ORIGINAL PAPER

The *Tvv1* 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

Abstract
Key message **Combining several different approaches, we have examined the structure, variability, and distribution of *Tvv1* retrotransposons. *Tvv1* 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 *Tvv1_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 *Tvv1_Sdem* with *Tvv1-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 *Tvv1* family members in potato, tomato, and poplar genomes, and led to the identification of full-length copies of *Tvv1* in these species. We were also able to identify polymorphism in UTL sequences between *Tvv1_Sdem* copies from wild and cultivated potatoes that are useful as molecular markers. Combining different approaches, our results suggest that the *Tvv1* 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, *Tvv1* represents an unusual plant retrotransposon metapopulation comprising highly similar members disjointedly dispersed among very distant species. The twin features of *Tvv1* 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.

C. Moisy · A. H. Schulman · R. Kalendar · J. P. Buchmann
MTT/BI Plant Genomics Lab, Institute of Biotechnology,
University of Helsinki, P.O. Box 65, Biocenter 3, Viikinkaari 1,
00014 Helsinki, Finland

C. Moisy · A. H. Schulman
Biotechnology and Food Research, MTT Agrifood Research
Finland, Myllytie 1, 31600 Jokioinen, Finland

C. Moisy (✉)
INRA, UMR AGAP 1334, DAVEM, 2 place Pierre Viala,
34060 Montpellier Cedex 1, France
e-mail: cedric.moisy@supagro.inra.fr

F. Pelsy
INRA, UMR1131, 68000 Colmar, France

F. Pelsy
Université de Strasbourg, UMR1131, 67000 Strasbourg, France

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 polyprotein containing aspartic 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 the 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 Schulman 2006; Schulman 2013; Schulman and Wicker 2013). This life cycle is very prone to errors, estimated at 1.4×10^{-5} mutations/bp/cycle, because both RNA polymerase II and reverse transcriptase lack proofreading activity (Gabriel et al. 1996; Preston 1996; Abram 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 Schulman 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; Schulman 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" (Hanski 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* (Tanskanen 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
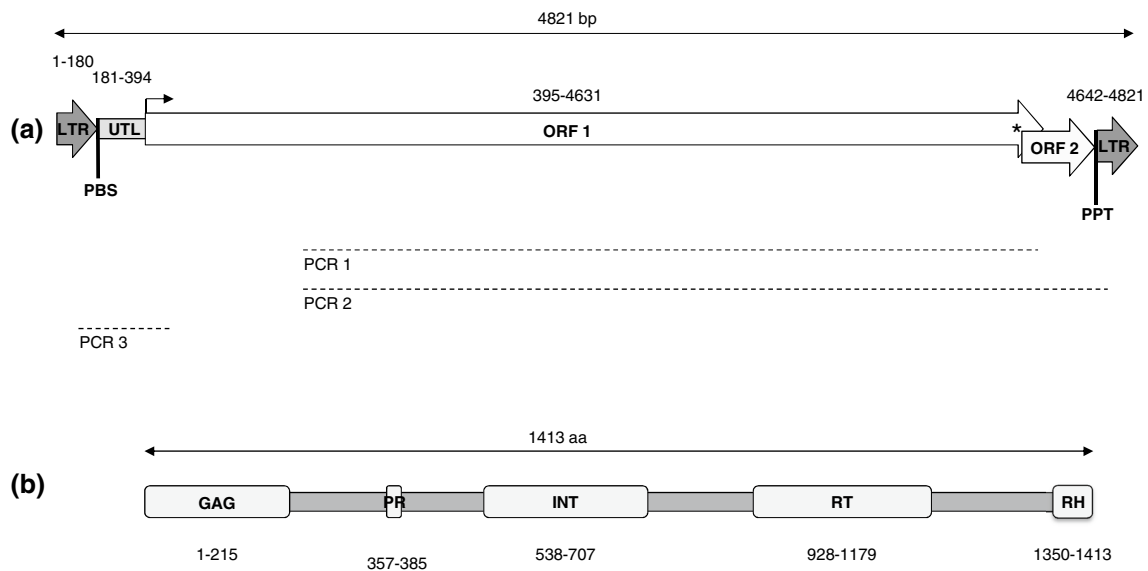
Fig. 1 Schematic representation of the structure of *Tvv1_Sdem*. **a** Position of domains and PCR amplifications on the nucleic sequence. Product PCR1 was amplified from the primer pair GAGsd-3F and RNASsd-2R and targets the internal region extending from the *gag* domain to the *rh*. Product PCR2 was amplified from the primer pair GAGsd-3F and LTRsd-4R and targets the internal region extending from the *gag* domain to the 3′ LTR. PCR3 was amplified from the primer pair LTRsd-utl and GAGsd-utl and amplifies the UTL region. **b** Putative protein sequence and functional domains encoded by ORF 1 and 2 considering the frameshift. *LTR* long terminal repeat, *PBS* primer binding site, *UTL* untranslated leader region, *ORF* open reading frame, *asterisk* frameshift

