The Tvv1 retrotransposon family is conserved between plant genomes separated by over 100 million years

Moisy, Cedric
2014-05-01


http://hdl.handle.net/10138/155667
https://doi.org/10.1007/s00122-014-2293-z

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.
The *Tvvl* retrotransposon family is conserved between plant genomes separated by over 100 million years

Cédric Moisy · Alan H. Schulman · Ruslan Kalendar · Jan P. Buchmann · Frédérique Pelsy

Received: 2 September 2013 / Accepted: 21 February 2014 / Published online: 4 March 2014 © Springer-Verlag Berlin Heidelberg 2014

Abstract

**Key message** Combining several different approaches, we have examined the structure, variability, and distribution of *Tvvl* retrotransposons. *Tvvl* is an unusual example of a low-copy retrotransposon metapopulation dispersed unevenly among very distant species and is promising for the development of molecular markers.

**Abstract** Retrotransposons are ubiquitous throughout the genomes of the vascular plants, but individual retrotransposon families tend to be confined to the level of plant genus or at most family. This restricts the general applicability of a family as molecular markers. Here, we characterize a new plant retrotransposon named *Tvvl* _Sdem_, a member of the *Copia* superfamily of LTR retrotransposons, from the genome of the wild potato _Solanum demissum_. Comparative analyses based on structure and sequence showed a high level of similarity of _Tvvl_Sdem_ with _Tvvl_VB_, a retrotransposon previously described in the grapevine genome _Vitis vinifera_. Extending the analysis to other species by in silico and in vitro approaches revealed the presence of _Tvvl_ family members in potato, tomato, and poplar genomes, and led to the identification of full-length copies of _Tvvl_ in these species. We were also able to identify polymorphism in UTL sequences between _Tvvl_Sdem_ copies from wild and cultivated potatoes that are useful as molecular markers. Combining different approaches, our results suggest that the _Tvvl_ family of retrotransposons has a monophyletic origin and has been maintained in both the rosids and the asterids, the major clades of dicotyledonous plants, since their divergence about 100 MYA. To our knowledge, _Tvvl_ represents an unusual plant retrotransposon metapopulation comprising highly similar members disjointedly dispersed among very distant species. The twin features of _Tvvl_ presence in evolutionarily distant genomes and the diversity of its UTL region in each species make it useful as a source of robust molecular markers for diversity studies and breeding.

Introduction

Transposable elements (TEs) have been found in virtually all eukaryotic species investigated so far. They are divided into two classes based on their means of transposition and replication; the DNA transposons (Class II) move by a “cut and paste” mechanism whereas retrotransposons (Class I) use a “copy and paste” strategy (Wicker et al. 2007). The retrotransposons are ubiquitous in the plant kingdom and...
comprise most of the DNA of large genomes (Kidwell and Lisch 2001; Schnable et al. 2009; Wicker et al. 2009).

Of the four orders of retrotransposons, the Long Terminal Repeat (LTR) retrotransposons are the most prevalent in plant genomes; in animals, their relatives are the retroviruses and endogenous retroviruses (Wicker et al. 2007).

The LTRs are found at either end of the element and contain the transcriptional promoter and RNA and cDNA processing signals. The internal domain between the LTRs includes one or more open reading frames (ORFs). The gag gene encodes the structural protein that forms the virus-like particles (VLPs), whereas pol encodes a polypeptide containing asparagine protease (AP), integrase (INT), reverse transcriptase (RT) and RNase H (RH) domains. According to the organization of the coding domains in pol and their sequence similarities, LTR retrotransposons are divided into two superfamilies: Copia and Gypsy (Wicker et al. 2007). In Copia elements, integrase is upstream of the RT and RH domains, whereas in Gypsy elements and the retroviruses it is downstream.

The replicative lifecycle of LTR retrotransposons and retroviruses involves transcription from the 5′-LTR to the 3′-LTR followed by packaging of the RNAs into VLPs, their reverse transcription into cDNA and finally integration of the cDNA into the host genome as a new copy (Kumar and Bennetzen 1999; Sabot and Schultman 2006; Schultman 2013; Schultman and Wicker 2013). This life cycle is very prone to errors, estimated at $1.4 \times 10^{-5}$ mutations/bp/cycle, because both RNA polymerase II and reverse transcriptase lack proofreading activity (Gabriel et al. 1996; Preston 1996; Abran et al. 2010). Moreover, retrotransposon copies are replicated also as a part of the chromosome by DNA-dependent DNA polymerase; most copies are likely to accumulate mutations at the neutral rate. This enables their insertion time to be dated, because the LTRs in a newly integrated retrotransposon are identical but then diverge (SanMiguel et al. 1998; Vitte et al. 2007b; Moisy et al. 2008b). Internal structural heterogeneity can also arise from strand switching during replication (Vicient et al. 2005; Moisy et al. 2008a), or from recombination subsequent to integration (Sabot and Schultman 2007). More generally, the observation of highly conserved domains within RT, INT, and other retrotransposon regions needed for function during the replicative lifecycle indicates the action of purifying selection (Stuart-Rogers and Flavell 2001; Smith et al. 2004; Gómez et al. 2006; Baucom et al. 2009; Schultman and Wicker 2013).

