the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

*The Plant Journal* (2009) **60**, 181–193 doi: 10.1111/j.1365-313X.2009.03946.x

# Genome organization of the tomato *sun* locus and characterization of the unusual retrotransposon *Rider*

**Ning Jiang[1], Dongying Gao[1,†], Han Xiao[2] and Esther van der Knaap[2,*]**

[1]*Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA, and*

[2]*Department of Horticulture and Crop Science, Ohio State University, Wooster, OH 44691, USA*

Received 15 May 2009; accepted 26 May 2009; published online 29 June 2009.

[*]For correspondence (fax +1 330 263 3887; e-mail vanderknaap.1@osu.edu).

[†]Present address: Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA.

## SUMMARY

**DNA sequences provide useful insights into genome structure and organization as well as evolution of species. We report on a detailed analysis of the locus surrounding the tomato (*Solanum lycopersicum*) fruit-shape gene *SUN* to determine the driving force and genome environment that foster the appearance of novel phenotypes. The gene density at the *sun* locus is similar to that described in other euchromatic portions of the tomato genome despite the relatively high number of transposable elements. Genes at the *sun* locus include protein-coding as well as RNA genes, are small in size, and belong to families that were duplicated at the locus an estimated 5–74 million years ago. In general, the DNA transposons at the *sun* locus are older than the RNA transposons, and their insertion pre-dates the speciation of *S. lycopersicum* and *S. pimpinellifolium*. Gene redundancy and large intergenic regions may explain the tolerance of the *sun* locus to frequent rearrangements and transpositions. The most recent transposition event at the *sun* locus involved *Rider*, a recently discovered high-copy retrotransposon. *Rider* probably arose early during the speciation of tomato. The element inserts into or near to genes and may still be active, which are unusual features for a high-copy element. *Rider* full-length and read-through transcripts past the typical transcription termination stop are detected, and the latter are required for mobilizing nearby sequences. *Rider* activity has resulted in an altered phenotype in three known cases, and may therefore have played an important role in tomato evolution and domestication.**

**Keywords: genome structure, transposons, tomato, evolution, *Rider*.**

## INTRODUCTION

Tomato (*Solanum lycopersicum*) is a member of the *Solanaceae* family, which also includes potato (*S. tuberosum*), pepper (*Capsicum* spp), eggplant (*S. melongena*), petunia (*Petunia* spp.) and tobacco (*Nicotiana* spp.). Tomato is a model plant for genome studies in the *Solanaceae*, as well as a model for studies of fruit development, ripening and crop domestication. Genomes evolve through structural changes such as inversions, duplications, deletions and transpositions, in addition to point mutation, which leads to single nucleotide polymorphisms (SNPs). Recent discoveries regarding diversity among humans have revealed an extraordinary level of structural variation within this species (Kidd *et al.*, 2008), some of which underlie phenotypic diversity (McCarroll and Altshuler, 2007). Structural changes of genomes are often mediated by unequal and non-homologous recombination, facilitated by repetitive elements such as transposons.

Transposable elements (TEs), DNA fragments that are capable of replication and movement, are major components of eukaryotic genomes. TEs are divided into two classes. Class I elements or RNA elements (retrotransposons) use element-encoded mRNA as the transposition intermediate. These transposons are either flanked by a long terminal repeat (LTR) or lack terminal repeat sequences (non-LTR transposons). Class II elements (DNA transposons) are often characterized by the terminal inverted repeats (TIRs) and transposition through a DNA intermediate. Autonomous DNA TEs encode a transposase and other proteins required for transposition, while non-autonomous elements lack functional transposition proteins and rely on

cognate autonomous TEs for their transposition. Upon transposition, transposons contribute to genome diversity as a result of insertional polymorphism among closely related species. In rice, 14% of the DNA polymorphisms in the two sub-species *indica* and *japonica* were due to transposon insertion polymorphisms (Huang *et al.*, 2008). For most TEs, integration of the element is accompanied by duplication of a small segment of flanking genomic sequence. This is called target site duplication (TSD), and is the hallmark of transposition. LTR retrotransposons are very abundant in plants, and are largely responsible for the genome size expansion in certain species (Bennetzen, 1996). However, unequal recombination between the LTR pair of a single element often leads to deletion of the internal region and one of the LTR, resulting in the formation of a 'solo' LTR. This process significantly reduces the size expansion caused by element amplification (Ma *et al.*, 2004).

Transposons are also known to duplicate and mobilize gene sequences. For example, the maize *Bs1* LTR retrotransposon carries part of a plasma membrane proton-translocating ATPase gene without its intron sequences (Bureau *et al.*, 1994; Jin and Bennetzen, 1994). In rice, over 1000 genes duplicated through retrotransposition (retrogenes) have been identified, and many recruited new exons from flanking regions, resulting in the formation of chimeric genes (Wang *et al.*, 2006a). In addition, there are thousands of *Mutator*-like elements (MULE) that carry genes or gene fragments in the rice genome (Jiang *et al.*, 2004; Juretic *et al.*, 2005). Due to their ability to duplicate genes or gene fragments, transposons themselves may represent the structural variation among species or individuals in the population. For example, duplication of gene fragments by *Helitron* elements in maize has contributed significantly to many sequences that are not shared among maize cultivars at the orthologous position (Fu and Dooner, 2002; Morgante *et al.*, 2005).

One of the features of plant domestication is the selection of altered morphology of the harvested organ (Pickersgill, 2007). In tomato, continued selection of fruit characteristics has resulted in a diverse array of accessions that differ in shape, size and color (Paran and van der Knaap, 2007). We are specifically interested in investigating the molecular basis of variation in tomato fruit shape in order to obtain insights into genes and developmental pathways that have been selected during domestication and resulted in altered patterning of the fruit. Recently we isolated *SUN*, a gene that controls tomato morphology by producing an elongated and oval-shaped fruit. Sequence analysis of the locus demonstrated that *SUN* arose from an interchromosomal duplication event mediated by a newly discovered retrotransposon, *Rider*. The duplicated gene was placed in a new genomic context, resulting in increased expression and hence altered fruit shape (Xiao *et al.*, 2008). This mutation is unusual as it was caused by a TE that duplicated and transposed a gene into a different genomic environment, which is in contrast with more common mutations underlying phenotypic diversity such as SNPs in coding and/or promoter regions (e.g. Frary *et al.*, 2000; Liu *et al.*, 2002; Fridman *et al.*, 2004) or TE insertional inactivation of genes (e.g. Fray and Grierson, 1993; Doebley *et al.*, 1997).

In this study, we sought to investigate the genome organization of the tomato *sun* locus, as well as the ancestral region from which the locus arose. A detailed analysis of the *sun* locus, which has undergone several structural rearrangements (Van der Knaap *et al.*, 2004), has the potential to provide insights regarding the evolution of plant genomes and how structural changes have an impact on the phenotype. Therefore, *SUN* is an excellent example to assess the effects of the driving force, *Rider*, and genome environment, the *sun* locus, that determine the appearance of novel phenotypes such as elongated fruit shape. Unlike previous studies on genome structure within the *Solanaceae* (Wang *et al.*, 2005, 2006b; Datema *et al.*, 2008; Zhu *et al.*, 2008), which are genome-wide and largely based on prediction programs and similarity to known gene and repeat sequences, our analysis involved a detailed and partly manual classification and identification of all features in the sequenced regions. Therefore, this study provides one of the most detailed and comprehensive descriptions of the structural organization of a locus in the tomato genome. In addition, we conducted a detailed comparison between orthologous regions of almost 32 kb in *S. lycopersicum* and one of its closest wild relatives (*S. pimpinellifolium*), providing an insight into the genic and intergenic evolution of an extended region of these two closely related red-fruited tomato species. Furthermore, we characterized the newly identified *Rider* retrotransposon, which may have contributed significantly to genome evolution in tomato in the past as well as the present.

## RESULTS

### The genome structure at *sun* and the ancestral locus

At the *sun* locus, the tomato cultivar Sun1642, which has oval-shaped fruit, harbors a gene-rich 24.7 kb segment that is missing in the round-fruited wild relative *S. pimpinellifolium* accession LA1589 (Xiao *et al.*, 2008) (Figure 1). This 24.7 kb fragment carries the *SUN* gene, which was duplicated by a retrotransposition event mediated by the *copia*-like retrotranposon *Rider*. Transformation experiments and expression studies indicate that the promoter of the disrupted defensin gene drives *SUN* expression on chromosome 7 (Xiao *et al.*, 2008). To investigate the genome environment that permitted the frequent transposition and other rearrangements at the *sun* locus, we performed a detailed molecular analysis on the structure and composition of the locus as well as its ancestor on chromosome 10.

