Rice (2010) 3:242–250
DOI 10.1007/s12284-010-9046-7

Assessing the Extent of Substitution Rate Variation of Retrotransposon Long Terminal Repeat Sequences in *Oryza sativa* and *Oryza glaberrima*

Andrea Zuccolo · Aswathy Sebastian · Yeisoo Yu · Scott Jackson · Steve Rounsley · Dean Billheimer · Rod A. Wing

Received: 26 February 2010 / Accepted: 8 July 2010 / Published online: 31 July 2010

Abstract Long Terminal Repeat retrotransposons (LTR-RTs) are a major component of several plant genomes. Important insights into the evolutionary dynamics of these elements in a genome are provided by the comparative study of their insertion times. These can be inferred by the comparison of pairs of LTRs flanking intact LTR-RTs in combination with an estimated substitution rate. Over the past several years, different substitution rates have been proposed for LTRs in crop plants. However, very little is known about the extent of substitution rate variation and the factors contributing to this variation, so the rates currently used are generally considered rough estimators of actual rates. To evaluate the extent of substitution rate variation in LTRs, we identified 70 orthologous LTRs on the short arms of chromosome 3 of both *Oryza sativa* and *Oryza glaberrima*, species that diverged ~0.64 Ma. Since these orthologous sequences were present in a common ancestor prior to species divergence, nucleotide differences identified in comparing these regions must correspond to mutations accumulated post-speciation, thereby giving us the opportunity to study LTR substitution rate variation in different elements across these short arms. As a control, we analyzed a similar amount of non-repeat-related sequences collected near the orthologous LTRs. Our analysis showed that substitution rate variation in LTRs is greater than 5-fold, is positively correlated with G+C content, and tends to increase near centromeric regions. We confirmed that in the vast majority of cases, LTRs mutate faster than their corresponding non-repeat-related neighboring sequences.

Keywords Substitution rate · Long terminal repeat · Retrotransposon

Electronic supplementary material The online version of this article (doi:10.1007/s12284-010-9046-7) contains supplementary material, which is available to authorized users.

A. Zuccolo · A. Sebastian · Y. Yu · R. A. Wing (✉)
Arizona Genomics Institute, University of Arizona,
Thomas W. Keating Bioresearch Building 1657 E. Helen Street,
Tucson, AZ 85721, USA
e-mail: rwing@ag.arizona.edu

A. Zuccolo · A. Sebastian · Y. Yu · S. Rounsley · R. A. Wing
Department of Plant Sciences, BIO5 Institute for Collaborative Research, University of Arizona,
Tucson, AZ 85721, USA

S. Jackson
Agricultural Genomics, Purdue University,
West Lafayette, IN 47907, USA

D. Billheimer
BIO5 Institute, University of Arizona,
Tucson, AZ 85721, USA

Introduction

Long Terminal Repeat (LTR) retrotransposons (RTs) are widespread and ubiquitous in the plant kingdom (Flavell et al. 1992; Voytas et al. 1992; Suoniemi et al. 1998) where they constitute significant portions of many genomes (Feschotte et al. 2002; IRGSP 2005; Tuskan et al. 2006; Jaillon et al 2007; Zuccolo et al. 2007; Ming et al. 2008). They contribute actively to genome size variation (Hawkins et al. 2006; Piegu et al. 2006; Neumann et al. 2006; Ammiraju et al. 2007) and gene expression (Varagona et al. 1992; Leprinc et al. 2001; Kashkush et al. 2003) and are involved in genome rearrangements (Ma et al. 2005). Because of their potentially mutagenic effects, transposable

elements (TEs) are strictly regulated by epigenetic silencing mechanisms (Lisch 2009) and are major targets for DNA methylation in plant genomes (Bender 2004).

Useful information about the "history" of LTR-RTs in a genome is provided by the comparative study of their insertion times in the host genome that can be inferred from the comparison between the two LTRs from each individual element. Due to the mechanisms of retrotranscription and insertion, LTR-RTs contain two identical LTRs at the moment of insertion (Lewin 1997). Over time, each LTR in a pair accumulate independent mutations and diverge (SanMiguel et al. 1998). Thus, sequence comparison between pairs of divergent LTRs allows one to estimate when an insertion in the host genome occurred when combined with an appropriate substitution rate.

Although substitution rates have been made for genes (Gaut et al. 1996) and LTR-RTs (Ma and Bennetzen 2004; Vitte et al. 2004), very little is known about the extent of LTR-RT substitution rate variation as well as the factors contributing to this variation. To better understand substitution rate variation associated with LTR-RTs and factors contributing to such variation, we took advantage of a unique plant within-genus sequence data set of chromosome 3 short arms from two cultivated *Oryza* species—*Oryza sativa* ssp. *japonica* (Asian rice) and *Oryza glaberrima* (West African rice). The length of the two chromosome 3 short arms studied is 17,111,432 bp in the case of *O. glaberrima* and 19,401,704 in that of *O. sativa*. We identified all orthologous LTR insertions between *O. sativa* and *O. glaberrima,* inferred their substitution rates, and then compared these rates with flanking sequences unrelated to TEs. We found that the substitution rates in LTRs vary by more than 5-fold, that this variation is positively correlated with the G+C content, and tends to increase near the centromere.