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. *Tvv1* represents a highly unusual example of a low-copy retrotransposon metapopulation with highly similar members dispersed unevenly among very distant species.

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-utl 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 MgCl$_2$), 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-VB* 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 (Kalendar 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-VB*, *Tvv1_Sdem*, *Tnt1*, *PDR1*) and frameshifts 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-VB* (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′ TGGTATCAGAGCC 3′ comprises the Primer Binding Site (PBS) complementary to tRNA[Met]. Immediately upstream to the 3′ LTR a polypurine tract (PPT), 5′ TGAGGGGGAG 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-VB* from *V. vinifera*

Tvv1_Sdem was compared to the canonical full-length copy *Tvv1-VB* previously described in grapevine (Moisy et al. 2008a). *Tvv1_Sdem* and *Tvv1-VB*, 4,821 and 5,222 bp, respectively, share 61 % identity overall. *Tvv1_Sdem* LTRs (180 bp) are longer than the identical *Tvv1-VB* 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-VB* 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-VB*.

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-VB*, which generates a +1 frameshift and likely a truncated

polyprotein (Fig. 1). The conceptual full-length *Tvv1_Sdem* *gag–pol* polyprotein is 1413 aa, compared to 1382 aa for *Tvv1-VB*, with the domain order (Fig. 1) and motifs (Fig. S1) typical for superfamily *Copia* (Bowen and Mcdonald 1999; Wilhelm and Wilhelm 2001; Wicker et al. 2007). The two proteins display 69 % identity and 79 % similarity overall, with locally higher similarity in functionally conserved domains. The predicted *Tvv1_Sdem* Gag (aa 1 to 215) is 67 % identical (87 % similar) to *Tvv1-VB* Gag in a gapless alignment (data not shown). The most conserved domains (Fig. S1) are INT (79 % identity over 170 aa), RT (81 % identity over 252 aa), and RH (86 % over 128 aa). Alignments of the NABD, INT, RT, and RH are consistent in showing that *Tvv1_Sdem* is more closely related to the *Tvv1* than to any other retrotransposon heretofore described (Fig. S1).

The *Tvv1* family in other asterids

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 results from PCR3 showed that the UTL is more variable than the rest of *Tvv1*. 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 695-174-106 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. tuberosum* (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 T_n 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

(a)

PCR #	Target	Exp. size (bp)	*S.demissum*				*S.tuberosum* 2X Rosa H1	*V.vinifera* PN40024	*C.annuum* Yolo Wonder	*S.lycopersicum* Vendor	*S.linnaeanum* MM195	*A.thaliana*	*N. tabacum*
			695-152-105	695-154-015	695-168-102	695-174-106							
PCR 1	gag - RH	3335	3300 +1800	3300 +1800	3300	3300 +1800	3300[1]	-	-	3300[2]	-	-	-
PCR 2	gag - LTR3'	3666	3600	3600	3600	3600	3600 weak	1600	-	-	-	-	-
PCR 3	UTL	419	419	-	419	419	419	-	-	-	-	-	-
		other	yes	yes	yes	yes	yes	>1kb	weak	weak	-	weak	weak