The opposing processes of error-prone replication and the accumulation of mutations and rearrangements on the one hand and purifying selection for replicative success on the other lead to the genomic copies of a retrotransposon forming a population of closely related elements referred to as a “quasispecies” (Casacuberta et al. 1995; Ojosnegros et al. 2011). Retrotransposon quasispecies generally are named and treated as “families” based on sufficient similarity between aligned sequences. It was first proposed that two elements showing more than 90 % homology between motifs I to VII of RT belong to the same family (Bowen and McDonald 1999). More recently, Wicker et al. (2007) proposed the “80–80–80 rule”: two elements belong to the same family if they share at least 80 % sequence identity in at least 80 % of their coding or internal domain, LTRs, or in both regions on segments longer than 80 bp. Lineages, by contrast, have been defined as large groups of families that are on a common branch of an RT domain-based tree with high bootstrap value (>95), sharing characteristics such as LTRs of similar sizes, and having a general tendency toward either mostly high-copy or mostly low-copy families (Wicker and Keller 2007). Families themselves are sometimes divided into subfamilies corresponding to groups of elements sharing specific features, but elements very often display, depending on their population dynamics, a continuum of sequence variation, making it difficult to define clear limits for families and subfamilies. Retrotransposon quasispecies, moreover, form “metapopulations” (Hanks 1998), as populations within each genome of the plant population and species that interact through gene flow at some level.

Variability within different families of retrotransposons has been extensively investigated but most studies have focused on small regions between conserved domains that can be amplified by PCR, cloned and sequenced, such as within rt (Dixit et al. 2006; Nielen et al. 2010), gag (Tanskren et al. 2007), the LTR (Vicient et al. 2005), and the untranslated leader (UTL) between the LTR and gag (Pelsy 2007). With the increasing availability of plant genome sequences, the evolution of complete elements and their families within genomes now can be examined. The genomes of the rosids offer a good opportunity for this; at least 22 are either available or currently being sequenced. The rosids are a major group of the eudicot clade of flowering plants and comprise a quarter of the angiosperm species (Wang et al. 2009). The asterids, the other major clade of eudicots, separated from the rosids 93–110 MYA; the two major rosid clades, the Fabidae and the Malvidae, diverged 83–108 MYA (Wang et al. 2009). The Vitaceae, which includes cultivated grapes, appeared in the fossil record 60 MYA and are considered as a sister clade to all other Rosids (Jansen et al. 2006; Wang et al. 2009) or even as its own order, Vitales (The Angiosperm Phylogeny Group 2009).

The Vitis vinifera genome contains at least thirteen different families of LTR retrotransposons (Moisy et al. 2008b). The structural variability of one family, Tvv1 (Pelsy and Merdinoglu 2002) of the superfamily Copia was investigated (Pelsy 2007) and a canonical copy, Tvv1-VB, identified (Moisy et al. 2008a). Here, we characterize
Tvv1_Sdem, a member of the Tvv1 family from the genome of Solanum demissum, an asterid. We have examined its dispersion across the Solanum genus and analyzed relationships with the Tvv1 family beyond Solanum and Vitis. Our results provide evidence for the family’s monophyletic origin across the rosid–asterid divide.

**Materials and methods**

**Plant material and DNA isolation**

DNA was extracted from four Solanum demissum accessions (69S-152-105, 69S-154-105, 69S-168-102, and 69S-174-106), and from the grapevine line PN40024. Young expanded leaves of shoot tips were ground into fine powder with liquid nitrogen. Total DNA was extracted with the Qiagen Dneasy TM Plant Mini Kit (cat. 69104, Qiagen, Hilden, Germany) as described by the manufacturer. DNA samples for the following were also tested: ‘Yolo wonder’, an inbred line of sweet pepper (Capsicum annuum); ‘Vendor’, a tomato cultivar (Solanum lycopersicum); MM195, an eggplant cultivar (Solanum linnaeanum); ‘Rosa H1’, a dihaploid clone of potato (Solanum tuberosum 2x); Arabidopsis thaliana (Col1); and tobacco Nicotiana tabacum.

**PCR amplifications and sequencing**

Primers (sequences in Table S1, products in Fig. 1) were designed using Primer3 software (Rozen and Skaletsky 2000) from the sequence of Tvv1_Sdem. Product PCR1 extends from gag to rh, PCR2 from gag to the 3’ LTR, and PCR3 spans the UTL region. Primer LTRsd-utr was labeled at its 5′ end with IRD 800 5′ (Eurofins MWG Operon, Ebersberg, Germany). PCR amplifications were carried out according to Pelsy and Merdinoglu (2002). PCR fragments were resolved by electrophoresis in a LiCor 4000L automated DNA sequencer (Lincoln, NB) using IRD41-labeled M13 fragments (50 to 1,206 bp) as size standards. PCR fragments of interest were cloned, and two randomly selected recombinant clones were sequenced.

**Quantitative real-time PCR and relative quantification**

Copy numbers of Tvv1-like elements were determined through genomic quantitative PCR in V. vinifera, S. tuberosum, S. demissum, S. linnaeanum, S. lycopersicum, N. tabacum, P. trichocarpa, and C. annuum, using primers 4147 and 4155 (Table S1) designed to match the Tvv1_Sdem gag-ap region. PCR reactions were performed in 15 μL containing: 7.5 ng DNA, 1× Phire® Reaction Buffer (including 2.5 mM MgCl2), 300 nM each primer, 200 μM dNTP, 0.15 μl Phire Hot Start II DNA Polymerase (Thermo Scientific), 0.5× SYBR Green I (Cambrex)
Bio Science Rockland, Inc). The amplification program, run on a LightCycler® 480 Real-Time PCR System (Roche), consisted of: 98 °C, 2 min; 30 cycles of 10 s at 98 °C, 10 s at 55 °C, 15 s at 72 °C. All measurements were repeated four times.