Materials and methods

Sequence data The short arm of chromosome 3 (Chr3S RefSeqs) in *O. sativa* was obtained from http://rgp.dna.affrc.go.jp/IRGSP/Build4/chr03.fasta.gz (IRGSP 2005). The corresponding orthologous sequences of *O. glaberrima* were obtained from BACs whose GenBank accession numbers are in Supplementary Table 1.

Mining orthologous LTR retrotransposons sequences from O. sativa and O. glaberrima The RepBase 13.02 library (Jurka et al. 2005) and an in-house collection of LTR retrotransposons isolated in *O. sativa* were used to search the *O. sativa* Chr3S pseudomolecule (IRGSP Version 4) for the presence of LTRs (from complete LTR-RTs, solo-LTRs, and LTRs from truncated elements (the latter defined as elements having internal coding domains but lacking of a clear second LTR)). Similarity searches were carried out using the program RepeatMasker (www.repeatmsker.org) and the algorithm BlastN (Altschul et al. 1997). All the LTRs identified in the *O. sativa* Chr3S were extracted along with significant portions of flanking genomic sequence (non-repeat-related flanking sequences: NRR) according to the rules depicted in Fig. 1. The tracts (LTR+NRRs) were then compared through similarity searches against the *O.*

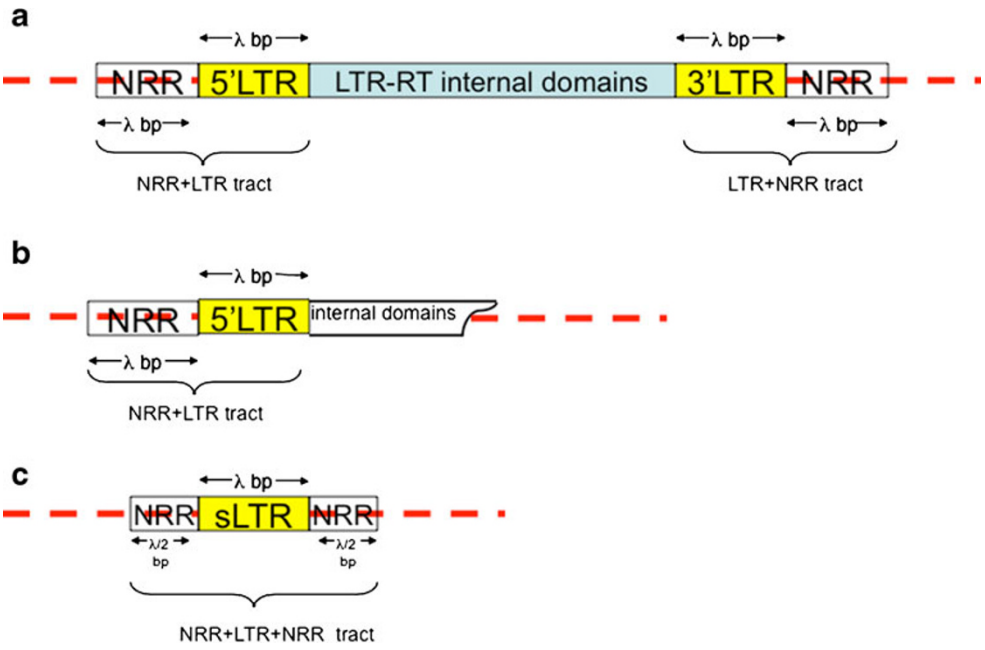

Fig. 1 Description of different kinds of orthologous tracts used for analysis. **a** Complete orthologous retrotransposons: LTR pairs were extracted along with two nearby genomic regions (NRRs) having a length similar to that of the LTR (λ bp). **b** Truncated orthologous retroelements: only the LTR was extracted plus a nearby genomic region (NRR) of similar length (λ bp). **c** Solo-LTRs: the LTR was extracted plus an overall similar amount of genomic sequence represented by two genomic tracts (NRR) flanking both sLTR ends (λ/2 bp each). Other details are in the text.

glaberrima BAC sequences spanning the complete *O. sativa* Chr3S to identify candidate orthologous insertions. Nucleotide similarity greater than 85% all over the complete length of the alignment was required to identify orthologous candidates (Supplementary Figure 1). Putative orthologous insertions were then confirmed by means of dot plot comparisons: sequence tracts including 10 Kbp upstream and 10 Kbp downstream from the orthologous insertion in both *Oryza* spp. were compared to each other using the program Dotter (Sonnhammer and Durbin 1995) and manually inspected (Supplementary Figure 2). Orthologous insertions were then extracted from both species. The NRR tracts were screened a second time using a public collection of repeats to remove all detectable transposable element-related sequences. In particular, we discarded from our analysis all the LTRs nested into other TEs.

Sequence data analyses The orthologous LTR+NRR tracts were aligned using the program "Stretcher" (EMBOSS package—Rice et al. 2000) and were edited using the program JalView 2.3 (Clamp et al. 2004). Genetic distances between orthologous tracts (i.e., LTRs and non-repeat flanking sequences) were estimated using the Kimura 2 parameters method (Kimura 1980) as implemented in the program "Distmat" (EMBOSS package—Rice et al. 2000). Mutations were analyzed using the program DNAsp V.4 (Rozas et al. 2003). The G+C content was determined using a custom PERL script available upon request.