 

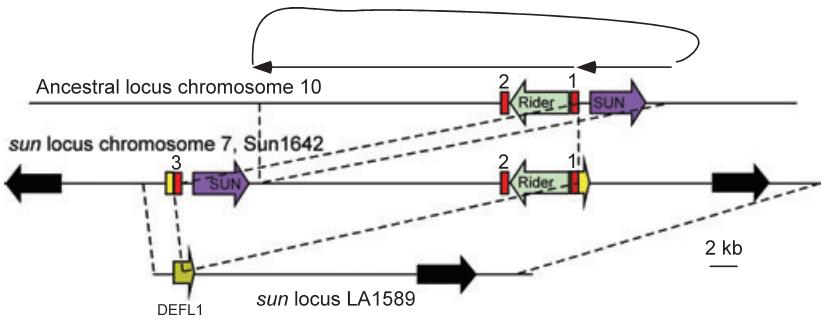


**Figure 1.** Comparative analysis of the *sun* locus on chromosome 7 and the ancestral locus on chromosome 10 (partially redrawn from Xiao *et al.*, 2008).
The green, purple and gold arrows indicate the retrotransposon *Rider*, the fruit-shape gene *SUN* and *DEFL1*, respectively. The yellow box and arrow represent the *DEFL1* gene that is disrupted by the transposition The black arrows indicates the inverted duplication flanking *sun*. The narrow red boxes represent the LTR sequences flanking the transposon. The numbers above the LTR indicate their order based on the proposed transcription of *Rider* on chromosome 10. The thin arrow above the chromosome 10 segment indicates *Rider* read-through transcription beyond LTR2 followed by the template switch. The dashed lines indicate homology between the regions.

There were 21 predicted genes with protein coding capacity in the sequenced regions (Figure 2 and Table S1). Eleven predicted genes were identified in the BAC clone representing the ancestral locus, giving a density of one gene per 5.6 kb in this region of chromosome 10, which is similar to other tomato euchromatic regions (Wang *et al.*, 2005). At the *sun* locus, there were four genes encoding defensins, one of which was interrupted by the insertion of *Rider* (Figure 2b). When the duplication from chromosome 10 was excluded, the gene density at the *sun* locus was one gene per 11 kb. The *sun* locus also contained six U2-snRNA genes that demonstrated sequence similarity of 78–96%. When counting the protein coding and RNA genes, the gene density in this region approached that of other euchromatic regions. However, the percentage of intergenic regions at the *sun* locus was much higher than that on chromosome 10 (87% versus 47%), due to the small size of both the defensin and RNA genes at the *sun* locus (Table S1).

Many of the repetitive sequences were recognizable as TEs, and included both DNA and RNA elements (Table 1, Figure 2 and Table S2). Although the number of TEs varied between chromosomes 7 and 10, DNA elements outnumbered RNA elements in the sequenced regions (47 versus 5). The overall density of both types of TEs was higher at the *sun* locus than at the ancestral locus on chromosome 10, possibly due to the larger proportion of intergenic regions at the *sun* locus. To determine whether other parts of the tomato genome also harbored a higher number of DNA TEs than RNA TEs, we examined genomic sequences from both euchromatic and heterochromatic regions that had been studied previously (Wang *et al.*, 2006b). Among the five BACs from euchromatic regions, we detected 81 DNA TEs (15.6 per 100 kb) and nine RNA TEs (1.7 per 100 kb). Among seven BACs from heterochromatic regions, we detected 62 DNA TEs (9.1 per 100 kb) and 87 RNA TEs (12.8 per 100 kb). This result strongly suggests that DNA elements are common repetitive sequences in the tomato genome, and that RNA TEs are more common in heterochromatic than euchromatic regions. In addition, the density of DNA and RNA TEs at the ancestral locus is close to that in other euchromatic regions, whereas the DNA TE density is much higher at the *sun* locus compared to other euchromatic and heterochromatic regions (39.4 versus 15.6 and 9.1, respectively, on average, per 100 kb).

Five LTR retrotransposons representing three elements were found in the sequenced regions (Table 1 and Figure 2). The LTR retrotransposons were generally larger in size than DNA transposons, which is why they comprise a relatively large percentage of the genome (Table 1). Forty-seven DNA TEs were found, of which 27 (57%) were intact with recognizable TSDs (Table 1 and Table S2). These TEs included the major types described in plants, such as *Mutator*-like elements, MITEs (*Tourist*/*Stowaway*), *hAT* and CACTA-like elements, but not *Helitron*. None of them encoded a transposase, indicating that these were non-autonomous elements.

### Conservation and divergence between *S. lycopersicum* and *S. pimpinellifolium*

The availability of sequences from both *S. lycopersicum* and *S. pimpinellifolium* provided a unique opportunity to observe the dynamics of sequence evolution in this region of the tomato genome. Of the three LTR retrotransposons at the *sun* locus, only the *copia*-like element fragment LeLTR002 was found in both, possibly representing a relic from an ancient TE insertion that was present before the two species diverged. A solo LTR showed sequence similarity to the *Jinling* retrotransposon (Wang *et al.*, 2006b), and was present only in *S. pimpinellifolium* (Figure 2b). Comparison of flanking sequences showed that this solo LTR is probably a new insertion in the *S. pimpinellifolium* accession LA1589 because the 5 bp TSD flanking the LTR was present once in *S. lycopersicum* Sun1642. Sun1642 harbored a *Rider* element, which was absent in LA1589 at the same location. Like the *Jinling* solo LTR, *Rider* also recently inserted, as the 5 bp TSD flanking the element in Sun1642 was represented once

 