Genome databases and sequence searches

To find Tvv1-related sequences in other plant genomes, searches were conducted in the NCBI plant databases (http://www.ncbi.nlm.nih.gov/) by BLASTn 2.2.24+ (Zhang et al. 2000) with default parameter settings using Tvv1-VP sequence (EU304807) and Tvv1-S_dem as the queries. In addition, 41 other plant genome databases and the TE database Repbase (Kapitonov and Jurka 2008; http://www.girinst.org/) were queried as well (Table S2). When significant similarities (>70 % identity, >15 % query coverage) were found, the regions, including 10 kb of flanking sequence on each side, were excised in silico from the corresponding genomic sequences and screened for the presence of Tvv1-related sequences. Individual Tvv1-related sequences were annotated for the position of LTRs, UTL, and ORFs using DNAsis 2.1 (Hitachi Software Engineering Co, Ltd) and FastPCR (Kalander et al. 2011).

Sequence analysis

Nucleic sequences were aligned using EMBL-EBI Pairwise Sequence Alignment tools (http://www.ebi.ac.uk/) and MEGA Version 5 (Tamura et al. 2011). The hypothetical protein sequences were deduced by comparison with proteins from other retrotransposons (Tvv1-VP, Tvv1_Sdem, Tnt1, PDR1) and framenashes were removed manually to obtain the putative amino acid sequences of the Nucleic Acid Binding Domain (NABD), INT, AP, RH, and RT. Amino acid sequences were aligned using MEGA Version 5 (Tamura et al. 2011) and figures were produced using Geneious Version 6.1.6 (Biomatters, http://www.geneious.com/). Phylogenetic and molecular evolutionary analyses based on amino acid sequences spanning reverse transcriptase motifs III to V were carried out using MEGA Version 5 (Tamura et al. 2011). LTR sequences were aligned using “needle”, which is from the eMBOSS package (Rice et al. 2000), and MUSCLE (Edgar 2004). All insertion and divergence estimates for the 5′ and 3′ LTR pairs as well as for individual LTRs were calculated as previously described (SanMiguel et al. 1998; Vitte et al. 2007b; Wicker and Keller 2007; Moisy et al. 2008b), applying a substitution rate of 1.5 × 10−8 substitutions per site per year (Koch et al. 2000). All accession numbers of plant retrotransposon sequences used in this study are given in Table S3.

Results

Tvv1_Sdem, a new LTR retrotransposon in Solanum demissum

A new retrotransposon of superfamily Copia, named Tvv1_Sdem and having the classification code RLC (Wicker et al. 2007), was identified in a BAC clone (Accession AC149291, nt 13231 to 18051) from wild potato or nightshade (Solanum demissum) by BLASTn, using grape retrotransposon Tvv1-VP (EU304807; Moisy et al. 2008a) as the query. The complete sequence of Tvv1_Sdem, 4,821 bp, comprises an internal region of 4,461 bp flanked by two short, 180 bp LTRs (Fig. 1). The two LTR sequences only differ by three nucleotides (98.3 % identity) and are flanked by perfect 5 bp direct repeats (CTCGA). Based on the sequence divergence between both LTRs and a molecular clock applied earlier (Vitte et al. 2007b), the insertion date of this particular copy was estimated at 0.65 MYA. The internal region of this element contains an untranslated leader (UTL) of 214 bp between the 5′LTR and the protein coding domain. Immediately downstream of the 5′LTR, the sequence 5′ TGTTATCAGAGCC 3′ comprises the Primer Binding Site (PBS) complementary to tRNAMet. Immediately upstream to the 3′LTR a polyurine tract (PPT), 5′ TGAGGGGGAGG 3′ was found. In between the PBS and the PPT, required for cDNA synthesis in LTR retrotransposons, a large ORF with a single frameshift is present (see below).

Comparison of asterid Tvv1_Sdem to rosid Tvv1-VP from V. vinifera

Tvv1_Sdem was compared to the canonical full-length copy Tvv1-VP previously described in grapevine (Moisy et al. 2008a). Tvv1_Sdem and Tvv1-VP, 4,821 and 5,222 bp, respectively, share 61 % identity overall. Tvv1_Sdem LTRs (180 bp) are longer than the identical Tvv1-VP LTRs (149 bp) but remain in the same length range as LTRs of Tvv1 elements previously described in grapevine (149 to 198 bp; Moisy et al. 2008a). Despite their differences in length, LTRs of Tvv1_Sdem and Tvv1-VP show 46 % identity overall, with higher conservation in blocks (Fig. S2). The U3 region, which contains the promoter in LTR retrotransposons, is more divergent than the R region. As described for Tvv1 in grape (Pelsy and Merdinoglu 2002), the TATA box and the polyadenylation site are also present in Tvv1_Sdem, but appear more degenerated than in Tvv1-VP.

The internal, protein-coding domain of the Tvv1_Sdem from the BAC clone contains a deletion of 35 bp (beginning at nt 4389, before rh) with respect to Tvv1-VP, which generates a +1 frameshift and likely a truncated…
The similarity of the asterid *Tvv1_Sdem* to *Tvv1*, a family otherwise known only from the Vitaceae, which is a sister group to all other rosids, prompted us to look in more detail at *Tvv1* members in the asterids and mainline rosids. We took three approaches: PCR amplification of major segments of *Tvv1* elements from related asterids, particularly in the Solanaceae; in silico searches of sequences available from genome projects; copy number determination by genomic quantitative PCR.