Statistical analyses All the statistical analyses were carried out using scripts implemented in R language (R Development Core Team 2009).

Results

Comparison of the *O. glaberrima* and *O. sativa* Chr3S sequences (see "Materials and methods") revealed the presence of 70 orthologous LTR insertions, comprising 28 LTRs from 14 complete retroelements, 24 solo LTRs, and 18 LTRs from trunctated LTR-RTs, totaling 103,963 bp from orthologous LTRs and 91,954 bp from orthologous NRRs. The 14 complete elements were checked for the presence of target site duplications (TSD): all but one have TSDs. Orthologous LTR sequences were collected along with genomic sequences (NRR) flanking each LTR sequence (similar in size to each LTR sequence; Fig. 1). Since the *Oryza* genome is riddled with repetitive elements and their remnants (IRGSP 2005), all flanking NRR sequences were carefully inspected to identify and remove any kind of repetitive sequence that could be detected. The orthologous tracts (LTR+NRR) were aligned and manually inspected to identify all substitutions that accumulated in both the LTR and NRR sequences, respectively, since speciation took place (Supplementary Figure 1). Since the orthologous tracts were present in a common ancestor of *O. sativa* and *O. glaberrima* before species divergence, substitutions identified in both the LTR and NRR sequences of both species should have accumulated during the same amount of time. This is true for all the orthologous tracts isolated and thus offers a unique opportunity to study substitution rate variation in LTRs across the Chr3S.

Assessing the extent of nucleotide distance variation in LTRs and NRRs

Analyzing these regions enabled us to compare the nucleotide distance between LTR and NRR sequences as well as between different LTRs in different chromosomal regions. Nucleotide distances were calculated using the Kimura 2 parameters (K2p) method (Kimura 1980) (Supplementary Table 1). As expected, in almost all cases (61 out of 70), LTRs accumulated greater nucleotide distance than their corresponding NRR sequences (Supplementary Table 1). The distribution of nucleotide distances in the two sets was analyzed using the Mann–Whitney test indicating that they were significantly different ($p<0.0001$). Furthermore, the variation in nucleotide distance for LTR and NRR regions was totally unrelated ($n=70$; $r=0.0214$; $p=0.8604$; Fig. 2). This result clearly indicates LTR–NRR pairs have different substitution rates. The nucleotide K2p distances for LTRs ranged 5.9-fold, from 0.0094 to 0.0552, with an average value of 0.0282, whereas the K2p distances for the NRR regions varied between 0 and 0.0298 (average=0.0130), less than half that of the LTRs distances (Fig. 3). The differences between nucleotide distance average values for NRRs and LTRs were statistically significant (Student's t test: $p<0.0001$).

Two thirds of LTR nucleotide distances fell within a 2-fold range (0.0162–0.0332), whereas 60% of the nucleotide distances varied 3-fold (0.005–0.0152) for the NRR sequences. This fact was reflected by the coefficients of variation that were 37.6% and 55.5% for LTR and NRR distances, respectively, indicating a greater degree of variation for nucleotide distances in the NRR flanking regions. To reduce the possible contribution of LTRs from complete LTR-RTs to the homogenization of nucleotide distances for LTRs alone, we recalculated the coefficient of variation for LTRs by removing all three LTRs from the complete elements. A very similar value for the coefficient of variation was obtained—38.52%. The ratio of nucleotide distances in LTR–NRR pairs varied between 0.422 and 14.609 reflecting the lack of correlation between variation in nucleotide distance for LTR and NRR regions.

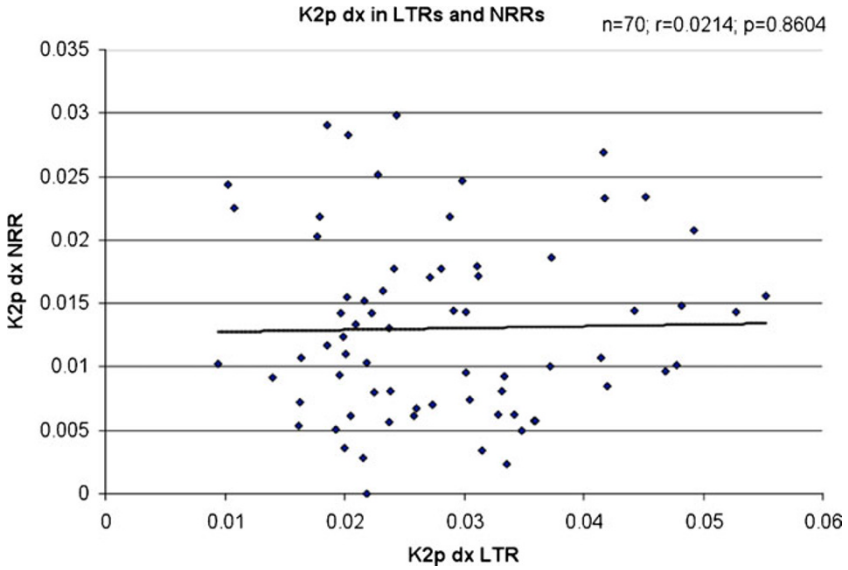

Fig. 2 Kimura 2p distance (dx) value for LTR and corresponding NRR sequences were plotted and correlation values were calculated.