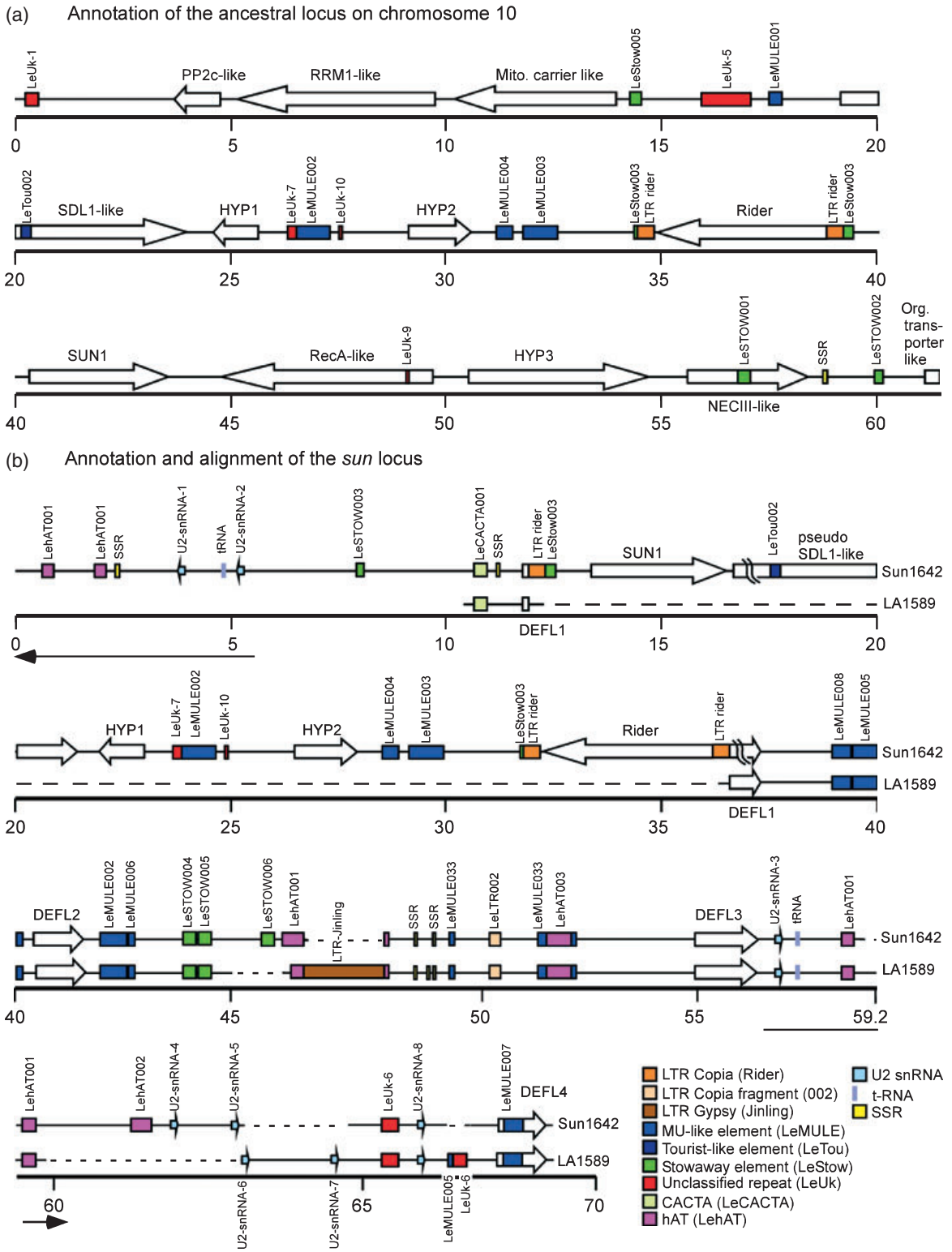


**Figure 2.** Detailed genome structure at the *sun* locus and the ancestral locus.
(a) Ancestral locus on chromosome 10.
(b) Comparative analysis of the *sun* locus in Sun1642 and LA1589. Dashed lines indicate gaps in the alignment. The inverted duplication is found at 1–5452 bp and 56 729–60 253 bp, and is indicated by the thin black arrow below the sequences. The region 2495–4198 bp is missing from the other duplicate region. White arrows indicate genes and the direction of transcription. Colored boxes depict the various transposable elements, unclassified repeats, tRNA and U2-snRNA genes (see legend in the bottom right-hand corner). The distances are given in kilobase pairs.

**Table 1** Repetitive sequences at the ancestral locus on chromosome 10 and the *sun* locus on chromosome 7

| | RNA TE | DNA TE | Other repeats | Total |
|---|---|---|---|---|
| Ancestral locus (61 605 bp, EF094939) | | | | |
| Copies | 1 | 9 | 5 | |
| Density[a] | 1.6 | 14.6 | 8.2 | |
| Size (fraction[b]) | 4867 (8.0) | 3557 (5.8) | 1692 (2.8) | 10 116 (16.6) |
| Sun locus Sun1642 (with duplication, 68 840 bp, EF094940) | | | | |
| Copies | 2 | 24 | 7 | |
| Density[a] | 2.9 | 34.9 | 10.2 | |
| Size (fraction[b]) | 5511 (8.0) | 10482 (15.2) | 757 (1.1) | 16 751 (24.3) |
| Sun locus Sun1642 (without duplication, 44 168 bp) | | | | |
| Copies | 1 | 18 | 5 | |
| Density[a] | 2.3 | 40.8 | 11.3 | |
| Size (fraction[b]) | 246 (0.6) | 8301 (18.8) | 558 (1.3) | 9105 (20.6) |
| Sun locus LA1589 (33 008 bp, EF094941) | | | | |
| Copies | 2 | 14 | 5 | |
| Density[a] | 6.1 | 42.4 | 15.1 | |
| Size (fraction[b]) | 2121(6.4) | 7991 (24.2) | 796 (2.4) | 10 908 (33.0) |

[a]Density refers to the number of elements per 100 kb genomic sequence.
[b]Fraction refers to the percentage of genomic sequence that the elements account for.

**Table 2** Sequence identity between Sun1642 and LA1589 at the *sun* locus

| Components | Coding region of genes | UTR | Intron | RNA genes | TE | All other regions | Average |
|---|---|---|---|---|---|---|---|
| Identity (%) | 99.7 | 99.5 | 98.9 | 98.9 | 98.4 | 98.7 | 98.7 |

in LA1589 at the corresponding site. Therefore, despite the few LTR elements at the *sun* locus, and with the exception of LeLTR002, the retrotransposons appear to have been active in recent times.

Despite the abundance of DNA elements, a comparison of the syntenic regions in Sun1642 and LA1589 indicated that only three of them were present/absent in the two species (Figure 2b). It is not clear whether the three polymorphic TEs represented new insertions or were deleted in the respective genomes, because their flanking sequences were absent from the other species in those cases. Therefore, there is no evidence that any of the DNA elements transposed into the *sun* locus after divergence of the two species.

The length of syntenic regions at the *sun* locus was approximately 58.3 kb for Sun1642 and 32.1 kb for LA1589 (Figure 2b). A large part of the size polymorphism was contributed by TEs, namely the 24.7 kb duplication and transposition in Sun1642 and the insertion of *Jinling* in LA1589. When excluding the retrotransposons, the lengths of sequence present in one species but not the other were 7.0 and 3.6 kb for Sun1642 and LA1589, respectively. Most of the polymorphic gaps (58 out of 66) were less than 20 bp long. In addition, both LA1589 and Sun1642 carried two large gaps (200 bp or longer). Excluding all gaps, the orthologous region spanned 26.6 kb, and this showed that the DNA sequences from the two species were highly similar. The extent of conservation is highest between coding regions (Table 2). Only three silent nucleotide changes were observed for each of the four defensin gene pairs, resulting in a sequence identity of 99.7%. The sequence identity was the lowest in the orthologous TEs, and this was significantly different from the genome component exhibiting the next lowest identity ($P < 0.005$, $\chi^2$ test) (Table 2). The pooled substitution rate for introns and UTRs between *S. lycopersicum* and *S. pimpinellifolium* is 1.0%. Based on this and a mutation rate of $6.03 \times 10^{-9}$ per year (Muse, 2000), the divergence of Sun1642 and LA1589 was estimated to have occurred 1.6 million years ago (MYA), which is close to what has been reported in the past for these two species (1.4 MYA) (Nesbitt and Tanksley, 2002).

### Age of the small- and large-scale genome rearrangements at the *sun* locus

Different mutation rates were applied for genes and TE to calculate the age of the genome rearrangements at the *sun* locus. In addition, we assumed that *S. lycopersicum* and *S. pimpinellifolium* diverged 1.4 MYA, based on the mutation rate at multiple loci (Nesbitt and Tanksley, 2002), and this may be a better estimate than our own, which was based on mutation rate at one locus. The most ancient series of duplications was found for the defensin genes. Overall, the nucleotide similarity of these genes ranged from 57 to 68%, with *DEFL1* and *DEFL2* being the most similar to one another. Thus, we estimated that the defensin gene amplification occurred more than 50 MYA (74 MYA based on coding regions and 59 MYA based on intron sequences),

 

possibly before the origin of the Solanaceae, which has been estimated to have occurred approximately 40 MYA (Wikstrom *et al.*, 2001; Wu *et al.*, 2006).

Based on sequence similarity among the members, the U2-snRNA genes were found to have duplicated more recently than the defensin genes, namely between 28 and 5 MYA. Moreover, four family members (U2-4, U2-5, U2-6 and U2-7, Figure 2b) were lost after the divergence of *S. lycopersicum* and *S. pimpinellifolium*, suggesting that birth and death for this gene family are highly dynamic, and their function may be redundant. The overall sequence similarity of the inverted duplication of approximately 4 kb (Figure 1) is 90%, which includes a U2-snRNA, a tRNA gene and a few TE fragments, as well as a 1.6 kb deletion (Figure 2b). The formation of the inverted duplication is estimated to have occurred at 9–11 MYA, depending on whether the TE or average mutation rate is used. The U2-snRNA genes present in the duplication display higher similarity to one another (96%) than to any other U2-snRNA gene in the region, suggesting that the inverted duplication may have occurred as recently as 5 MYA. This result supports the view that the inverted duplication is more recent than the amplification of other U2-snRNA genes at the *sun* locus.

The most recent rearrangement is the 24.7 kb transposition event, which showed 100% identity at the nucleotide level with the exception of three mismatches at the presumed template switch, which we assumed occurred during transposition (Xiao *et al.*, 2008). If we take the mutation rate of the TE as a guide, the transposition of *Rider* that created the *sun* locus occurred within the last 3500 years.