PCR amplifications of *Tvv1* were carried out on genomic DNA from four *S. demissum* accessions, diploid potato (*S. tuberosum*), grapevine (*V. vinifera*), sweet pepper (*C. annuum*), tomato (*S. lycopersicum*), eggplant (*S. linnaeanum*), and tobacco (*N. tabacum*) from the Solanaceae as well as on Arabidopsis (*A. thaliana*). Three primer pairs were used, yielding three expected products based on *Tvv1_Sdem*: PCR1, 3,300 bp, spanning *gag–rh*; PCR 2, 3,600 bp, spanning *gag–3′ LTR*; PCR3, 419 bp, extending from the 5′ LTR to the *gag* (Figs. 1, 2a; Table S1).

In the four *S. demissum* accessions, PCR1 amplified a fragment of expected size; three yielded as well an extra fragment of ~1,800 bp (Fig. 2a), probably due to the presence of a truncated copy as described in Moisy et al. (2008a) for grapevine. Two PCR1 products from *S. tuberosum* and *S. lycopersicum* were sequenced and showed the expected size (~3,300 bp). The potato sequences were 94 % identical to each other, and showed, respectively, 93 and 95 % identity with *Tvv1_Sdem*. The tomato sequences were 99 % identical to each other and 86 % to *Tvv1_Sdem*. Consistent with these results, the expected PCR2 fragment was detected in all *S. demissum* accessions, *S. tuberosum*, and *S. lycopersicum* (Fig. 2a). Together, the data establish that *Tvv1* is present in at least three *Solanum* species and has a conserved organization from the *gag* to the 3′ LTR.

The expected 419 bp fragment, corresponding to an UTL sequence of 214 bp, was amplified from all potato DNAs except from *S. demissum* accession 69S-154-105, the one that had displayed a deletion for the PCR1 region as well (Fig. 2b, lanes 1 to 5). Additional polymorphic fragments were also detected; two, 490 and 500 bp, were amplified from all *S. demissum* accessions but not from *S. tuberosum*; two, 440 and 455 bp, were observed only in *S. tuberosum*. Finally, a 710 bp fragment was observed only in three *S. demissum* accessions while absent from 69S-174-106 and from *S. tuberosum*. Larger bands of over 1 kbp are also present, and were previously observed for amplification of *Tvv1* UTLs in grapevine (Pelsy 2007). Forty-two UTL regions from *S. demissum* (28) and *S. tuberosum* (14) were cloned and sequenced, ranging from 210 to 505 bp. Aligned together with *Tvv1* UTLs from grapevine used as the outgroup, the sequences did not cluster according to their genome of origin (Fig. 2c); one *S. demissum* (Sd_211) and one *S. tuberoussum* (rosa_56) UTL sized 212 bp were even 100 % identical.

Several UTLs of *S. demissum* were almost identical to *Tvv1_Sdem*, only varying in the length of an internal T₃ microsatellite. The comparison of the UTL sequences indicates that at least 14 and 10 different groups of *Tvv1* elements are present, respectively, in the genomes of wild and cultivated potatoes (Fig. 2c).

More broadly in the Solanaceae, PCR2 and PCR3 yielded no products of the expected size from pepper, tomato, eggplant, Arabidopsis, or tobacco. PCR3 only resulted in weak amplifications of ~500 bp fragments (Fig. 2a).

Distribution of *Tvv1_Sdem* by in silico analysis

PCR amplifications can fail due to mismatches over a very few nucleotides, even if sequences related to the target are present in the query genomes. Therefore, we searched available sequence resources in silico with *Tvv1_Sdem* as the query (List of genomes in Table S2). Starting within the Solanaceae, in *S. tuberosum* databases many matches were found in searches of the assembly of chromosome pseudo-molecules (12 hits), scaffolds (202 hits), and unanchored scaffolds (14 hits). Scaffolds datasets are generally rich in redundant sequences such as TEs that are difficult to anchor onto the physical map, resulting in a higher number of hits than in the pseudomolecules.

Coincidentally, the 12 best hits on the pseudomolecule assemblies matched each of the 12 potato chromosomes. This indicates that the *Tvv1* family is distributed widely within the potato genome but probably in low copy number. The best match covered 87 % of *Tvv1_Sdem* with 97 % identity and was located on the chromosome 4 of...
S. tuberosum Group Phureja. It led to the identification of one particular element, Tvv1_Stub (HE614294; Table 1), which is a complete copy flanked by two 5 bp direct target-site duplications (TSDs). The whole sequence of Tvv1_Stub shows more than 95% identity with Tvv1_Sdem and can be considered as a member of the Tvv1 family.