Characterization of the forces behind nucleotide distance variation

To better understand the cause of differential substitution rates in LTR vs. NRR sequences, we cataloged base pair changes that occurred in each orthologous LTR and NRR sequence (Table 1). For LTRs, the overall frequency of G:C <-> A:T mutations was 0.01097 per nucleotide (95% confidence interval=0.0103–0.0116), whereas in NRR sequences, it was less than half that of LTRs at 0.0043 (confidence interval=0.0038–0.0047). Although we lack direct information about the ancestral state of LTR nucleotides to determine mutation directionality (C to T or vice versa?), we tried to infer it by exploiting the availability of LTR quadruplets provided by the 14 pairs of intact orthologous LTR retroelements that we identified. Since all LTRs from a single pair of intact orthologous elements are predicted to be highly similar, it is possible to

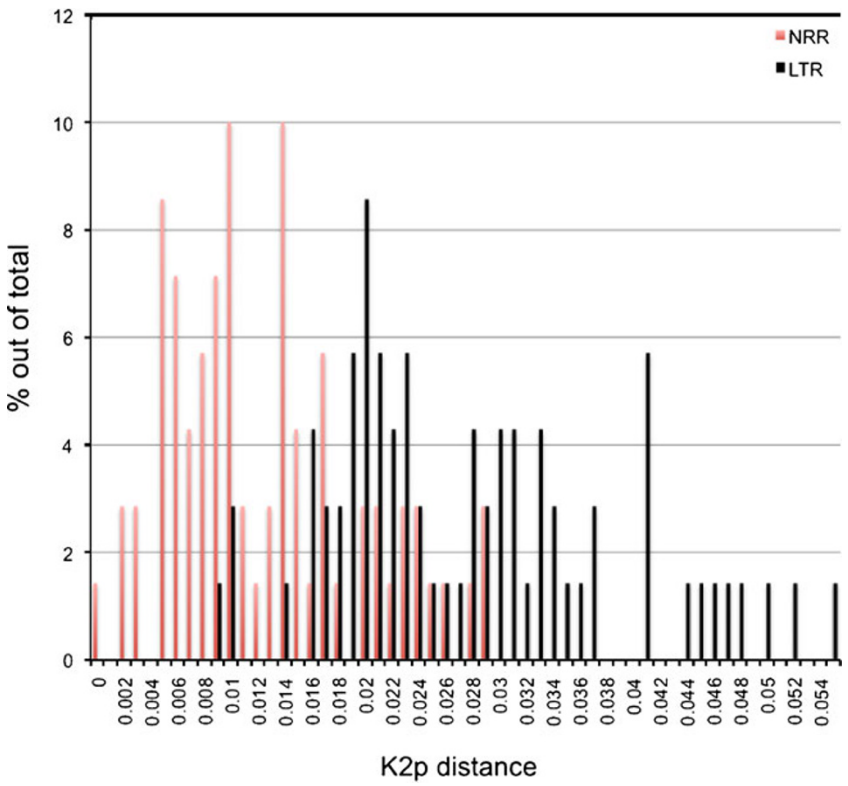

Fig. 3 Frequency of different Kimura 2p distance were plotted for both LTRs (*black bars*) and NRRs (*red bars*).

Table 1 Frequency of Different Substitutions

Subsitution	NRR	LTR	NRRf	LTRf
C<->A (G<->T)	88 (70.58–108.42)	132 (110.44–156.53)	0.0010 (0.0008–0.0012)	0.0014 (0.0012–0.0017)
C<->T (G<->A)	391 (353.20–431.74)	1,140 (1,074.78–1,208.15)	0.0043 (0.0038–0.0047)	0.0110 (0.0103–0.0116)
A<->T	74 (58.11–92.90)	57 (43.17–73.85)	0.0008 (0.0006–0.0010)	0.0006 (0.0005–0.0008)
C<->G	29 (19.42–41.65)	49 (36.25–64.78)	0.0003 (0.0002–0.0005)	0.0005 (0.0004–0.0007)

Values in parenthesis are the confidence intervals calculated assuming the substitutions occurring according a Poisson distribution

Substitution type of substitution; *NRR* occurrences in NRR sequences; *LTR* occurrences in LTRs, *NRRf and LTRf* frequencies of substitution per nucleotide

infer the original base in a substitution event by comparing the corresponding position in the four LTRs: if three out of four sites have, for instance a C, and the fourth position has a T, we consider this is good evidence that the original base was a C.

Parsing a total of 341 informative sites, we found a preponderance (315 to 26) of C to T (or G to A) mutations over the reverse suggesting that the mutation path from C to T remains the most common scenario in the case of LTR retrotransposon-related sequences. This possibly reflects the effects of cytosine methylation (Duncan and Miller 1980).

All remaining substitution types did not exhibit significant differences between NRRs and LTRs.

The G+C content of LTRs was, on average, greater than that of the NRRs (45.66% vs. 41.56%; p=9e-04 derived from a t test comparing distributional means). When LTR and flanking NRRs were considered together, the G+C content was greater for LTRs than NRRs in 73% of the cases (51 of 70). For the remaining 19 cases, five had higher substitution rates in NRRs than in LTRs, suggesting a possible effect of G+C content on substitution rate.