### *Rider* retrotransposon distribution, insertion preference and LTR polymorphisms

The TE responsible for the creation of the *sun* locus is the *Rider* retrotransposon, with 398 bp LTR. The internal region of the TE was 4071 bp and encoded a single protein of 1307 amino acids. The protein sequence was similar to that of other *copia*-like elements in plants, such as *Tnt1* (Grandbastien *et al.*, 1989). The internal region also contained other essential features that are required for transposition, including the primer binding site and polypurine track (Lewin, 2008). Thus, *Rider* could be an autonomous *copia*-like element. Because of the unusual gene duplication event mediated by *Rider*, we investigated this element in more detail while also comparing *Rider* to another known tomato retrotransposon *Jinling* (Wang *et al.*, 2006b). Based on LTR sequence similarity and searches through the available BAC sequences, we found 106 *Rider* elements, including 48 intact elements, 44 solo LTRs and 14 truncated elements from 53.4 Mb of genomic DNA. Assuming that the analyzed BAC sequences were representative of the genome, we predict that 1900 ($106 \times 950$ Mb/53.4) copies of *Rider* exist in tomato. Using the same tomato BAC sequence data, we identified 254 *Jinling* elements (96 complete elements, 68 solo LTRs and 90 truncated elements), which would result in 4500 *Jinling* elements in the tomato genome (see Appendix S1 for further discussion).

As the LTR of a single retrotransposon is identical upon insertion (Lewin, 2008), sequence divergence between LTR of the same element provides a measure of the time of insertion when an estimate of the nucleotide substitute rate is available (SanMiguel *et al.*, 1998). The *Rider* elements on chromosomes 7 and 10 have identical LTR, indicating they inserted into the tomato genome relatively recently. Of the 48 complete *Rider* elements in the available BAC sequences, all but one had a LTR identity of more than 96%. In addition, 24 (50%) of the intact elements have LTR with a nucleotide identity of 99% or higher, and five elements have identical LTR. In contrast, only three out of the 96 intact *Jinling* elements had such high sequence conservation, and none have identical LTR. This suggests that *Rider* elements are generally much younger than *Jinling* elements. Assuming a pooled substitution rate for silent sites of 0.8% (Nesbitt and Tanksley, 2002), over one-third (18 of 48) of the intact *Rider* elements for which the LTR is 99.2% identical or higher formed after the divergence of the two species and within 1.4 MYA. However, if we use the TE divergence estimation obtained in this study and the sequence similarity of 98.4% for TEs between *S. lycopersicum* and *S. pimpinellifolium*, nearly two-thirds (30 out of 48) of the intact elements are found to have inserted after divergence of the species.

If the above estimates are correct, one would expect many polymorphic *Rider* insertions between *S. lycopersicum* and *S. pimpinellifolium*. To test this, we used transposon display (TD), which generates PCR products anchored in a transposable element and a flanking restriction site (Van den Broeck *et al.*, 1998; Casa *et al.*, 2000). Insertional polymorphism, as defined by the presence of a PCR product in one accession but not in another, was approximately 70% between *S. lycopersicum* and *S. pimpinellifolium* (Figure 3a), demonstrating a burst of amplification of *Rider* after the divergence of the two species. Also, a few polymorphic fragments have been observed among various cultivars of tomato, suggesting that transposition of *Rider* occurred after domestication (Figure 3a). Although we cannot rule out the possibility that the intra-specific polymorphisms were due to introgressions from wild relatives, these polymorphisms mean that *Rider* is a useful marker in genetic studies of intra-specific comparisons in tomato. To survey the presence of *Rider* in more distant wild species, DNA blot analysis was performed using DNA from *S. pennellii* and *S. habrochaites* accessions, with the *Rider* LTR as a probe. As shown in Figure 3(b), the banding pattern varied between species, and the hybridization signal was reduced with DNA from *S. habrochaites* compared to the other distant relative *S. pennellii*. However, the banding pattern and hybridization signal was much more homogenous among the tomato species when using *Jinling* LTR as probe (Figure 3c). Again, the data

 

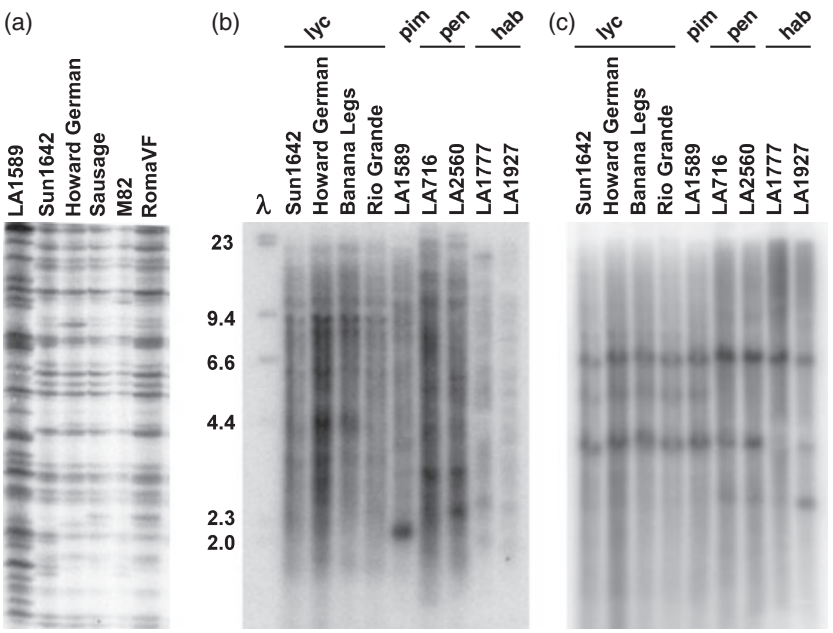


**Figure 3.** *Rider* and *Jinling* elements in the genome of tomato and its wild relatives.
(a) Transposon display using *Rider* LTR primers with DNA from LA1589 (*S. pimpinellifolium*) and five tomato cultivars. Note the strikingly different band pattern in LA1589 compared to the other accessions.
(b) Southern blot of *Eco*RV-digested genomic DNA from accessions indicated above the lanes. *Rider* LTR was used as the probe. 'lyc', *S. lycopersicum* accessions; 'pim', *S. pimpinellifolium* accession; 'pen', *S. pennellii* accessions; 'hab', *S. habrochaites* accessions. Note the lower hybridization signal with the *S. habrochaites* accessions.
(c) Southern blot of *Eco*RV-digested genomic DNA from accessions indicated above the lanes. *Jinling* LTR was used as the probe.
The final wash conditions for (b) and (c) were 63°C at 0.25 × SSC.

suggest that transposition and amplification of *Rider* in cultivated and wild tomatoes is more recent than that of *Jinling*. When the *Rider* sequence was used as a query to search the available potato genomic sequence (8 Mb BAC sequences and 108 Mb BAC end sequences), no sequences similar to the *Rider* LTR ($E < 1e^{-5}$) were retrieved. There were potato sequences with less than 70% similarity to the internal region of *Rider*, and these were presumably distant relatives of the element. These data suggest that amplification of *Rider* occurred after the divergence of potato and tomato.

The LTR sequence of a retrotransposon is composed of three regions comprising essential *cis*-elements for transcription start and termination, and integration of the element (Lewin, 2008). Those include the promoter (U3), the polyadenylation signal (R), and the sequence for termination of synthesis of element RNA (U5). Figure 4 shows an alignment of representative LTRs of the *Rider* family from intact elements, indicating that they are highly similar in the R and U5 regions while the U3 region is variable. The *Rider* element at the *sun* locus and its ancestral locus had the longest LTR among all family members. The shortest LTR lacked most of the U3 region including the TATA box. The second shortest LTR, which was 213 bp in length, retained the putative promoter, but most of the U3 region was absent. As each of these two shortest LTRs were associated with only one individual element in the analyzed genomic sequence (while all other variant LTRs were represented multiple times), it was possible that these LTRs were not competent for transposition and represent relic TEs. Other LTRs varied by an insertion or deletion (indels) in a single location within the U3 region, suggesting this is an error-prone region in duplication or transposition.

Another interesting feature of the *Rider* family LTRs were the two copies of the TTGT sequence in the U5 region, separated by 1 bp (Figure 4). This region is believed to be an important RNA synthesis termination signal (Temin, 1981). Interestingly, one copy of TTGT was mutated to TTAT in both LTR of *Rider* at the *sun* locus and its ancestral locus. In addition, four of the 48 intact *Rider* elements found in the available BAC sequences also carried the mutation in both LTRs, indicative of the presence of the mutation upon or before transposition of these elements (Table 3). For the remainder of the elements, the mutation was either polymorphic in the two LTR flanking the same element (TTGT/TTAT) or the element was truncated. Moreover, there was only one TTAT variant among the 44 solo LTRs (Table 3), suggesting that the mutation is relatively new.