For tomato, searches in the genome assembly gave 30 significant hits. One copy, Tvv1_Slyco (HE647701), was retrieved from the best match on chromosome 12. It is surrounded by imperfect 5 bp TSDs and contains LTRs having 90% identity, 6 mismatches and 2 short gaps (Table 1). The ORF is interrupted by several stop codons, making difficult to reconstruct the polyprotein in silico. The whole sequence of Tvv1_Slyco shows 81% of identity with Tvv1_Sdem and can be considered a Tvv1 family member in the cultivated tomato S. lycopersicum, according to the 80–80–80 rule (Wicker et al. 2007). Analysis of the currant tomato S. pimpinellifolium database gave 31 hits. The best
match led to the identification of *Tvv1_Spimp* (HE647700), a complete copy with LTRs showing 91 % identity, flanked by two 5 bp perfect TSDs (Table 1). Its internal sequence is interrupted by several stop codons. The whole sequence of *Tvv1_Spimp*, including LTRs and UTL, shows 81 % identity with *Tvv1_Sdem* and is thereby a member of the *Tvv1* family. The potato and tomato insertions can be dated to 150 Maunouri, InrA, personal communication). Based on BLAST searching were investigated (Table S2). One full-length *Tvv1_Sdem* was also conducted in a specialized database containing retrotransposon sequence. Moreover, a BLAST search was retrieved (Table 1). *Tvv1_Sdem* is closer to *V. vinifera* previ-
ously described (Moisy et al. 2008b).

Among the rosid Fabidae, we searched the genome assembly of poplar (*Populus trichocarpa*) using *Tvv1_Sdem* as the query. Only 4 hits were found, with a maximum 15 % of query coverage and 73 % identity. However, one copy of 3,919 bp, *Tvv1_Pt-pop004-B08* (Accession AC214995), was retrieved (Table 1). *Tvv1_Pt-pop004-B08* shows perfect 5 bp TSDs (AAAAA), and two identical LTRs making it a likely recent insertion (less than 0.3 MYA), but carries a deletion bridging the *gag* and *ap* domains that would render these proteins non-functional. In the Repbase database, a retrotransposon from *P. trichocarpa*, named *Copia39-PTR* (3,941 bp) is otherwise identical to *Tvv1_Pt-pop004-B08* except for a 22 bp indel (Table 1), and bears the same large internal deletion. Its Repbase annotation indicates a UTL of 1,045 bp, but the many stop codons make it difficult to identify the precise *gag–pol* region. Nevertheless, the overall similarity to *Tvv1_Sdem* places these elements in the *Tvv1* family. Because the Repbase accessions do not include the TSDs or flanking sequences, it is not possible to ascertain whether or not *Tvv1_Pt-pop004-B08* and *Copia39-PTR* are the same copy or two highly conserved but distinct members of *Tvv1*.

Other Fabidae and Malvideae rosid species for which finished or ongoing genome sequences are available for BLAST searching were investigated (Table S2). One full-length relatively divergent *Tvv1* element (*Tvv1_Cpap*, Table 1) was found from papaya (*Carica papaya*), a member of the Malvideae. Other matches were limited to small segments and cannot be attributed to more than generically conserved domains of LTR retrotransposons.

We also have tested 34 other genomes (Table S2), including *Brachypodium distachyon*, *Coffea arabica*,

---

**Table 1** *Tvv1* in investigated genomes (Accession numbers in Table S3)

<table>
<thead>
<tr>
<th>Name</th>
<th>Genome</th>
<th>Total length (bp)</th>
<th>LTRs</th>
<th>Insertion date (MYA)</th>
<th>UTL (bp)</th>
<th><em>gag–pol</em> (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Tvv1_Cpap</em></td>
<td><em>C. papaya</em></td>
<td>5,193</td>
<td>159/198</td>
<td>85.4</td>
<td>4,383</td>
<td>179</td>
</tr>
<tr>
<td><em>Copia39-PTR</em></td>
<td><em>P. trichocarpa</em></td>
<td>3,941</td>
<td>225</td>
<td>100.0</td>
<td>1,145</td>
<td>2,046</td>
</tr>
<tr>
<td><em>Tvv1_Pt-pop004-B08</em></td>
<td><em>P. trichocarpa</em></td>
<td>3,919</td>
<td>225</td>
<td>100.0</td>
<td>1,145</td>
<td>2,046</td>
</tr>
<tr>
<td><em>Tvv1_Spimp</em></td>
<td><em>S. pimpinellifolium</em></td>
<td>4,884</td>
<td>150/151</td>
<td>91.4</td>
<td>2,457</td>
<td>4,276</td>
</tr>
<tr>
<td><em>Tvv1_Sdem</em></td>
<td><em>S. demissum</em></td>
<td>4,821</td>
<td>180</td>
<td>98.3</td>
<td>214</td>
<td>4,461</td>
</tr>
<tr>
<td><em>Tvv1_Slyco</em></td>
<td><em>S. lycopersicum</em></td>
<td>4,941</td>
<td>148/145</td>
<td>90.0</td>
<td>398</td>
<td>4,259</td>
</tr>
<tr>
<td><em>Tvv1_Stub</em></td>
<td><em>S. tuberosum</em></td>
<td>4,916</td>
<td>180</td>
<td>96.1</td>
<td>276</td>
<td>4,276</td>
</tr>
<tr>
<td><em>Tvv1_40024-Sd</em></td>
<td><em>V. vinifera</em></td>
<td>4,957</td>
<td>179</td>
<td>96.6</td>
<td>466</td>
<td>4,123</td>
</tr>
<tr>
<td><em>Tvv1-VB</em></td>
<td><em>V. vinifera</em></td>
<td>5,222</td>
<td>149</td>
<td>100.0</td>
<td>768</td>
<td>4,146</td>
</tr>
<tr>
<td><em>Tvv1 family</em></td>
<td><em>V. vinifera</em></td>
<td>4,640–7,162</td>
<td>149–198</td>
<td>79.4–100.0</td>
<td>179–474</td>
<td>3,863–6,114</td>
</tr>
</tbody>
</table>

---

*a* From Repbase

*b* From the 23 full-length *Tvv1* copies identified in the grapevine genome (Moisy et al. 2008b)
Medicago truncatula, Oryza sativa and Zea mays assemblies, but no Tvv1 members were found.