[1] Fragments cloned and sequenced under the names *Tvv1_St-ROSA49* and *Tvv1_St-ROSA54* ; [2] *Tvv1_Sl-VENDOR70F3* and *Tvv1_Sl-VENDOR80*
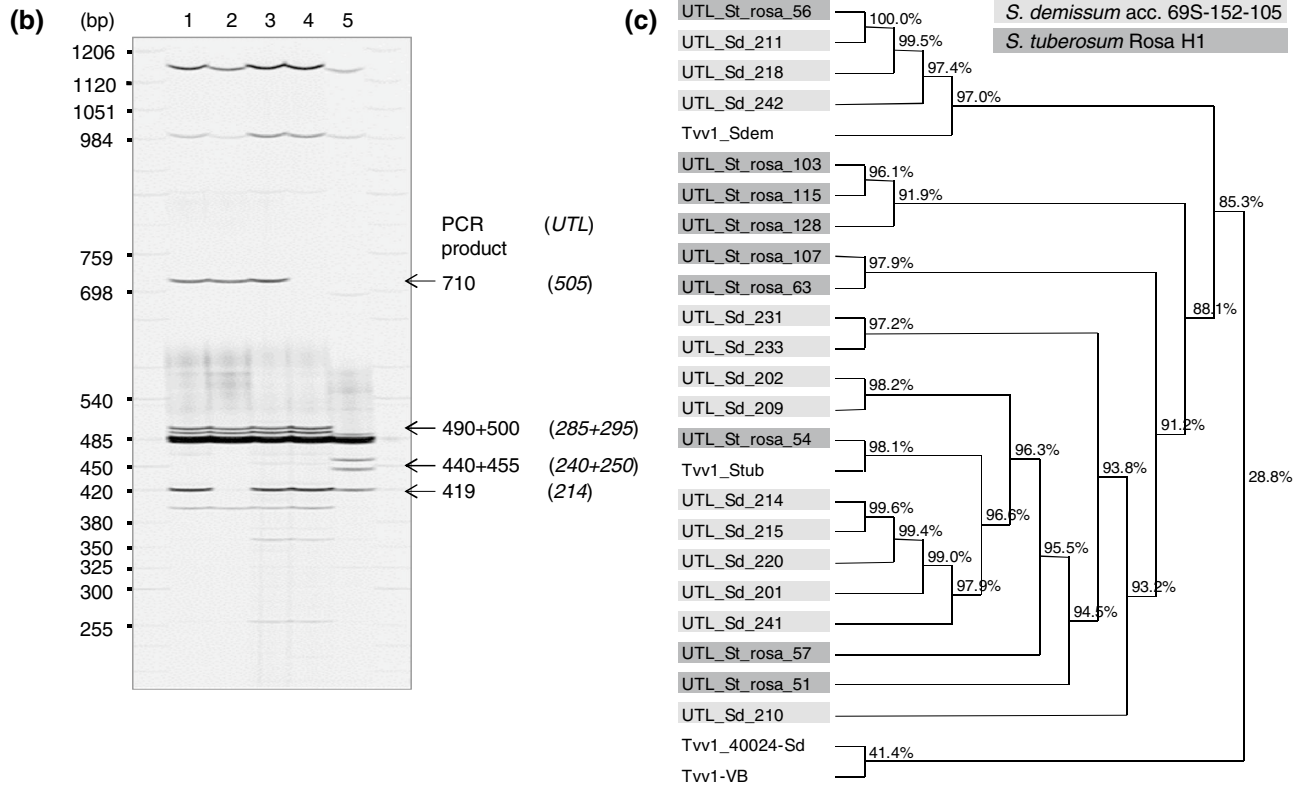
Fig. 2 Internal PCR amplifications of *Tvv1* in *Solanaceae* and other plants. **a** Sizes of the PCR products resulting from PCR 1, 2 and 3 on different plant DNAs. **b** PCR3 (UTL) products separated by electrophoresis on a long 5.5 % acrylamide gel in a LiCor (Lincoln, NB) 4000L automated DNA sequencer. *Line 1*: *S. demissum*, accession 69S-152-105, *2* accession 69S-154-105, *3* accession 69S-168-102, *4* accession 69S-174-106 and *5 S. tuberosum* 2x, accession Rosa H1. Size markers in bp are indicated on the left margin. Sizes of the polymorphic bands (and corresponding UTL sizes) are indicated on the right margin. **c** Comparison of only non-redundant UTL sequences retrieved by cloning from *S. demissum* accession 69S-152-105 (13 sequences) and from *S. tuberosum* 2x, accession Rosa H1 (9 sequences). Percentage of identity is given for each comparison