To rigorously determine if nucleotide variation rates for both the LTR and NRR flanking sequences were associated (and to which extent) with their positions along the Chr3S and G+C content, we used a multiple linear regression model. The dependent variable was the nucleotide distance, and the predictive variables were the position of the sequence on the chromosome and G+C content. In the case of LTRs, the regression results indicated that both position and G+C content are significantly associated with substitution rate. For every one million base pairs in position change along the chromosome towards the centromeric region, nucleotide distance increases by about 0.0006 (95% confidence interval, 0.00007 to 0.00011; Fig. 4). Similarly, a 1% increase in G+C content is associated with a 0.0007 increase in nucleotide distance. The ANOVA results indicate that G+C content accounts for about twice the variability of sequence position along the chromosome. Together, they account for 14% of total variation (Table 2). Nucleotide distances in the NRR regions are also positively associated with position along the chromosome and G+C content. Nucleotide distance increases by about 0.0004 for every one million base pairs of position change along the chromosome towards the centromeric region (95% confidence interval, 0.00004 to 0.00075). A 1% increase in G+C% content is associated with a 0.0002 increase in nucleotide distance. About 12% of the total variation in nucleotide distance is accounted for by these two factors.

To determine if K2p nucleotide distance variation was different between LTR-RT families, we identified and analyzed nine families that contained three or more members on Chr3S. For all families, we detected a large amount of variation of K2p nucleotide distances ranging from 1.43 fold for the Ty1-*copia* element family OSR_8 to 4.00 fold for the Ty3-*gypsy* RIRE10 family (Table 3): this result indicates that whatever the mechanism(s) causing mutations is, it does not appear to be family specific.

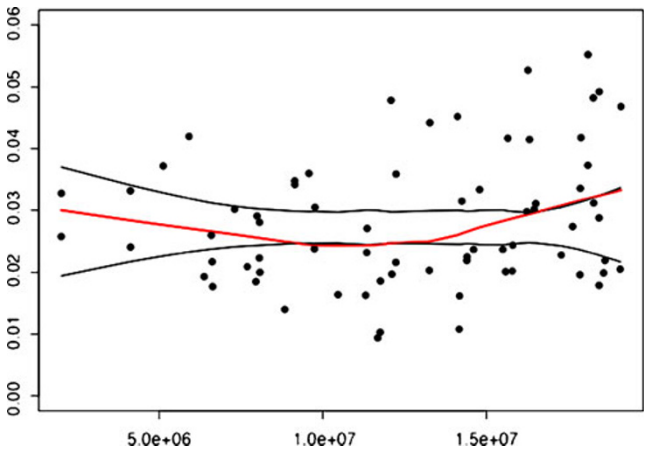

Fig. 4 Distribution of Kimura 2p nucleotide distances along *O. sativa* chromosome 3 short arm RefSeq for LTRs orthologous sequences. *Black dots* denote the nucleotide distance value (*Y*-axis) for each position. The *red line* is a nonparametric regression estimate of the local values. *Black lines* indicate upper and lower 95% confidence bounds on the smooth line computed assuming no positional preference (10,000 permutations). Position (*X*-axis) is expressed as nucleotides on the chromosome 3 short arm RefSeq. Centromere position is estimated at 1.94e+07 bp.

Table 2 Multiple Linear Regression Analysis of LTR Nucleotide Distance Variation

Residuals					
Min	1Q	Median	3Q	Max	
−0.02013	−0.00678	−0.00182	0.00692	0.02200	
Coefficients					
	Estimate	Std. Error	*t* value	Pr(>|t|)	
(Intercept)	−0.0099256	0.0119506	−0.83100	0.40918	
posmbp	0.0005910	0.0002610	2.26400	0.02679	**
GCLTR	0.0006723	0.0002377	2.82800	0.00617	***
Residual S.E.	0.00996	67 DF			
Multiple R^2	0.1432000				
Adjusted R^2	0.1176000				
F statistic	5.5980000	(2 and 67 DF)			
Analysis of variance table					
	Df	Sum Sq	Mean Sq	*F* value	Pr(>*F*)	
posmbp	1	0.0003171	0.0003171	3.1983	0.07823	*
GCLTR	1	0.000793	0.000793	7.9982	0.00617	***
Residuals	67	0.006643	0.0000991			

Q quartile; *S.E.* standard error; *sq.* square; *DF* degrees of freedom; *posmbp* position along the chromosome (expressed as 1 Mbp bins), *GCLTR* LTR G+C content

***p=0.001, **p=0.01, *p=0.05

Evaluating the errors introduced by the current LTR-RT dating methodology

Since 1998 (SanMiguel et al. 1998), analysis of sequence divergence between paired LTRs from intact LTR-RTs, coupled with a molecular clock, has been the accepted method used to estimate the insertion times of LTR-RTs in plant genomes. Our analysis revealed the presence of 14 complete orthologous LTR-RTs between *O. glaberrima* and *O. sativa* which must have inserted into the genome of the common ancestor of these species. Insertion time estimates for these elements should, therefore, yield the same result in the two species if a single substitution rate really applies, and indeed, in 10 of 14 cases, the estimated insertion time varied less than 0.3 million years between the two species when the substitution rate of 2×10^{-8} mutations per synonymous site per year is used (Vitte et al. 2004). However, in the four remaining cases, we found one LTR-pair having a K2p distance almost twice as high in *O. glaberrima* than *O. sativa* (0.0629 vs. 0.0331; Table 4), leading to approximate insertion times of 1.57 and 0.83 million years, respectively. The significant divergence of some of these estimates is indicative of an error introduced by using a single substitution rate for these species, which shared a common ancestor some 640,000 years ago, but have since experienced different evolutionary histories.

