.1%	1.89
	Copia	309	68	22.0%	−2.16	6	1.9%	0.07	132	42.7%	−0.20	103	33.3%	**2.45**
	Gypsy	723	167	23.1%	**−3.00**	15	2.1%	0.14	306	42.3%	−0.45	235	32.5%	**3.45**
	LINE	325	90	27.7%	−1.10	3	0.9%	−0.12	150	46.2%	0.63	82	25.2%	0.68
	SINE	637	193	30.3%	−0.78	5	0.8%	−0.20	294	46.2%	0.88	145	22.8%	0.23
Tandem	All pairs	9428	3483	36.9%		100	1.1%		3964	42.0%		1881	20.0%	
	All retros	627	224	35.7%	−0.38	5	0.8%	−0.07	257	41.0%	−0.34	129	20.6%	0.18
	Copia	100	31	31.0%	−0.72	1	1.0%	−0.01	43	43.0%	0.13	25	25.0%	0.58
	Gypsy	242	88	36.4%	−0.11	3	1.2%	0.03	98	40.5%	−0.31	53	21.9%	0.34
	LINE	135	50	37.0%	0.01	0	0.0%	–	57	42.2%	0.03	28	20.7%	0.10
	SINE	279	99	35.5%	−0.30	2	0.7%	−0.06	123	44.1%	0.46	55	19.7%	−0.04

Numbers in the columns labelled *Z* are test statistics from the binomial test, comparing the fraction of the various types of retrotransposon-associated gene pairs in each conservation/rearrangement class with the fraction of all gene pairs in the same class. Bold numbers denote a statistically significant difference ($P<0.01$).

with an increase in the frequency of missing homologs (Table 3). Maize showed a decrease in rearranged divergent pairs coupled with an increase in missing homologs. These data suggest that *Gypsy* LTR-retrotransposons are likely to be powerful agents of gene pair disruption compared to other retrotransposons. With some exceptions, their association with rice gene pairs correlates with an increase in the frequency of gene pair rearrangement and homolog deletion.

The frequency of convergent and tandem gene pair conservation in sorghum and *Brachypodium* was significantly reduced when LINEs were located between rice genes (Table 2). However, LINEs are less likely to interfere with conservation than LTR-retrotransposons. In addition, their association with rice gene pairs appears to influence gene pairs in other grass genomes with higher effect on deletion of homologs, and little effect on their rearrangement.

SINE insertions show very similar pattern as LINEs with intergenic insertions showing the greatest effect leading to reduced conservation of convergent and tandem pairs in sorghum and *Brachypodium* (Table 2). Rice gene pairs with SINEs between their genes are more likely to be rearranged rather than deleted in other grass species which is statistically significant among convergent pairs.

3. Discussion

Significant variation and similarities exist among different families of retrotransposons with regard to the insertional preferences as well as their influence on the conservation, rearrangement and deletion of gene pairs. Intergenic retrotransposon insertions which affected gene pair conservation the most were commonly found within divergent pairs (31.6%), followed by tandem pairs (26.3%) and convergent pairs (13.6%). These results may appear counterintuitive if one considers the likelihood of the retrotransposon insertion interfering with the genes' promoters (since both promoters are in the intergenic region of divergent pairs while neither promoter is there in convergent pairs). However, the fraction of pairs with intergenic insertions correlates quite well with the mean intergenic distances of each pair type, which are 4371 bp, 3734 bp, and 2562 bp for divergent, tandem, and convergent pairs, respectively. Thus there appears to be little selective pressure for or against intergenic retrotransposon insertions based on pair type, and insertion frequency may simply be determined by the available space. The variation in flanking insertion frequency cannot be explained by differences in the size of the intergenic regions flanking each pair type, which are more consistent in size, ranging from 3211 bp on average for divergent pairs to 4114 bp for convergent pairs, while insertion frequency varied greatly, from 6.7% of tandem pairs to 41.1% of divergent pairs. The frequency of this type of insertion may be influenced by the possibility of disrupting regulatory elements, as only divergent pairs have no promoters in the pair's flanking region. Furthermore, the enhanced rate of gene deletions in divergent pairs can be explained by the highest percentage retrotransposon insertions flanking them (Fig. 1).

Intergenic insertions of *Gypsy*, *Copia*, LINE and SINE retrotransposons in rice divergent pairs showed least effect on gene pair conservation and rearrangement in other grass genomes compared to significant reduction in gene pair conservation in convergent and tandem pairs, which can be attributed to the presence of bidirectional promoters in divergent pairs. Most of the promoters located between divergent pair of genes separated by <1 kb are expected to be bidirectional promoters [16]. Therefore, only a subset of divergent gene pairs analyzed in this study are likely to harbor bidirectional promoters.