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

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

Name	Genome	Total length (bp)	LTRs		Insertion date (MYA)	UTL (bp)	*gag–pol* (bp)
			5'/3' (bp)	Identity (%)			
Tvv1_Cpap	*C. papaya*	5,193	195/198	85.4	4.83	672	4,118
Copia39-PTR[a]	*P. trichocarpa*	3,941	225	100.0	0.0	1,045	2,446
Tvv1_Pt-pop004-B08	*P. trichocarpa*	3,919	225	100.0	0.0	1,023	2,446
Tvv1_Spimp	*S. pimpinellifolium*	4,884	150/151	91.4	2.45	307	4,276
Tvv1_Sdem	*S. demissum*	4,821	180	98.3	0.65	214	4,461
Tvv1_Slyco	*S. lycopersicum*	4,941	148/145	90.0	1.68	389	4,259
Tvv1_Stub	*S. tuberosum*	4,916	180	96.1	1.54	276	4,270
Tvv1_40024-Sd	*V. vinifera*	4,957	179	96.6	1.32	466	4,123
Tvv1-VB[b]	*V. vinifera*	5,222	149	100.0	0.0	768	4,146
Tvv1 family[b]	*V. vinifera*	4,640–7,162	149–198	79.4–100.0	0–7.1	179–747	3,863–6,114

[a] From Repbase

[b] From the 23 full-length *Tvv1* copies identified in the grapevine genome (Moisy et al. 2008b)

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 1.54 to 2.45 MYA (Table 1).

In tobacco databases, including Genbank Genome Survey Sequences (GSS), the best hit displayed more than 80 % identity with *Tvv1_Sdem* but covers only 15 % of its sequence; we were not able to reconstruct a complete retrotransposon sequence. Moreover, a BLAST search was also conducted in a specialized database containing retrieved tobacco LTR retrotransposon sequences, but no significant similarity with *Tvv1_Sdem* was detected (Nicolas Maunouri, INRA, personal communication). Based on BLAST and PCR results, we consider that the *Tvv1* family has no full-length autonomous member in the tobacco genome. Given that the *Tvv1* family was first identified in *Vitis*, a search of the grapevine genome sequence using *Tvv1_Sdem* as the query was expected to yield matches. A total of 134 were found. The best covered 28 % of the sequence with an overall identity of 77 %. One particular element, *Tvv1_40024-Sd* (HE614295, Table 1), was retrieved from the grapevine genome sequence of chromosome 10 (reverse strand, nt 574301 to 579257) available on the Genoscope server. The *Tvv1_Sdem* and *Tvv1_40024-Sd* ORFs are 71 % identical while their full sequences (including LTRs and UTL) are 66 % identical. Comparison of *Tvv1_40024-Sd* with other *Tvv1* copies of grapevine indicates that *Tvv1_40024-Sd* is a distant member of the *Tvv1* family. Surprisingly, although present in the grapevine genome, *Tvv1_40024-Sd* is closer to *Tvv1_Sdem* from wild

potato than to any full-length *Tvv1* from grapevine previously 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 (AAAAC), 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*,