In addition to the identification of significant variation in K2p distance rates between pairs of orthologous LTR between species, we also observed LTR pairs from the same elements that accumulated different numbers of mutations during the same time period, post-speciation. For example, we found one case where the K2p distance rates varied by more than 2-fold (0.0442 vs. 0.0203), and in only two cases out of 14 where the K2p distance rates less that 10% between LTR pairs (Table 4).

Discussion

The half-life of LTR retrotransposable elements in cereals has been estimated to be approximately six million years (Ma et al. 2004); thus, efforts to measure LTR substitution

Table 3 Kimura 2p Distance Variation within LTR Retrotransposon Families

Family	Elements	Lowest k2p	Highest k2p	Variation	%GC
Atlantys	4	0.014	0.024	1.71	38.94
BAJIE	17	0.016	0.05	3.13	45.01
COPI1	3	0.031	0.048	1.55	42.38
Copia_Ecgs	5	0.021	0.055	2.62	46.49
GYPSY1	3	0.022	0.042	1.91	47.38
OSR_8	4	0.021	0.03	1.43	46.20
RIRE10	11	0.011	0.044	4.00	47.15
RIRE2	3	0.020	0.045	2.25	39.30
SZ-50	3	0.009	0.027	3.00	49.42

Only families having at least three elements in the Chr3S were considered. Variation is the ratio between the highest and the lowest K2p distance values for the elements of that family. %GC is the G+C content: value is the average of the orthologous LTRs

Table 4 Kimura 2p Nucleotide Distances of Complete Retrotransposons

No	Element family	Kdx	Kdx (G)	Kdx(S)	MY (G)	MY (S)
1		0.0332				
2	Copia_Ecgs	0.0241	0.0629	0.0331	1.5725	0.8275
8		0.0232				
9	SZ-50	0.0271	0.0382	0.0528	0.9550	1.3200
10		0.0336				
11	RIRE 2	0.0196	0.0341	0.0318	0.8525	0.7950
16		0.0288				
17	BAJIE	0.0492	0.0507	0.0401	1.2675	1.0025
21		0.0442				
22	RIRE10	0.0203	0.0533	0.0403	1.3325	1.0075
23		0.0281				
24	RIRE10	0.0200	0.022	0.0261	0.5500	0.6525
25		0.0328				
26	Gypsy/T	0.0258	0.0452	0.0443	1.1300	1.1075
31		0.0348				
32	BAJIE	0.0342	0.0475	0.0572	1.1875	1.4300
42		0.0216				
43	[a]GYPSY1	0.0359	0.0468	0.0452	1.1700	1.1300
64		0.0108				
65	RIRE10	0.0162	0.0217	0.0199	0.5425	0.4975
66		0.0552				
67	Copia_Ecgs	0.0373	0.0346	0.0464	0.8650	1.1600
69		0.0103				
70	Copia/L	0.0186	0.0144	0.0102	0.3600	0.2550
71		0.0225				
72	Copia/I	0.0219	0.0355	0.0358	0.8875	0.8950
73		0.0482				
74	BAJIE	0.0312	0.0341	0.0482	0.8525	1.2050

No LTRs as described in Supplementary Table 1; *Element family* retrotransposon classification; *Kdx* K2p nucleotide distance calculated between orthologous LTRs in *O. sativa* and *O. glaberrima*; *Kdx* (*G*) Kimura 2p nucleotide distance calculated between the LTRs of the complete retrotransposon in *O. glaberrima*; *Kdx* (*S*) Kimura2p nucleotide distance calculated between the LTRs of the complete retrotransposon in *O. sativa*; *MY* (*G*) insertion time estimate in million years in *O. glaberrima*: substitution rate of 2×10^{-8} mutations per synonymous site per year is used (Vitte et al. 2004); *MY* (*S*) insertion time estimate in million years in *O. sativa*: substitution rate of 2×10^{-8} mutations per synonymous site per year is used (Vitte et al. 2004)

[a] Does not have TSDs

rate variation across genera (e.g., rice-maize-sorghum) are virtually impossible. Here, we used a within-genus model system to assess the extent of LTR substitution rates by scanning highly accurate Sanger-sequenced Chr3S pseudomolecules from *O. sativa* and *O. glaberrima* for orthologous transposable elements and their derivatives. These species diverged from a last common ancestor 0.64 Ma (Ma et al. 2004) and offer an ideal evolutionary vista to analyze nucleotide rate variation in both LTR-RTs and neighboring sequences.