**Tvv1 fingerprints**

Alignments of the hypothetical polyprotein domains NABD, AP, PR, INT, RT, and RH (Fig. S1) have been established from the nucleic sequences of Tvv1-like elements, and frameshifts were removed manually by comparison with other polyproteins. The results support Tvv1_Cpap, Copia39_PTR, Tvv1_Pt-pop004-B08, Tvv1_Sdem, Tvv1_Slyco, Tvv1_Spimp, Tvv1_Sub, Tvv1_40024-Sd, and Tvv1-VR as being members of the Tvv1 family, which is clearly distinct from any other retrotransposon family here-tofore described. Despite divergence of distant Tvv1-like copies, numerous residues specific to Tvv1 members have been identified in each functional domain (online resource Fig. S1), and can therefore be considered as fingerprints of use for diagnostic identification of this family.

**Phylogenetic position of Tvv1_Sdem**

An alignment was made of the RT domains (between motifs III to V) of all Tvv1 elements identified in this study. Other retrotransposon sequences previously identified in Solanaceae species and in other plants were added to it (see Table S3 for accession numbers). A neighbor-joining tree constructed from the alignment (Fig. 3) shows that the Tvv1 family forms a distinct clade with a high bootstrap value (80). Tvv1_Sdem is closer to grapevine Tvv1 than to any other retrotransposon previously identified in Solanum. The Tvv1 elements identified in potato species (Tvv1_Sdem and Tvv1-Sub) and in wild tomato (Tvv1_Spimp), all from Solanum, are closely related to each other. Tvv1_Slyco is found slightly apart, as is the more divergent Tvv1_Cpap copy from papaya; those two sequences are more degenerate than the other Tvv1 copies. However, Tvv1_40024-Sd from grapevine is more closely related to those of potato, tomato, and poplar than to the other Tvv1 copies from grapevine, which are on a strongly supported node (bootstrap value of 96).

The two elements Copia39_PTR and Tvv1_Pt-pop004-B08 from poplar, a rosid as is grapevine, cluster apart from...
the other Tvv1 family members (bootstrap value 100). These two sequences may represent a distinct Tvv1 line, or may derived from a single degenerate copy; whereas other Tvv1-like segments can be found in the poplar genome (data not shown), none are sufficiently long to be incorporated into alignments. The Tvv1 family is distantly connected to four other sequences from Glycine, Medicago and grapevine (Fig. 3), forming a branch with a high bootstrap value (98), which could belong to the same distant lineage according to Wicker and Keller (2007). However, those four sequences do not share strong identity with Tvv1 apart from their RT domain, and therefore were not considered as members of the Tvv1 family.

Copy number estimation

To complement BLAST and classical PCR experiments, copy numbers of Tvv1-like elements were determined in several plant genomes through quantitative PCR using specific primers (Table S1) located in the gag-ap region. Results (data not shown) support Tvv1 being absent from C. annuum, S. linnaeanum, A. thaliana, and N. tabacum. Conversely, Tvv1 is rare in S. lycopersicum (9 copies) and P. trichocarpa (9 copies); it is of intermediate abundance in S. tuberosum (36 copies) and in the four S. demissum accessions (80–100 copies). As expected, the highest copy number (260) of Tvv1 was observed in V. vinifera PN40024. The ranges of Tvv1 copy numbers estimated by qPCR are consistent with the numbers deduced from BLAST searches and previous findings in grape (Moisy et al. 2008b). They also strongly corroborate our assumption that there are at least 14 and 10 different Tvv1 copies in wild and cultivated potatoes, respectively, based on comparisons of UTL sequences (Fig. 2c).

Divergence time of Tvv1 copies

If all the Tvv1 copies evolved from a common and unique ancestral copy, they should have had the same original founder lTrs that would have diverged over time. The lTr sequences of all the Tvv1 copies (Table 1, Fig. S3) were compared to estimate the probable time of divergence (Fig. 4) based on the calculation of the pairwise distance between the lTRs (Vitte et al. 2007b; Moisy et al. 2008b). The data indicate that Tvv1 diverged between grapevine, poplar, and Solanum first, then between potato and tomato, and finally independently between wild and cultivated species of these two Solanaceae.

Discussion

Tvv1 is a metapopulation of retrotransposons present across the asterid–rosid divide

Tvv1 was first characterized as a well-conserved family of LTR retrotransposons of superfamily Copia (classification code RLC; Wicker et al. 2007) in grapevine (Pelsy and Merdinoglu 2002; Moisy et al. 2008a). Here, we have identified and characterized a member of this family, Tvv1_Sdem, from the phylogenetically distant wild potato S. demissum. The Tvv1 family thus bridges the major
phylogenetic division in the dicotyledonous plants separating the rosids (Vitis) from the asterids (Solanum). While this particular copy of \textit{Tvv1}_Sdem is no longer autonomous because its coding region contains a frameshift in the \textit{rh} domain, all the functional domains for retrotransposition are present and well conserved between the potato and grapevine members.