The lower conservation rates of gene pairs observed due to LTR-retrotransposon insertions especially *Gypsy* elements compared to non-LTR elements is likely due to illegitimate recombination and unequal homologous recombination, which created the current state of the gene pairs. Thus a rice divergent pair labeled as "rearranged" in sorghum may in fact have been created by a retrotransposon-mediated rearrangement in the rice genome that brought together two previously non-adjacent genes. Similarly, so-called deleted genes may in fact be a reflection of gene creation in ancestor of rice. Most of LTR-retrotransposon deletion and amplification have probably occurred in rice in the last 5 million years while truncated elements were >10 million years old [7,17,18]. Furthermore, retrotransposon insertions in rice genes tend to be older than those in promoters [19,20]. Illegitimate recombination between elements can result in the deletion of any sequence between the two retrotransposons providing a mechanism for retrotransposon-mediated deletion of genes [6,18]. Another possible mechanism of retrotransposon-related gene pair rearrangement is the repair of double-stranded DNA breakage, which can be induced by the presence of transposable elements [4]. Depending on the repair mechanism used, these breaks can result in the duplication or deletion of sequence near the break, or the insertion of seemingly unrelated genomic sequence at the breakage point [21–23]. Retrotransposon cDNA sequences have also been found to be inserted during such repairs, so in some cases retrotransposon insertions may be the result of double-stranded break repair, rather than a cause [22,24].

LINEs and SINEs differ both from each other and from the LTR-retrotransposons with regard to their correlation with particular events in gene pair conservation and rearrangement. While all four types of insertions reduce the frequency of gene pair conservation, the reductions associated with LINEs and SINEs is mostly limited to insertions between convergent and tandem genes with the effect of LINEs being the weaker of the two. Rice gene pairs with SINE elements are more likely to be rearranged than have missing homologs, while the opposite is true for LINEs. Both LINEs and SINEs have been found to cause several types of genomic rearrangements via recombination, although most of the studies have been in animal genomes. Homologous recombination between LINEs has produced deletions in the human genome [25,26]. Segmental duplications have been attributed to SINE-SINE recombination in the human and mouse genomes [27,28]. LINEs and SINEs use an alternative endonuclease independent pathway for insertions which suggest their involvement in double strand break repair [29].

The effects of retrotransposon insertions within and between genes are both more profound and widespread than those flanking gene pairs. If we assume that recombination between retrotransposons is responsible for the majority of retrotransposon-mediated gene pair alterations, as described above, then it follows that retrotransposon insertions within the gene pair would be associated with more deletions and rearrangements, as the recombined region between the insertion in the pair and the outside retrotransposon would always include all or part of at least one gene.

4. Materials and methods

Rice genome sequence and annotation data were downloaded from http://rice.plantbiology.msu.edu (MSU rice pseudomolecules release 6) while sorghum, maize, and *Brachypodium* data, gene pair identification and comparative analysis were described in Krom and Ramakrishna [13]. Pairs were considered 'rearranged' if both genes of a pair were separated by >50 kb or other genes were inserted between them or they were part of different contigs. Retrotransposon insertions in rice gene pairs study were identified using RepeatMasker (www.repeatmasker.org). For each pair, five sequences were analyzed: the two genes' unspliced genomic sequence, the intergenic region between them, and the two intergenic regions flanking the pair. The evolutionary status (conserved, inverted, rearranged, or deleted) of the pairs containing *Copia* or *Gypsy* LTR-retrotransposons, Long Interspersed Nuclear Elements (LINEs), or Short Interspersed Nuclear Elements (SINEs) within, between, or flanking their genes was determined via cross-reference with the results of our previous study [13].

The normal approximation of the binomial test was used to test the statistical significance of the differences in conservation, inversion, rearrangement, or deletion frequency between the complete sets of gene pairs and the sets of retrotransposon-associated pairs. Differences with a *P*-value less than 0.01 ($|Z|>2.3267$) were considered significant.

RepeatMasker was also used to identify retrotransposon insertions in and around the sorghum, maize, and *Brachypodium* homologs of rice gene pairs conserved or rearranged in those species. For conserved pairs, the sequence analyzed included both of the homologs, their intergenic region, and 2000 bp of flanking sequence upstream and downstream of the pair. For each rearranged pair, two sequences composed of the homologous genes and 2000 bp of flanking sequence were analyzed. The RepeatMasker output was then examined to determine if the homologous sequences contained the same type of retrotransposon insertions originally identified in rice.

References

[1] J.L. Bennetzen, M. Chen, Grass genomic synteny illuminates plant genome function and evolution, Rice 1 (2008) 109–118.

[2] J.L. Bennetzen, Patterns in grass genome evolution, Curr. Opin. Plant Biol. 10 (2007) 176–181.