To determine the insertion preference of *Rider*, the LTR sequence was used as a query to search against nucleotide sequences in GenBank, which resulted in three *Rider*-like elements in known genes. In addition to its insertion in the intron of *DEFL1* on chromosome 7, *Rider* also inserted into the first exon of tomato phytoene synthase gene *PSY1*, resulting in the abolishment of the gene and creating the yellow flesh mutation (Fray and Grierson, 1993). *Rider* also inserted into the tomato *FER* gene, leading to iron deficiency (Ling *et al.*, 2002; Cheng *et al.*, 2009). The presence of *Rider* in genes raised the question of whether the association of *Rider* with genes is a coincidence or whether it reflects target specificity. To this end, the 1 kb flanking sequences of complete *Rider* elements and solo LTRs were used to search Tomato Gene Index database from the Institute for Genomic Research, and compared with the flanking sequence of *Jinling*. Among the 92 *Rider* elements examined, 40 (43%)

 

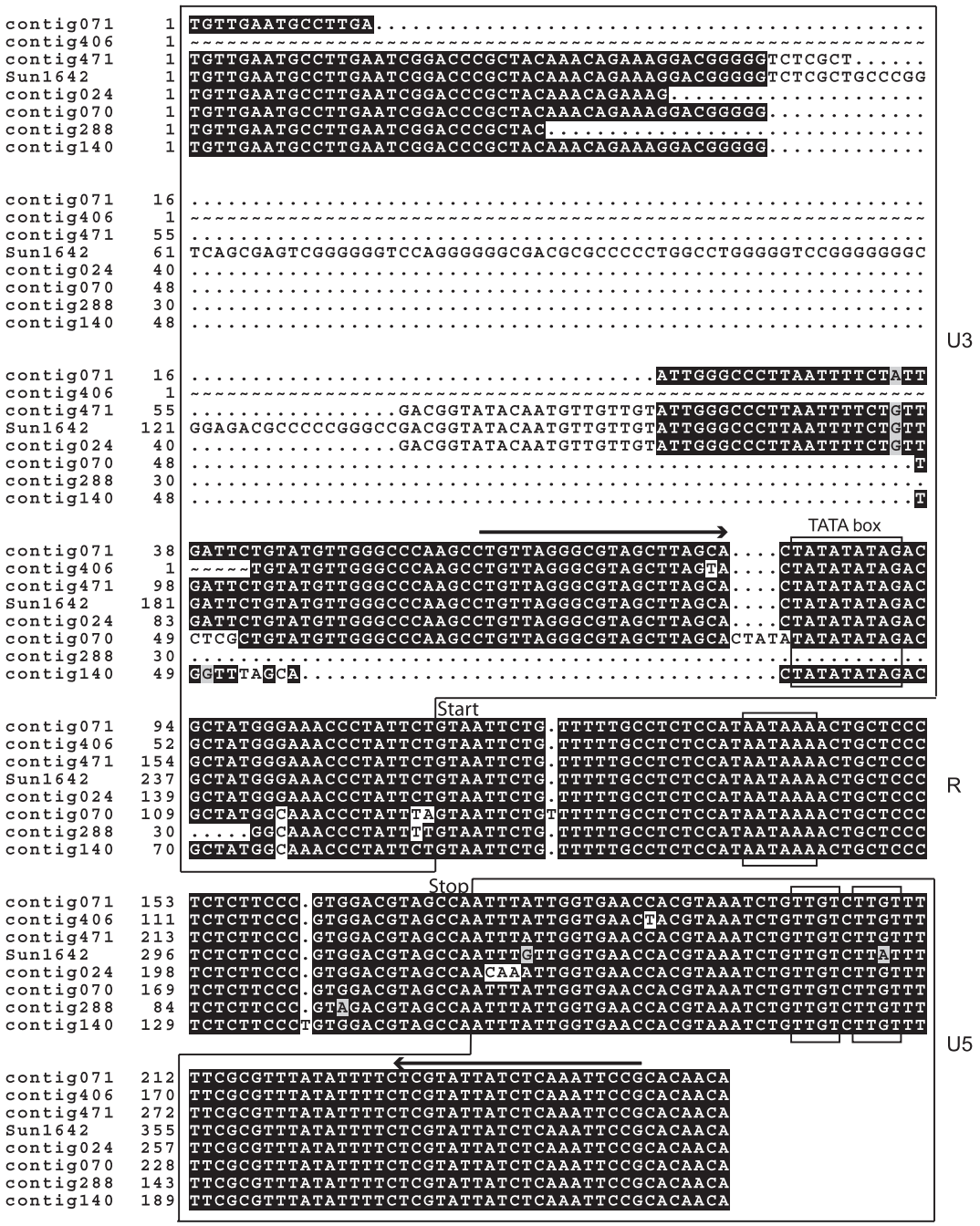


**Figure 4.** Alignment of the LTR sequences from individual *Rider* elements.
The U3, R and U5 regions are indicated. 'Start' indicates the putative transcription start site in 5′ LTR, and 'Stop' indicates the position where *Rider* transcription normally terminates. The arrows above the sequences correspond to the primers used for the detection of read-through *Rider* transcripts. The two adjacent boxed regions of four nucleotides each in the U5 region indicate the region thought to be involved in transcription termination and putatively involved in transcript read-through. The contigs represent the following BAC sequences: HBa0031A21 (contig024), HBa0303I24 (contig070), HBa0320D04 (contig071), HBa0107M13 (contig140), HBa0304I22 (contig288), HBa0057A19 (contig406) and SLm0066I05 (contig471).

were located within 1 kb of a putative gene and 19 (21%) were associated with a gene on both sides of the element. In comparison, only 20% of the *Jinling* elements (33 out of 164) were associated with a putative gene and four (2%) were flanked by genes on both sides. As *Jinling* elements are generally much older than *Rider* elements, it is possible that the insertion preference is an artifact caused by the age difference. To test this notion, we compared the intact *Rider* and *Jinling* elements that probably transposed in the same time frame, based on the similarity of their LTR. Among the

 

**Table 3** Frequency of TTAT in the U5 region of *Rider* LTRs

| | | TTAT variant | Total LTRs | Percentage of LTRs |
|---|---|---|---|---|
| Intact elements | TTGT/TTAT[a] | 10 | 48 × 2 | 18.8 |
| | TTAT/TTAT[b] | 4 × 2 = 8 | | |
| Solo LTRs | TTAT | 1 | 44 | 2.3 |
| Truncated elements | TTAT | 3 | 14 | 21.4 |
| Total | | 22 | 154 | 13.0 |

[a]TTAT mutation in only one LTR.
[b]TTAT mutation in both LTR flanking the element.

23 *Rider* elements with LTR similarity of 96–99%, 11 (48%) are associated with genes. In contrast, only 16 (21%) of the 75 *Jinling* elements exhibiting 96–99% LTR identity are associated with genes. This strongly suggests that *Rider* inserts much more frequently into genic regions than *Jinling* does.

To test whether *Rider* inserts into a specific sequence, we examined 78 *Rider* elements (including solo LTRs) with perfect TSDs. It appears that *Rider* inserts into various sequences; however, the target sites were generally TA-rich. The GC profile of the 15 bp target sequence (including the 5 bp TSD and 5 bp surrounding sequence on each side) is shown in Figure 5, together with that of gene and non-gene sequences, and the target sequence of *Jinling*. Based on the distribution, the gene sequence was significantly more GC-rich than the non-gene sequence (mean GC content 41% versus 32%, $P < 0.001$, *t* test). The target sequence of *Jinling* largely overlapped with that of the non-gene sequence (mean GC content 31% versus 32%). The target sequence of *Rider* had an even lower GC content than the non-gene sequence (25%, $P < 0.001$, *t* test).

### Expression of *Rider* elements and read-through transcription

Database searches led to the identification of four ESTs matching the internal regions of *Rider* with 97% similarity or higher, and thus belonging to the *Rider* family. These ESTs were from leaf, shoot/meristem, and isolated leaf trichomes (accession numbers BG127728, ES894312, AI778038 and EX487903). To further investigate *Rider* expression, we performed semi-quantitative RT-PCR analysis of RNA isolated from various tissues (Figure 6a). In contrast to *SUN*, which was highly expressed in developing fruit, *Rider* was constitutively expressed in all tissues examined. To further test whether the products detected in RT-PCR represent intact transcripts from *Rider*, an mRNA blot was probed with the same internal region as used in RT-PCR. As shown in Figure 6(b), transcripts of approximately 4.5 kb were found, representative of full-length transcripts from intact *Rider* elements. The fuzzy bands may be indicative of the variable length of *Rider* LTRs. In addition, transcripts with a size similar to that of *SUN* (2.1 kb) were also observed. These transcripts might correspond to an internal deletion derivative of *Rider*, one of which was found on chromosome 8 (BAC accession number AP009268). Nevertheless, a considerable proportion of *Rider* transcripts appeared to be intact and should be capable of inducing transposition.