*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 1*, *Tvv1_Pt-pop004-B08*, *Tvv1_Sdem*, *Tvv1_Slyco*, *Tvv1_Spimp*, *Tvv1_Stub*, *Tvv1_40024-Sd*, and *Tvv1-VB* as being members of the *Tvv1* family, which is clearly distinct from any other retrotransposon family heretofore 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-Stub*) 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
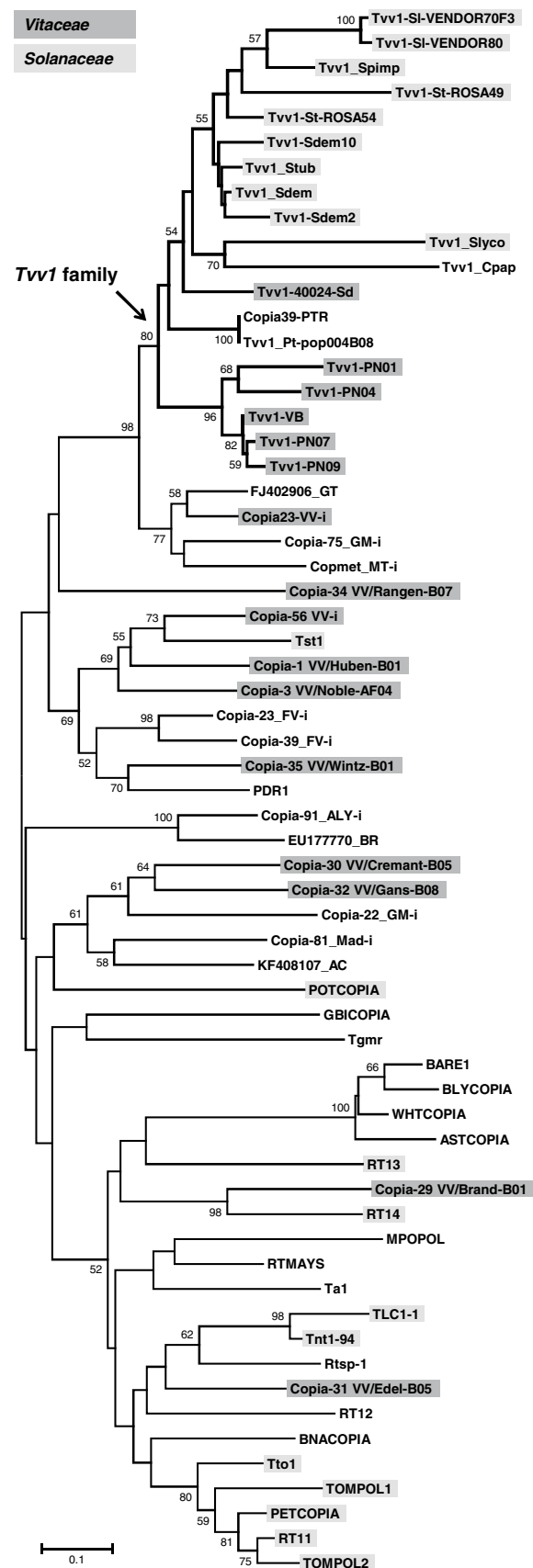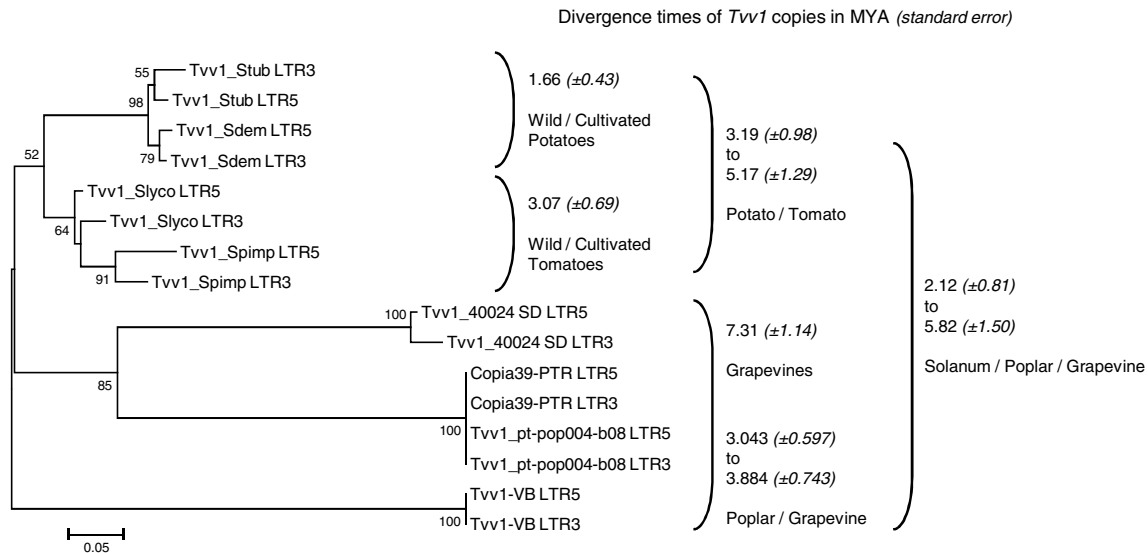
Fig. 4 Phylogenetic analysis of *Tvv1*-related LTR sequences and estimation of divergence time. Neighbor-Joining tree based on multiple sequence alignment of LTRs using the p-distance model and 1,000 replicates. Only bootstrap values greater than 50 % are shown. Estimation of divergence time is based on the pairwise distance comparison of LTR sequences as described in the text