The family shares short, 180 bp LTRs, numerous diagnostic residues in the polyprotein, and a fairly low copy number. Besides those of the non-autonomous TRIMs (Witte et al. 2001), LTRs of less than 200 nt appear to be fairly rare among plant retrotransposons; the best characterized example is \textit{PDR1} of pea, having 156 bp LTRs (Vershinin and Ellis 1999). The potato and grapevine elements display 69 % identity overall, with the highest conservation in the enzymatic regions (INT, 79 %; RT, 81 %) compared to the structural Gag (67 %), consonant with general trends for retrotransposons and retroviruses (Suoniemi et al. 1998; Eickbush and Jamburuthugoda 2008; Llorens et al. 2009). The UTLs of retrotransposons generally show high diversity, making them suitable for molecular markers (Pelsy 2007; Macas et al. 2009).

Based on the UTL diversity, at least 14 and 10 distinct \textit{Tvv1} copies are present in the genome of wild and cultivated potatoes, respectively. This trend was confirmed by quantitative PCR, attesting that wild accessions contain more \textit{Tvv1} copies than the cultivated one. This difference may be due either to founder effects or those of selection in the domestication of potato. Overall, \textit{Tvv1} UTLs in the Solanaceae have diverged mainly by accumulation of indels as was previously described for \textit{Tvv1} copies in grapevine (Pelsy 2007).

Interestingly, UTL sequences did not necessarily cluster based on their origin; some were perfectly identical between wild and cultivated potatoes, although markers based on the \textit{Tvv1} UTL are more polymorphic between \textit{Vitis} species than are microsatellites (Pelsy 2007). The UTL data indicate that \textit{Tvv1} copies in \textit{Solanum} have most probably arisen from several ancestors. The dispersion of the \textit{Tvv1} family among evolutionarily distant genomes belonging to both rosids and asterids provides the basis for developing a common system of molecular markers. Retrotransposon UTL Polymorphism (RUP, Pelsy 2007) for \textit{Tvv1}, moreover, can serve to describe genetic diversity in distant species and, combined with LTR-based dating of individual insertions, helps to retrace the evolutionary history of the \textit{Tvv1} family in genomes. The usefulness of \textit{Tvv1} as a molecular marker in disparate species is rare for a retrotransposon-based system.

Extending the analysis to other species among asterids, rosids, and also monocotyledonous plants by both in silico and in vitro approaches revealed the presence of \textit{Tvv1} members in both asterid (potato and tomato) and rosid (poplar and grapevine) genomes, and led to the identification of full-length copies of \textit{Tvv1}-like sequences in these species. Excepting \textit{Tvv1}_{Pt-pop004-808} from poplar, which shows a large deletion but retains \textit{Tvv1}-specific motifs, and the more divergent copy \textit{Tvv1-Cpap} from papaya, all other \textit{Tvv1} relatives belong to the same family according to the 80–80–80 rule proposed by Wicker et al. (2007). Compared against the \textit{Tnt1}-like elements within \textit{Solanum} (Manetti et al. 2009) that show specific amino acid sequences conserved among \textit{Tnt1}, \textit{TLC1}, and \textit{Retrosol1}, we have identified several specific motifs that can be considered as fingerprints of the \textit{Tvv1} family. These signatures, together with the detection of several distinct copies through in silico and in vitro approaches, establish that the \textit{Tvv1} family forms a metapopulation of retrotransposons that is unusually distributed among the dicotyledonous plants.

The \textit{Tvv1} family, while widely dispersed, does not appear to be ubiquitous. Here, \textit{Tvv1} members were detected in \textit{Populus}, but not in \textit{Medicago}, although both of them belong to the clade Fabidae of the rosids. No \textit{Tvv1}-like elements were found in \textit{Arabidopsis}, which belongs to the Malvidae, the other major lineage of rosids, and a relatively divergent element was found in papaya. In the asterids, \textit{Tvv1} members were found in potato and tomato, but no full-length copy could be identified in tobacco. Although comparative sequence analyses indicate that \textit{Coffea} (asterids) and \textit{Vitis} (rosids) derive from the same paleohexaploid ancestral genome (Cenci et al. 2010), \textit{Tvv1} was not detected in the coffee genome database.

Phylogenetic analysis of \textit{Copia} elements from barley, wheat, rice, and Arabidopsis established that they can be classified into six ancient lineages that existed before the divergence of monocots and dicots, 150 MYA (Wicker and Keller 2007). Lineages were defined as large groups of families that, within an RT-based tree, are found on a common branch having a high bootstrap value, share characteristics such as LTRs of similar sizes, and have a general tendency to contain mostly high-copy or mostly low-copy families (Wicker and Keller 2007). Surprisingly, \textit{Tvv1} as a family is widely dispersed within the plant kingdom; it has an origin as ancient (>100 MY) as the six lineages themselves (Wicker and Keller 2007). This has not been described for any of the individual families from those six lineages. Hence, \textit{Tvv1}, unusually, is both more conserved and less ubiquitous than retrotransposon families heretofore described.