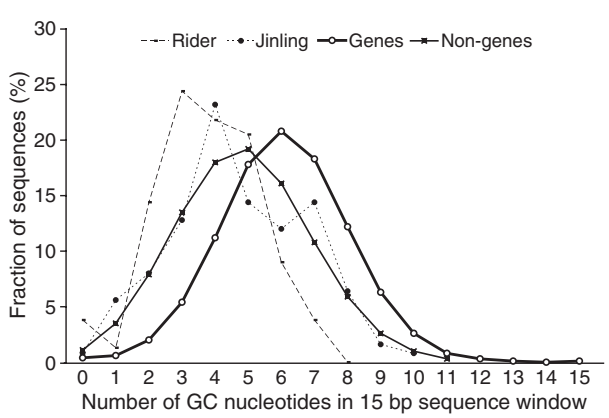


**Figure 5.** The GC content of the target sequences of *Rider*, *Jinling*, genic and non-genic sequences in a 15 bp window.

Two additional ESTs (accession numbers DB687421 and BF097567) contained LTR sequences that were 98 and 95% similar to *Rider*. Interestingly, both ESTs were chimeric in sequence, i.e. part of the sequences matched the *Rider* LTR but the remainder did not match the internal region of the element. These EST could be the result of read-through transcription or an aberrant transcription start in or near the *Rider* LTR. To investigate this possibility further, primers that would only generate RT-PCR products in the case of read-through transcription (i.e. transcripts that would not terminate in the R region and instead extend to the U5 region) were designed and used (Figure 4). We detected read-through transcription in all the tissues examined, even in mRNA samples that were purified twice using oligo(dT) (Figure 6c).

The *Rider* element that created the *sun* locus carried a mutation in one of the two TTGT sequences required for transcript termination (see above and Figure 4). To examine whether the mutation of TTGT to TTAT might influence the termination accuracy of *Rider* transcripts, we sequenced the RT-PCR products generated in Figure 6(c). As expected from database searches, few TTAT variants were recovered from genomic PCR (note the small green peak indicative of A in Figure 6d). However, the relative representation of TTAT compared to the TTGT variant was dramatically increased in all RT-PCR products analyzed, including those derived from mRNA (the signal for A is increased at the expense of G in Figure 6d). This strongly suggests that read-through transcripts are indeed associated with *Rider* elements and are more prevalent when the LTR carries the TTAT mutation in one of the TTGT copies in the U5 region. However, using locus-specific primers, read-through transcription

 

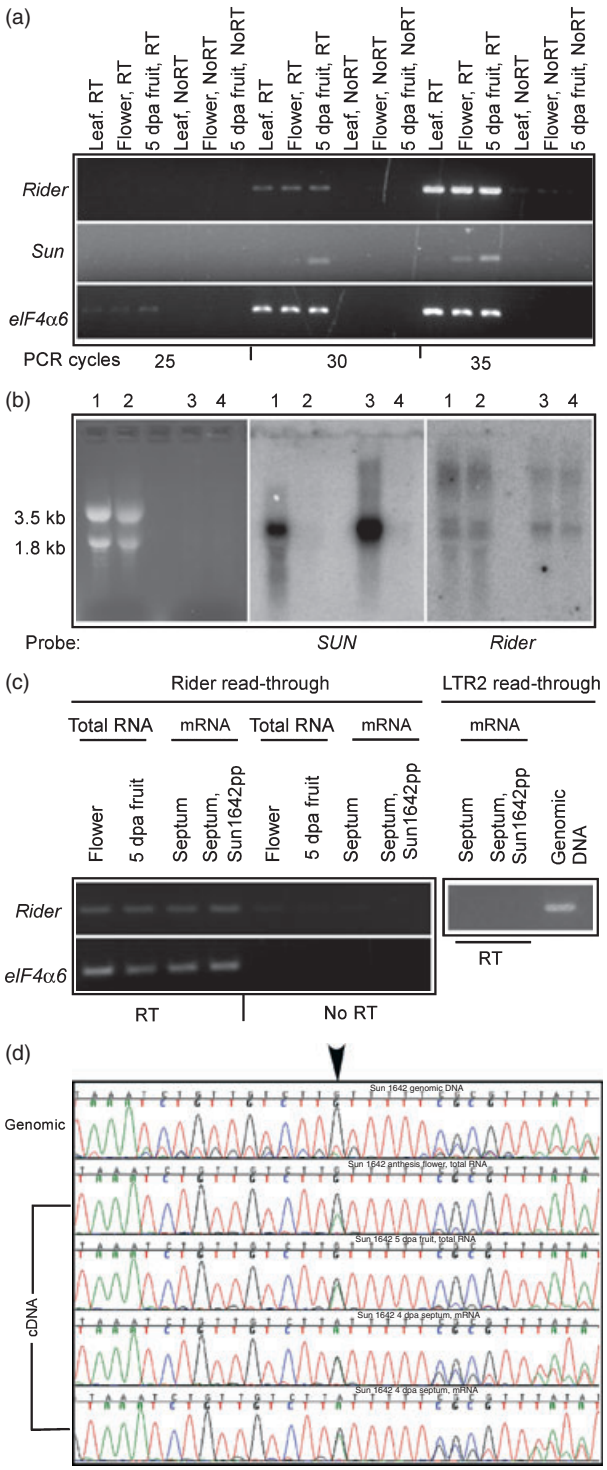


**Figure 6.** *Rider* expression and read-through transcription in Sun1642.
(a) Expression of *Rider* as determined by semi-quantitative RT-PCR of total RNA isolated from Sun1642 tissues indicated above the lanes.
(b) *Rider* expression revealed by Northern blot analysis. Lanes 1 and 3 contain RNA from septum tissues collected 4 days post-anthesis. Lanes 2 and 4 contain RNA from septum tissues collected 4 days post-anthesis from Sun1642pp without the *sun* mutation and *Rider* transposition on chromosome 7. Lanes 1 and 2, 10 μg of total RNA; lanes 3 and 4, 500 ng of mRNA. The left panel shows an ethidium bromide-stained RNA gel. The right two panels show the Northern blots probed as indicated.
(c) RT-PCR (35 cycles) using total and mRNA isolated from tissues indicated above the lanes. The left panel shows general read-through transcription and the right panel shows the lack of LTR2 read-through transcription.
(d) Sequence analysis of the PCR products generated in (c). The arrow indicates the location of the TTGT to TTAT mutation. Note the much higher A peak in the chromatogram of the RT-PCR products compared to the genomic PCR product.

from LTR2 at the *sun* locus and the ancestral locus on chromosome 10 was not detected (Figure 6c).

## DISCUSSION

Analyses of the *sun* locus and the ancestral locus, and comparisons to other euchromatic regions of the tomato genome, indicate that a significant proportion, 17–33%, of the gene-rich regions is comprised of TEs and other repeats. As shown in this and other plant studies, the most prevalent repeats in genic regions are small non-autonomous DNA elements (Feschotte *et al.*, 2002). The success of small DNA elements in genic regions may be attributed, at least in part, to their small sizes. However, despite their abundance, there is no evidence of recent transposition activity for DNA elements in tomato as opposed to RNA elements, which are more likely to have been active in recent times. The latter finding is consistent with the conclusion from a previous study that retrotransposon-based markers demonstrate a high degree of polymorphism in Solanaceae (Tam *et al.*, 2005).