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 *Tvv1_Sdem* is no longer autonomous because its coding region contains a frameshift in the *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 *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 *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 *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, *Tvv1* UTLs in the Solanaceae have diverged mainly by accumulation of indels as was previously described for *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 *Tvv1* UTL are more polymorphic between *Vitis* species than are microsatellites (Pelsy 2007). The UTL data indicate that *Tvv1* copies in *Solanum* have most probably arisen from several ancestors. The dispersion of the *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 *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 *Tvv1* family in genomes. The usefulness of *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 *Tvv1* members in both asterid (potato and tomato) and rosid (poplar and grapevine) genomes, and led to the identification of full-length copies of *Tvv1*-like sequences in these species. Excepting *Tvv1_Pt-pop004-B08* from poplar, which shows a large deletion but retains *Tvv1*-specific motifs, and the more divergent copy *Tvv1-Cpap* from papaya, all other *Tvv1* relatives belong to the same family according to the 80–80–80 rule proposed by Wicker et al. (2007). Compared against the *Tnt1*-like elements within *Solanum* (Manetti et al. 2009) that show specific amino acid sequences conserved among *Tnt1*, *TLC1*, and *Retrosol1*, we have identified several specific motifs that can be considered as fingerprints of the *Tvv1* family. These signatures, together with the detection of several distinct copies through in silico and in vitro approaches, establish that the *Tvv1* family forms a metapopulation of retrotransposons that is unusually distributed among the dicotyledonous plants.

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

Phylogenetic analysis of *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, *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, *Tvv1*, unusually, is both more conserved and less ubiquitous than retrotransposon families heretofore described.

The apparently spotty distribution of *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 *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 repetitious 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 *Vitis* 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 (Kalendar 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; Tanskanen 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 (Kalendar 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.