[3] J.S.S. Ammiraju, F. Lu, A. Sanyal, Y. Yeisoo, X. Song, N. Jiang, et al., Dynamic evolution of *Oryza* genomes is revealed by comparative genomic analysis of a genus-wide vertical data set, Plant Cell 20 (2008) 3191–3209.

[4] J.L. Bennetzen, Transposable elements, gene creation and genome rearrangement in flowering plants, Curr. Opin. Genet. Dev. 15 (2005) 621–627.

[5] C. Vitte, O. Panaud, H. Quesneville, LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss, BMC Genomics 8 (2007) 218.

[6] K.M. Devos, J.K.M. Brown, J.L. Bennetzen, Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*, Genome Res. 12 (2002) 1075–1079.

[7] J. Ma, K.M. Devos, J.L. Bennetzen, Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice, Genome Res. 14 (2004) 860–869.

[8] T. Zhixi, R. Carene, D. Jianchang, L. Zhu, J.L. Bennetzen, S.S. Jackson, et al., Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res. 19 (2009) 2221–2230.

[9] E. Franck, T. Hulsen, M.A. Huynen, W.W. de Jong, N.H. Lubsen, O. Madsen, Evolution of closely linked gene pairs in vertebrate genomes, Mol. Biol. Evol. 25 (2008) 1909–1921.

[10] L.D. Hurst, C. Pal, M.J. Lercher, The evolutionary dynamics of eukaryotic gene order, Nat. Rev. Genet. 5 (2004) 299–310.

[11] N. Krom, W. Ramakrishna, Comparative analysis of divergent and convergent gene pairs and their expression patterns in rice, Arabidopsis, and *Populus*, Plant Physiol. 147 (2008) 1763–1773.

[12] E.J.G. Williams, D.J. Bowles, Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*, Genome Res. 14 (2004) 1060–1067.

[13] N. Krom, W. Ramakrishna, Conservation, rearrangement, and deletion of gene pairs in four grass genomes, DNA Res. 17 (2010) 343–352.

[14] N. Krom, J. Recla, W. Ramakrishna, Analysis of genes associated with retrotransposons in the rice genome, Genetica 134 (2008) 297–310.

[15] K. Okamura, K. Nakai, Retrotransposition as a source of new promoters, Mol. Biol. Evol. 25 (2008) 1231–1238.

[16] S.R. Dhadi, N. Krom, W. Ramakrishna, Genome-wide comparative analysis of putative bidirectional promoters from rice, *Arabidopsis* and *Populus*, Gene 429 (2009) 65–73.

[17] J. Ma, J.L. Bennetzen, Rapid recent growth and divergence of rice nuclear genomes, Proc. Natl. Acad. Sci. 101 (2004) 12404–12410.

[18] C. Vitte, J.L. Bennetzen, Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution, Proc. Natl. Acad. Sci. 103 (2006) 17638–17643.

[19] Z. Xu, W. Ramakrishna, Retrotransposon insertion polymorphisms in six rice genes and their evolutionary history, Gene 412 (2008) 50–58.

[20] Z. Xu, S. Rafi, W. Ramakrishna, Polymorphisms and evolutionary history of retrotransposon insertions in rice promoters, Genome 54 (2011) 629–638.

[21] M. Mitch, L.S. Eun, MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings, Trends Genet. 24 (2008) 529–538.

[22] H. Puchta, The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution, J. Exp. Bot. 56 (2005) 1–14.

[23] S. Salomon, H. Puchta, Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells, EMBO J. 17 (1998) 6086–6095.

[24] J.K. Moore, J.E. Haber, Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks, Nature 383 (1996) 644–646.

[25] C.R. Beck, J.L. Garcia-Perez, R.M. Badge, J.V. Moran, LINE-1 elements in structural variation and disease, Annu. Rev. Genomics Hum. Genet. 12 (2011) 187–215.

[26] B. Burwinkel, M.W. Kilimann, Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease, J. Mol. Biol. 277 (1998) 513–517.

[27] J. Jurka, O. Kohany, A. Pavlicek, V.V. Kapitonov, M.V. Jurka, Duplication, coclustering, and selection of human Alu retrotransposons, Proc. Natl. Acad. Sci. 101 (2003) 1268–1272.

[28] J. Jurka, O. Kohany, A. Pavlicek, V.V. Kapitonov, M.V. Jurka, Clustering, duplication, and chromosomal distribution of mouse SINE retrotransposons, Cytogenet. Genome Res. 110 (2005) 117–123.

[29] D. Srikanta, S.K. Sen, C.T. Huang, E.M. Conlin, R.M. Rhodes, M.A. Batzer, An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair, Genomics 93 (2009) 205–212.