Despite high sequence identity of the syntenic DNA sequence of the *sun* locus of Sun1642 and LA1589, there are many small- and large-scale insertions, deletions and transpositions in this region. Although the majority of these rearrangements involved intergenic regions and TEs, the most significant difference at the locus is created by *Rider*, which transposed a 24.7 kb gene-rich region from chromosome 10 (Xiao *et al.*, 2008). Thus, our analysis of the *sun* locus suggests that rearrangements such as indels and transpositions are likely to be more powerful forces in the genome evolution of tomato than point mutations. A variety of reasons might explain the frequent DNA rearrangements at the *sun* locus. On the one hand, the genes in this region are relatively small, and consequently 87% of the sequence in this region is intergenic. Moreover, many genes in this region belong to gene families, and therefore a mutation or interruption in one of these genes will probably not be lethal. For example, tomato cultivars that carry *Rider* at the *sun* locus, interrupting one of the defensin genes, appear to be normal except for the altered fruit shape. Accordingly, two factors were indispensable for the appearance of the novel fruit shape: (i) the target specificity of *Rider*, allowing insertion in genic regions and placement of *SUN* under the control of new regulatory elements, and (ii) tolerance of a gene knock-out insertion into *DEFL1*. On the other hand, the abundance of repeats, including the large number of U2-snRNA genes in

this area, may promote various types of illegitimate recombination, replication error or other rearrangements. For example, a crossover between the inverted repeat will lead to inversion of the entire internal region flanked by the repeat. Consequently, the intrinsic structural features of this locus may lead to a higher frequency of rearrangement, and the tolerance to mutations would allow individuals carrying the rearrangement to survive and spread in the population. From this point of view, the emergence of long-shaped tomato is an excellent example of evolution of novel phenotypes under relaxed selective constraint.

The newly discovered *Rider* retrotransposon is abundantly present in tomato and its wild relatives. Surveys of tomato and its relatives indicate that initial amplification of this element occurred prior to divergence of the tomato species but after the divergence of tomato and potato, which is estimated to have occurred between 5.1 and 7.3 MYA (Wang *et al.*, 2008). Additionally, the transposon display analysis among closely related accessions suggested that the main amplification wave of this element occurred after speciation. The absence of *Rider* elements from potato indicates that either the ancestor of *Rider* has been lost in this species, or *Rider* might have arisen in tomato via a horizontal transfer event (Cheng *et al.*, 2009).

LTR elements vary significantly in copy number, transposition activity and chromosomal distribution. Some elements, such as *Tos17* in rice, are currently active and insert into genic regions (Miyao *et al.*, 2003). Although *Tos17* can amplify rapidly under artificial conditions, there are only a few *Tos17* copies in the natural population of rice (Hirochika *et al.*, 1996). Other elements, such as *Jinling* of tomato, have thousands of copies in the genome and are mainly clustered in the heterochromatic regions (Wang *et al.*, 2006b). The newly discovered *Rider* element represents an unusual LTR retrotransposon that is different from most reported thus far with respect to its high copy number and insertion preference near to or in active genes. Moreover, given the large size of *Rider* (approximately 5 kb), it is interesting to speculate how this element reached such a high copy number in genic regions, and, based on transcription analysis, how it may still be active at the present time. Among other features, the target sequence preference of *Rider* may play an important role in its successful and largely unnoticed amplification through the genome of tomato. This target preference may reflect the fact that direct insertion of *Rider* into GC-rich coding regions, resulting in gene knock outs and possible deleterious effects, is not frequent.

In addition to its avoidance of coding regions, there are factors that may favor retention of *Rider* in or near genic regions. One apparent feature is its ability to duplicate and relocate genomic fragments (Xiao *et al.*, 2008) via its tendency to carry out read-through transcription (Figure 6c). The failure to detect read-through from LTR2 at the *sun* locus and the ancestral locus might be due to the fact that the level of transcripts from these specific loci are too low to detect. However, our analysis indicates that read-through transcripts of Rider LTRs are indeed present, and the over-representation of TTAT variant in the read-through transcripts seems to favor the possibility that this mutation promotes errors during mRNA termination in 3′ LTR. Alternatively, individual *Rider* elements with TTAT are more likely to be expressed, or expressed at higher levels, which could also lead to the over-representation of TTAT variants. The two hypotheses are not mutually exclusive. Despite the presence of various possibilities, the fact that the length of the U3 region varies dramatically between individual elements suggests that the replication in this region is error-prone, and that the structure of U3 region may affect the termination of RNA synthesis, promoting read-through and the incorporation of other genomic sequences into the transposition process, as was found to occur at the *sun* locus (Xiao *et al.*, 2008). These facts, combined with its abundance, recent amplification, transcription and association with genes, clearly suggest that *Rider* may be an important player in the evolution of the tomato genome.

## EXPERIMENTAL PROCEDURES

### Plant material, genomic DNA analyses and transposon display

The Sun1642 accession was obtained from Harris Moran (http://www.harrismoran.com/) and carries the *sun* locus duplication. The nearly isogenic line Sun1642pp lacks the duplication and *Rider* transposition on chromosome 7. M82, LA716, LA1589, LA1777, LA1927 and LA2560 were obtained from the Tomato Genetic Resource Center at the University of California at Davis. Howard German, Banana Legs, Rio Grande, Sausage and Roma were purchased from the Tomato Growers Supply Company (http://www.tomatogrowers.com/). DNA extractions and Southern blot analysis were performed as described previously (Bernatzky and Tanksley, 1986; Fulton *et al.*, 1995).

Transposon display was performed as described by Casa *et al.* (2000). Tomato DNA was digested with *Mse*I and ligated using adapters GACGATGAGTCCTGAG and ATGAGTCCTGAGTA. The element-specific primers used were RiderR1 for pre-amplification and RiderR2 for selective amplification (primer sequences are given in Table S3). The annealing temperature for pre-amplification was 56°C, and that for selective amplification was 58°C.

### Rider expression and read-through transcription

Total RNA was isolated using Trizol reagent (Invitrogen, http://www.invitrogen.com/) and mRNA was purified twice using a Poly(A) Purist™ MAG kit (Ambion, http://www.ambion.com). Northern blot analyses were performed using 10 μg total RNA and 500 ng mRNA. Templates for probe labeling with $^{32}$P-dCTP were prepared using primers EP388 and EP575 (*SUN*) and EP1181 and EP1182 (*Rider*).

RNA (5 μg total RNA or 100 ng mRNA) was DNaseI-treated and either reverse-transcribed using oligo(dT) as the primer and SuperScript III (Invitrogen) (RT) or incubated without SuperScript III (no RT; control). Of each 20 μl RT reaction mixture, 0.2 μl was subjected to PCR using primers EP1181 and EP1182 corresponding to *Rider* internal regions. To detect read-through transcription, primers EP1256 and EP1257 corresponding to the LTR U3 and U5 region

 

and flanking the TTGTCTTAT site were used. The PCR products were gel-isolated and sequenced. Primers EP1256 and EP1258 were used for the locus-specific read-through transcription from LTR2, with the former primer corresponding to the LTR and the latter to the region outside the LTR on chromosomes 7 and 10. Primers CME2F and EP826 were used to detect *SUN* expression by RT-PCR. Tomato *eIF4a6* was used to assess equal loading and mRNA and cDNA quality. Primer sequences are listed in Table S3.

### Annotations

The gene sequences were identified using the *ab initio* gene prediction program FGENESH (http://www.softberry.com/berry.phtml). Final annotation of the genomic sequences using National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov) and Institute for Genomic Research (http://www.tigr.org) databases was performed at the Institute for Genomic Research by the group of Dr Robin Buell (Department of Plant Biology, Michigan State University).

For identification of the repetitive sequences and TEs, the tomato genomic sequences were downloaded from the SOL Genomics Network (http://www.sgn.cornell.edu/) on 26 November 2007 (BAC sequences version: bacs.v177.seq) and used for estimation of copy number as well as determining the identity of the flanking sequences (see Appendix S2 for details on TE annotation). To determine the GC content of the various types of sequences, 'gene' sequences were generated by masking the tomato gene index sequences with RepeatMasker using the sequences of TEs, including known TEs and newly identified from this study. The 'non-gene' sequence was generated by masking the genomic sequence using 'gene' sequences. The potato BAC (8 Mb) and BAC end (108 Mb) sequences were provided by the group of Dr Robin Buell (Michigan State University) on 21 April 2008.

## ACKNOWLEDGEMENTS

This work was supported by funds from Ohio State University (H.X. and E.v.d.K.) and Michigan State University (N.J. and D.G.). We thank Dr Dan Voytas (Department of Genetics, Cell Biology and Development, University of Minnesota) for valuable discussions, and Drs David Mackey and David Francis (Department of Horticulture and Crop Science, Ohio State University) for comments on the manuscript. We also thank two anonymous reviewers for their comments and suggestions.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

**Table S1.** Identified and predicted genes.

**Table S2.** Transposable elements and unclassified repeats.

**Table S3.** Primers used in this study.

**Appendix S1.** Variables that may influence the estimate of *Jinling* copy numbers.

**Appendix S2.** Annotation of TEs.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

## REFERENCES

**Bennetzen, J.L.** (1996) The contributions of retroelements to plant genome organization, function and evolution. *Trends Microbiol.* **4**, 347–353.