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

Abram M, Ferris A, Shao W, Alvord W, Hughes S (2010) Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. J Virol 84:9864–9878

Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL (2009) Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res 19:243–254

Boutabout M, Wilhelm M, Wilhelm FX (2001) DNA synthesis fidelity by the reverse transcriptase of the yeast retrotransposon Ty1. Nucleic Acids Res 29:2217–2222

Bowen NJ, Mcdonald JF (1999) Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. Genome Res 9:924–935

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

Cenci A, Combes M-C, Lashermes P (2010) Comparative sequence analyses indicate that Coffea (Asterids) and Vitis (Rosids) derive from the same paleo-hexaploid ancestral genome. Mol Gen Genet 283:493–501

The Angiosperm Phylogeny Group (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc 161:105–121

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

Dixit A, Ma K-H, Yu J-W, Cho E-G, Park Y-J (2006) Reverse transcriptase domain sequences from Mungbean (*Vigna radiata*) LTR retrotransposons: sequence characterization and phylogenetic analysis. Plant Cell Rep 25:100–111

Domingo E (2002) Quasispecies theory in virology. J Virol 76:463–465

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinforma 5:113. doi:10.1186/1471-2105-5-113

Eickbush TH, Jamburuthugoda VK (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134:221–234

Fortune PM, Roulin A, Panaud O (2008) Horizontal transfer of transposable elements in plants. Commun Integr Biol 1:74–77

Gabriel A, Willems M, Mules EH, Boeke JD (1996) Replication infidelity during a single cycle of Ty1 retrotransposition. Proc Natl Acad Sci USA 93:7767–7771

Gómez E, Schulman AH, Martínez-Izquierdo JA, Vicient CM (2006) Integrase diversity and transcription of the maize retrotransposon Grande. Genome 49:558–562

Hanski I (1998) Metapopulation dynamics. Nature 396:41–49

Jansen RK, Kaittanis C, Saski C, Lee S-B, Tomkins J, Alverson AJ, Daniell H (2006) Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol Biol 6:32

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH (2000) Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc Natl Acad Sci USA 97:6603–6607

Kalendar R, Tanskanen J, Chang W, Antonius K, Sela H, Peleg O, Schulman AH (2008) Cassandra retrotransposons carry independently transcribed 5S RNA. Proc Natl Acad Sci USA 105:5833–5838

Kalendar R, Lee D, Schulman AH (2011) Java web tools for PCR, in silico PCR, and oligonucleotide assembly and analysis. Genomics 98:137–144

Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9:411–412

Kidwell MG, Lisch DR (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution 55:1–24

Koch MA, Haubold B, Mitchell-Olds T (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). Mol Biol Evol 17:1483–1498

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

Le Rouzic A, Boutin TS, Capy P (2007) Long-term evolution of transposable elements. Proc Natl Acad Sci USA 104:19375–19380

Llorens C, Muñoz-Pomer A, Bernad L, Botella H, Moya A (2009) Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct 4:41

Lohe AR, Moriyama EN, Lidholm DA, Hartl DL (1995) Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. Mol Biol Evol 12:62–72

Macas J, Koblížková A, Navrátilová A, Neumann P (2009) Hypervariable 3′ UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. Gene 448:198–206

Manetti ME, Rossi M, Nakabashi M, Grandbastien MA, Van Sluys MA (2009) The Tnt1 family member Retrosol copy number and structure disclose retrotransposon diversification in different Solanum species. Mol Gen Genet 281:261–271

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

Moisy C, Garrison KE, Meredith CP, Pelsy F (2008b) Characterization of ten novel Ty1/copia-like retrotransposon families of the grapevine. BMC Genomics 9:1–14

Nielen S, Campos-Fonseca F, Leal-Bertioli S, Guimarães P, Seijo G, Town C, Arrial R et al (2010) FIDEL-a retrovirus-like retrotransposon and its distinct evolutionary histories in the A- and B-genome components of cultivated peanut. Chromosome Res 18(2):227–246

Ojosnegros S, Perales C, Mas A, Domingo E (2011) Quasispecies as a matter of fact: viruses and beyond. Virus Res 162(1–2):203–215