The apparently spotty distribution of \textit{Tvv1} could be due to three non-exclusive causes: analytical limitations; horizontal transfer to only some plant clades; vertical descent combined with stochastic loss. Concerning experimental limitations, efforts to isolate \textit{Tvv1} by PCR could have been hampered by sequence divergence; even minor sequence variation at the priming sites could have blocked amplification. On the other hand, our PCR data were confirmed using four different sets of primers located in domains...
showing variable levels of conservation among TEs (LTR, gag, UTL, ap and rt), strengthening the conclusions. Conversely, in silico searches of sequenced genomes, while not sensitive to local divergence, are at the mercy of the quality of the genome assemblies and incorporation of repetitive sequences therein (Wicker et al. 2006; Otto et al. 2008). We restricted our definition of \( Tvv1 \) presence to the 80–80–80 rule described above; relatives of \( Tvv1 \) of arbitrary degrees of divergence (as members of lineages) nevertheless may have been widespread.

The second scenario explaining the distribution of the \( Tvv1 \) metapopulation over plant genomes assumes at least one horizontal transfer event such as that suggested for several other plant retrotransposons (Wicker and Keller 2007; Fortune et al. 2008; Roulin et al. 2008; Cheng et al. 2009; Roulin et al. 2009). To match the phylogeny, we must assume that the \( Tvv1 \) family would have arisen in the common ancestor of poplar and grapevine and later been horizontally transferred into the potato and tomato ancestor after the divergence of the rosids. At present, however, the available data are insufficient to confirm a claim of horizontal transfer. The low copy number of \( Tvv1 \) in the species harboring members of this family helps to explain its sporadic occurrence. Moreover, the \( Tvv1 \) family is found in several asterid and rosid clades, in contrast to the proposed horizontally transferred Rider element (Cheng et al. 2009), necessitating for \( Tvv1 \) either more than one horizontal transfer or a combination of horizontal transfer and vertical descent. Furthermore, the pattern of divergence within the LTRs of \( Tvv1 \) matches the species phylogeny, requiring secondary hypotheses about adaptive evolution within the LTRs. More high-quality genome sequences will be needed for a better picture of \( Tvv1 \) distribution.

Alternatively, \( Tvv1 \) could have been passed vertically through the descendants of the last common ancestor of the rosids and asterids. In this interpretation, \( Tvv1 \) copies would have been positively selected over time, at least in particular species, to maintain the high level of similarity observed, considering that \( Viis \) and \( Solanum \) have diverged for 100 MY or more. In other species, the \( Tvv1 \) family would have been lost stochastically as proposed for quasispecies or metapopulations generally (Hanski 1998; Domingo 2002) or more specifically for TEs (Le Rouzic et al. 2007) such as Mariner in \( Drosophila melanogaster \) (Lohe et al. 1995). Such a loss is even more probable given the low copy number of \( Tvv1 \), even if observations for \( Copia \) lineages primarily in the grasses suggest that individual families become extinct only very rarely in a species (Wicker and Keller 2007). In support of vertical radiation, pairwise comparisons of \( Tvv1 \) LTRs indicate that their pattern of similarity matches the successive divergence of the asterids and rosids followed by that of the clades within the rosids and the species within \( Solanum \). Nevertheless, the analysis may be biased by possible differing degrees of transpositional activity in the various species and clades, because for retrotransposons the error rates for replication via reverse transcription and as integral chromosomal components differ greatly (Gabriel et al. 1996; Boutabout et al. 2001; Abram et al. 2010).

If \( Tvv1 \) has propagated vertically, the question that arises is how a retrotransposon family having such low numbers of copies has persisted over 100 MY. Conservation within the family suggests that purifying selection has been at work. Whereas genetic drift takes its toll on retrotransposon families, individual copies become inactive through accumulation of point mutations and indels and through LTR–LTR recombination (Shirasu et al. 2000; Vitte et al. 2007a). Loss of retrotransposons by LTR–LTR recombination seems to be positively correlated with LTR length (Vitte et al. 2007a), and is common for long retrotransposons with long LTRs (Kalander et al. 2000). In this regard, the unusually short LTRs of \( Tvv1 \) should make it relatively resistant to loss by LTR–LTR recombination.

Retrotransposon families such as \( Tvv1 \) that are present in low copy number also face the challenge of how to replicate should the ORFs encoding the requisite proteins be lost from their members. However, complementation by other retrotransposons (Sabot and Schulman 2006; Tanksanen et al. 2007) of truncated ORFs appears to permit non-autonomous elements to play a major role in shaping plant genomes (Wawrzynski et al. 2008). A striking example is Cassandra, a group of retrotransposons, which is non-autonomous but highly conserved across the vascular plants (Kalander et al. 2008). Hence, it appears that effective replication and curtailed recombinational loss have permitted \( Tvv1 \) to survive at least 100 MY in plant evolution despite its low copy number.

Acknowledgments This work was supported by funding from Région Alsace, INRA, and the Academy of Finland (Project 123074). We thank Emilie Haegy and Romain Guyot (IRD, Montpellier) for technical assistance, and the members of the experimental unit of INRA-Colmar for the production of plants in the greenhouse. We also thank Véronique Lefebvre (INRA, Avignon), Gilles Pilate (INRA, Orléans), Veronique Brault (INRA-Colmar), Florence Lahogue-Esnull, Michel Renard and Jean-Paul Dantec (INRA, Rennes) who kindly provided us DNA and plant samples.

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standards The experiments comply with the current laws of the countries in which they were performed.

References


Casacuberta JM, Vernettes S, Grandbastien MA (1995) Sequence variability within the tobacco retrotransposon Tnt1 population. EMBO J 14:2670–2678


Moisy C, Blanc S, Merdinger D, Pelsy F (2008a) Structural variability of Tvv1 grapevine retrotransposons can be caused by illegitimate recombination. Theor Appl Genet 116:671–682


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