The orthologous tracts used in this analysis comprised two contiguous LTR and NRR sequences of approximately the same lengths and were assumed to be subjected to the same "environmental pressures" since both sequence types are physically linked in the same genomic location. Thus, different mutational behavior between LTR and NRR sequences could be ascribed directly to the different nature of the sequences (i.e., LTR-RT vs. intergenic). Comparison of nucleotide distances in LTRs versus NRRs showed that in most cases (61 of 70) LTRs were evolving more rapidly than NRRs. We found only nine cases where NRR sequences were evolving faster, five of which had higher G+C contents that could explain the higher K2p distance rates. For the four remaining cases, an alternative hypothesis could be that the flanking LTRs are under evolutionary constraint or that the NRR regions actually contain uncharacterized repeats that were not detected during similarity searches.

Calculated nucleotide distances showed significant variation across the Chr3S RefSeq for both LTR and NRR sequences with a tendency to increase towards the centromeric region. The magnitude of variation in the case of LTRs spanned an almost 6-fold range. Variation of nucleotide distances is largely expected for coding genes (Wolfe et al. 1989; Zhang et al. 2002) and has recently proved to be significant also for intergenic regions in Arabidopsis (DeRose-Wilson and Gaut 2007). The nucleotide distances calculated for LTRs in most of the cases studied here were higher than that of the nearby NRR sequences; however, the degree of variation for nucleotide distances was smaller for LTRs than that for NRRs. This evidence can be explained considering that the species

studied are close enough in evolutionary time that an appreciable amount of variance among loci is going to be due to coalescence processes, even if every locus has an identical substitution rate. However, if the LTR-RT elements are active or have been active in recent evolutionary time, they can only have a lower variance due to coalescence because of new or recent insertions. On the other hand, old insertions, just like NRR loci, can have long coalescences. Thus, on average, LTRs will have less variance among loci than NRRs (assuming homogenous mutation rates among loci). Yet, the amount of variability for substitution rate in LTRs remain high to the point that even LTRs belonging to the same element could accumulate mutations with a 2-fold differential rate between them. The clear lack of a relationship between the nucleotide distance in LTRs and flanking NRR sequences suggests that whatever the mechanism(s) acting on LTRs is, it does not extend its effects to the immediate flanking sequences (as far as mutations are involved). Our data also demonstrated that the mechanism(s) inducing these mutations was not specific to one LTR-RT family over another.

We identified at least two features of LTRs that appear to be important contributors to nucleotide distance variation—namely, G+C content and sequence position along the chromosome. It is clearly evident that nucleotide distances are positively correlated with G+C content, not only with LTRs but also with NRRs. Our results are consistent with previous findings that showed a positive correlation with G+C content and nucleotide substitution rates in *Arabidopsis thaliana* and *Arabidopsis lyrata* (DeRose-Wilson and Gaut 2007). Similarly, it is evident that substitution rate variation for both LTRs and NRRs increases towards the centromeric regions. This evidence agrees with recent findings in *A. thaliana* where a higher mutation rate in pericentromeric regions has been demonstrated (Ossowski et al. 2010). However, G+C content and position along the chromosome alone cannot explain all the variation. GC dinucleotides, CpNpG, and CpHpHp trinucleotides are known to be one of the major targets for DNA methylation (Gruenbaum et al. 1981), and DNA methylation in turn is one of the methods used by host genomes to control the chaotic effects of transposable element proliferation (Kumar and Bennetzen 1999; Zilberman and Henikoff 2004). Importantly, a certain amount of direct correlation between methylation and the position of genes along the chromosome has been demonstrated in *A. thaliana* where genes near centromeres were found to have a higher likely hood of being methylated (Zilberman et al. 2006). Our analysis demonstrated that the vast majority of differences in the frequency of substitution between LTRs and flanking NRRs could be ascribed to mutations possibly affecting cytosines (G:C ->A:T) which are targets of DNA methylation. In contrast, when the frequency of substitutions is compared between bases that are normally not methylated, no significant differences could be identified between LTRs and flanking NRR tracts. It is easy to speculate that the major driving force behind the mutation rate variation patterns described in this paper, including the positive correlation with G+C content and the increasing trends towards centromeres, is possibly associated with methylation. This hypothesis can be tested by performing a detailed characterization of methylation patterns in the LTR-RT pool which is beyond the scope of this work.

The molecular paleontology method of LTR-RT insertion dating has never claimed to provide rigorous and exact insertion time estimates because the pitfalls of using a single substitution rate for a population of LTR retrotransposons are well known (SanMiguel et al. 1998; Pereira 2004; Ma and Bennetzen 2006; Piegu et al. 2006). This work provides the first assessment of the extent of substitution rate variation affecting LTRs in a population of LTR retrotransposons in two closely related species. It should, however, be noted that this work, because of its experimental design, focused only on elements older than 0.64 million years. The amount of nucleotide distance variation on younger elements remains to be assessed. Our data confirm the cautious approaches that have characterized the use of the molecular paleontology dating so far. Since LTR variation rates were shown to span a nearly 6-fold range, LTR-RT insertion time dating that relies on a very general and approximate substitution rate is prone to severe errors. Such errors not only occur when different elements are analyzed in the same species but also, although to a lesser extent, when the same element is studied in two different but closely related species. These limitations indicate that LTR-RT insertion time estimate should be considered as a general qualitative assay rather than a quantitative estimation.

Acknowledgments This work was supported by the National Science Foundation (Grant DBI-0638541 to R.A.W., S.J., and S.R.) and the Bud Antle Endowed Chair (to R.A.W.).