Otto TD, Gomes LH, Alves-Ferreira M, De Miranda AB, Degrave WM (2008) ReRep: computational detection of repetitive sequences in genome survey sequences (GSS). BMC Bioinforma 9:366

Pelsy F (2007) Untranslated leader region polymorphism of *Tvv1*, a retrotransposon family, is a novel marker useful for analyzing genetic diversity and relatedness in the genus Vitis. Theor Appl Genet 116:15–27

Pelsy F, Merdinoglu D (2002) Complete sequence of *Tvv1*, a family of Ty 1 copia-like retrotransposons of *Vitis vinifera* L., reconstituted by chromosome walking. Theor Appl Genet 105:614–621

Preston BD (1996) Error-prone retrotransposition: rime of the ancient mutators. Proc Natl Acad Sci USA 93:7427–7431

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16:276–277

Roulin A, Piegu B, Wing R, Panaud O (2008) Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon RIRE1 within the genus Oryza. Plant J 53:950–959

Roulin A, Piegu B, Fortune PM, Sabot F, D'Hont A, Manicacci D, Panaud O (2009) Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon Route66 in Poaceae. BMC Evol Biol 9:58

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365–386

Sabot F, Schulman AH (2006) Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity 97:381–388

Sabot F, Schulman AH (2007) Template switching can create complex LTR retrotransposon insertions in Triticeae genomes. BMC Genomics 8:247

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Schulman AH (2013) Retrotransposon replication in plants. Curr Opin Virol 3(6):604–614

Schulman AH, Wicker T (2013) A field guide to transposable elements. In: Fedoroff NV (ed) Plant transposons and genome dynamics in evolution. Wiley, Hoboken, pp 15–40

Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res 10:908–915

Smith RA, Anderson DJ, Preston BD (2004) Purifying selection masks the mutational flexibility of HIV-1 reverse transcriptase. J Biol Chem 279:26726–26734

Stuart-Rogers C, Flavell AJ (2001) The evolution of Ty1-copia group retrotransposons in gymnosperms. Mol Biol Evol 18:155–163

Suoniemi A, Tanskanen J, Pentikäinen O, Johnson MS, Schulman AH (1998) The core domain of retrotransposon integrase in Hordeum: predicted structure and evolution. Mol Biol Evol 15:1135–1144

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28(10):2731–2739

Tanskanen J, Sabot F, Vicient CM, Schulman AH (2007) Life without GAG: the BARE-2 retrotransposon as a parasite's parasite. Gene 390:166–174

Vershinin AV, Ellis TH (1999) Heterogeneity of the internal structure of PDR1, a family of Ty1/copia-like retrotransposons in pea. Mol Gen Genet 262:703–713

Vicient CM, Kalendar R, Schulman AH (2005) Variability, recombination, and mosaic evolution of the barley BARE-1 retrotransposon. J Mol Evol 61:275–291

Vitte C, Chaparro C, Quesneville H, Panaud O (2007a) Spip and Squiq, two novel rice non-autonomous LTR retro-element families related to RIRE3 and RIRE8. Plant Sci 172:8–19

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

Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, Davis CC, Latvis M, Manchester SR, Soltis DE (2009) Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci USA 106:3853–3858

Wawrzynski A, Ashfield T, Chen NWG, Mammadov J, Nguyen A, Podicheti R, Cannon SB, Thareau V, Ameline-Torregrosa C, Cannon E et al (2008) Replication of nonautonomous retroelements in soybean appears to be both recent and common. Plant Physiol 148:1760–1771

Wicker T, Keller B (2007) Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res 17(7):1072–1081

Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N (2006) 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7:275

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell AJ, Leroy P, Morgante M, Panaud O et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982

Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M, Stein N (2009) A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. Plant J 59:712–722

Wilhelm M, Wilhelm FX (2001) Reverse transcription of retroviruses and LTR retrotransposons. Cell Mol Life Sci 58:1246–1262

Witte CP, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci USA 98:13778

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214