**Bernatzky, R. and Tanksley, S.D.** (1986) Toward a saturated linkage map in tomato based on isozymes and random cDNA sequences. *Genetics*, **112**, 887–898.

**Bureau, T.E., White, S.E. and Wessler, S.R.** (1994) Transduction of a cellular gene by a plant retroelement. *Cell*, **77**, 479–480.

**Casa, A.M., Brouwer, C., Nagel, A., Wang, L., Zhang, Q., Kresovich, S. and Wessler, S.R.** (2000) Inaugural article: the MITE family heartbreaker (Hbr): molecular markers in maize. *Proc. Natl Acad. Sci. USA*, **97**, 10083–10089.

**Cheng, X., Zhang, D., Cheng, Z., Keller, B. and Ling, H.-Q.** (2009) A new family of Ty1-*copia*-like retrotransposon originated in the tomato genomes by a recent horizontal transfer event. *Genetics*, **181**, 1183–1193.

**Datema, E., Mueller, L.A., Buels, R., Giovannoni, J.J., Visser, R.G., Stiekema, W.J. and van Ham, R.C.** (2008) Comparative BAC end sequence analysis of tomato and potato reveals overrepresentation of specific gene families in potato. *BMC Plant Biol.* **8**, 34.

**Doebley, J., Stec, A. and Hubbard, L.** (1997) The evolution of apical dominance in maize. *Nature*, **386**, 485–488.

**Feschotte, C., Jiang, N. and Wessler, S.R.** (2002) Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* **3**, 329–341.

**Frary, A., Nesbitt, T.C., Grandillo, S., van der Knaap, E., Cong, B., Liu, J., Meller, J., Elber, R., Alpert, K.B. and Tanksley, S.D.** (2000) fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. *Science*, **289**, 85–88.

**Fray, R.G. and Grierson, D.** (1993) Identification and genetic analysis of normal and mutant phytoene synthase genes of tomato by sequencing, complementation and co-suppression. *Plant Mol. Biol.* **22**, 589–602.

**Fridman, E., Carrari, F., Liu, Y.S., Fernie, A.R. and Zamir, D.** (2004) Zooming in on a quantitative trait for tomato yield using interspecific introgressions. *Science*, **305**, 1786–1789.

**Fu, H. and Dooner, H.K.** (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl Acad. Sci. USA*, **99**, 9573–9578.

**Fulton, T.M., Chunwongse, J. and Tanksley, S.D.** (1995) Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Mol. Biol. Rep.* **13**, 207–209.

**Grandbastien, M.A., Spielmann, A. and Caboche, M.** (1989) Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. *Nature*, **337**, 376–380.

**Hirochika, H., Sugimoto, K., Otsuki, Y., Tsugawa, H. and Kanda, M.** (1996) Retrotransposons of rice involved in mutations induced by tissue culture. *Proc. Natl Acad. Sci. USA*, **93**, 7783–7788.

**Huang, X., Lu, G., Zhao, Q., Liu, X. and Han, B.** (2008) Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. *Plant Physiol.* **148**, 25–40.

**Jiang, N., Bao, Z., Zhang, X., Eddy, S.R. and Wessler, S.R.** (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, **431**, 569–573.

**Jin, Y.K. and Bennetzen, J.L.** (1994) Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs1 retroelement of maize. *Plant Cell*, **6**, 1177–1186.

**Juretic, N., Hoen, D.R., Huynh, M.L., Harrison, P.M. and Bureau, T.E.** (2005) The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Res.* **15**, 1292–1297.

**Kidd, J.M., Cooper, G.M., Donahue, W.F. *et al.*** (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.

**Lewin, B.** (2008) *Genes IX*, 9th edn. Sudbury, MA: Jones and Bartlett Publishers.

**Ling, H.Q., Bauer, P., Bereczky, Z., Keller, B. and Ganal, M.** (2002) The tomato *fer* gene encoding a bHLH protein controls iron-uptake responses in roots. *Proc. Natl Acad. Sci. USA*, **99**, 13938–13943.

**Liu, J., Van Eck, J., Cong, B. and Tanksley, S.D.** (2002) A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proc. Natl Acad. Sci. USA*, **99**, 13302–13306.

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

**McCarroll, S.A. and Altshuler, D.M.** (2007) Copy-number variation and association studies of human disease. *Nat. Genet.* **39**, S37–S42.

**Miyao, A., Tanaka, K., Murata, K., Sawaki, H., Takeda, S., Abe, K., Shinozuka, Y., Onosato, K. and Hirochika, H.** (2003) Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. *Plant Cell*, **15**, 1771–1780.

 

**Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A. and Rafalski, A.** (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.* **37**, 997–1002.

**Muse, S.V.** (2000) Examining rates and patterns of nucleotide substitution in plants. *Plant Mol. Biol.* **42**, 25–43.

**Nesbitt, T.C. and Tanksley, S.D.** (2002) Comparative sequencing in the genus *Lycopersicon*: implications for the evolution of fruit size in the domestication of cultivated tomatoes. *Genetics*, **162**, 365–379.

**Paran, I. and van der Knaap, E.** (2007) Genetic and molecular regulation of fruit and plant domestication traits in tomato and pepper. *J. Exp. Bot.* **58**, 3841–3852.

**Pickersgill, B.** (2007) Domestication of plants in the Americas: insights from Mendelian and molecular genetics. *Ann. Bot.* **100**, 925–940.

**SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L.** (1998) The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45.

**Tam, S.M., Mhiri, C., Vogelaar, A., Kerkveld, M., Pearce, S.R. and Grandbastien, M.A.** (2005) Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR. *Theor. Appl. Genet.* **110**, 819–831.

**Temin, H.M.** (1981) Structure, variation and synthesis of retrovirus long terminal repeat. *Cell*, **27**, 1–3.

**Van den Broeck, D., Maes, T., Sauer, M., Zethof, J., De Keukeleire, P., D'Hauw, M., Van Montagu, M. and Gerats, T.** (1998) Transposon display identifies individual transposable elements in high copy number lines. *Plant J.* **13**, 121–129.

**Van der Knaap, E., Sanyal, A., Jackson, S.A. and Tanksley, S.D.** (2004) High-resolution fine mapping and fluorescence *in situ* hybridization analysis of *sun*, a locus controlling tomato fruit shape, reveals a region of the tomato genome prone to DNA rearrangements. *Genetics*, **168**, 2127–2140.

**Wang, Y., van der Hoeven, R.S., Nielsen, R., Mueller, L.A. and Tanksley, S.D.** (2005) Characteristics of the tomato nuclear genome as determined by sequencing undermethylated *Eco*RI digested fragments. *Theor. Appl. Genet.* **112**, 72–84.

**Wang, W., Zheng, H., Fan, C. *et al.*** (2006a) High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell*, **18**, 1791–1802.

**Wang, Y., Tang, X., Cheng, Z., Mueller, L., Giovannoni, J. and Tanksley, S.D.** (2006b) Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome. *Genetics*, **172**, 2529–2540.

**Wang, Y., Diehl, A., Wu, F., Vrebalov, J., Giovannoni, J., Siepel, A. and Tanksley, S.D.** (2008) Sequencing and comparative analysis of a conserved syntenic segment in the Solanaceae. *Genetics*, **180**, 391–408.

**Wikstrom, N., Savolainen, V. and Chase, M.W.** (2001) Evolution of the angiosperms: calibrating the family tree. *Proc. Biol. Sci.* **268**, 2211–2220.

**Wu, F., Mueller, L.A., Crouzillat, D., Petiard, V. and Tanksley, S.D.** (2006) Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. *Genetics*, **174**, 1407–1420.

**Xiao, H., Jiang, N., Schaffner, E.K., Stockinger, E.J. and Van der Knaap, E.** (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science*, **319**, 1527–1530.

**Zhu, W., Ouyang, S., Iovene, M., O'Brien, K., Vuong, H., Jiang, J. and Buell, C.R.** (2008) Analysis of 90 Mb of the potato genome reveals conservation of gene structures and order with tomato but divergence in repetitive sequence composition. *BMC Genomics*, **9**, 286.

 