References

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

Ammiraju JS, Zuccolo A, Yu Y, Song X, Piegu B, Chevalier F, et al. Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus Oryza. Plant J. 2007;52(2):342–51.

Bender J. DNA methylation and epigenetics. Annu Rev Plant Biol. 2004;55:41–68.

Clamp M, Cuff J, Searle SM, Barton GJ. The Jalview Java Alignment Editor. Bioinformatics. 2004;20:426–7.

DeRose-Wilson LJ, Gaut BS. Transcription-related mutations and GC content drive variation in nucleotide substitution rates across the genomes of *Arabidopsis thaliana* and *Arabidopsis lyrata*. BMC Evol Biol. 2007;7:66.

Duncan BK, Miller JH. Mutagenic deamination of cytosine residues in DNA. Nature. 1980;287(5782):560–1.

Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 2002;3:329–41.

Flavell JA, Dunbar E, Anderson R, Pearce SR, Hartley R, Kumar A. Ty1-copia group retrotransposons are ubiquitous and heterogeneous in higher plants. Nucleic Acids Res. 1992;20:3639–44.

Gaut BS, Morton BR, Mccaig BC, Clegg MT. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc Natl Acad Sci USA. 1996;93 (19):10274–9.

Gruenbaum Y, Naveh-Many T, Cedar H, Razin A. Sequence specificity of methylation in higher plant DNA. Nature. 1981;292(5826):860–2.

Hawkins J, Kim H, Nason JD, Wing RA, Wendel JF. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res. 2006;16:1252–61.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449(7161):463–7.

Jurka J, Kapitonow VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.

Kashkush K, Feldman M, Levy AA. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet. 2003;33(1):102–6.

Kimura M. Simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16:111–20.

Kumar A, Bennetzen JL. Plant retrotransposons. Annu Rev Genet. 1999;33:479–532.

Leprinc AS, Grandbastien MA, Meyer C. Retrotransposons of the Tnt1B family are mobile in *Nicotiana plumbaginifolia* and can induce alternative splicing of the host gene upon insertion. Plant Mol Biol. 2001;47:533–41.

Lewin B. Genes VI. Oxford: Oxford University Press; 1997.

Lisch D. Epigenetic regulation of transposable elements in plants. Annu Rev Plant Biol. 2009;60:43–66.

Ma J, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA. 2004;101 (34):12404–10.

Ma J, Bennetzen JL. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. Proc Natl Acad Sci USA. 2006;103(2):383–8.

Ma J, Devos KM, Bennetzen JL. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. 2004;14(5):860–9.

Ma J, SanMiguel P, Lai J, Messing J, Bennetzen JL. DNA rearrangement in orthologous orp regions of the maize, rice and sorghum genomes. Genetics. 2005;70(3):1209–20.

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature. 2008;452(7190):991–6.

Neumann P, Koblizkova A, Navratilova A, Macas J. Significant expansion of Vicia pannonica genome size mediated by amplification of a single type of giant retroelement. Genetics. 2006;173:1047–56.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, et al. The Rate and Molecular Spectrum of Spontaneous Mutations in *Arabidopsis thaliana*. Science. 2010;327(5961):92–4.

Pereira V. Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. Genome Biol. 2004;5(10): R79.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res. 2006;16(10):1262–9.

R Development Core Team. R: A Language and Environment for Statistical Computing R. Austria: Foundation for Statistical Computing Vienna; 2009.

Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 2000;16:276–7.

Rozas J, Sanchez-Delbarrio JC, Messegyer X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 2003;19:2496–7.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. Nat Genet. 1998;20:43–5.

Sonnhammer EL, Durbin R. A dot matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 1995;167(1–2):GC1–10.

Suoniemi A, Tanskanen J, Schulman AH. Gypsy-like retrotransposons are widespread in the plant kingdom. Plant J. 1998;13:699–705.

The International Rice Genome Sequencing Project Project. The Map Based Sequence of the Rice Genome. Nature. 2005;436:793–800.

Tuskan GA, Difazio S, Jansson S, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science. 2006;313(5793):1596–604.

Varagona MJ, Purugganan M, Wessler SR. Alternative splicing induced by insertion of retrotransposons into the maize waxy gene. Plant Cell. 1992;4:811–20.

Vitte C, Ishii T, Lamy F, Brar D, Panaud O. Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L.). Mol Genet Genomics. 2004;272(5):504–11.

Voytas DF, Cummings MP, Koniczny A, Ausubel FM, Rodermel SR. Copia-like retrotransposons are ubiquitous among plants. Proc Natl Acad Sci USA. 1992;89:7124–8.

Wolfe KH, Sharp PM, Li WH. Rates of synonymous substitution in plant nuclear genes. J Mol Evol. 1989;29(3):208–11.

Zhang L, Vision TJ, Gaut BS. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. Mol Biol Evol. 2002;19(9):1464–73.

Zilberman D, Henikoff S. Silencing of transposons in plant genomes: kick them when they're down. Genome Biol. 2004;5:249.

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet. 2006;39:61–9.

Zuccolo A, Sebastian A, Talag J, Yu Y, Kim H, Collura K, et al. Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. BMC Evol Biol. 2007;7:152.