Supplementary Note 2

Analysis of *Petunia vein clearing virus* (PVCV) sequences, retroelements and tandem repeats in *Petunia axillaris* N and *P. inflata* S6

Trude Schwarzacher[1*], J.S. (Pat) Heslop-Harrison[1] and Katja R. Richert-Pöggeler[2]

1) University of Leicester, Department of Genetics, Leicester LE1 7RH, UK.
2) Institute for Epidemiology and Pathogen Diagnostics, Julius Kühn-Institut, (JKI) – Federal Research Centre for Cultivated Plants, 38104 Braunschweig, Germany

*) Address correspondence to TS32@le.ac.uk

Short Title: PVCV, retroelements and tandem repeats in *Petunia*

ABSTRACT

Within the genome sequence assemblies of *P. axillaris* (*PaxiN*) and *P. inflata* (*PinfS6*) and unassembled reads, we analysed the occurrence of endogenous *Petunia vein clearing virus* (PVCV) sequences, other endogenous pararetrovirus (EPRV) sequences, LTR-retroelements, and tandem repeats. Petunia genomes show substantial diversity in their pararetroviral sequences as revealed in searches using the polymerase motif. Homologies to two genera of *Caulimoviridae*, *Petu*- and *Florendoviruses,* with more than 60% amino acid identity, were present in both genomes. Almost complete PVCV copies, fragments, and degenerate copies, sometimes in tandem arrays, were found. PVCV motifs were more frequent in *P. axillaris*, with the results seen in the assemblies confirmed by *in situ* hybridization of PVCV fragments to metaphase chromosomes indicating that *P. axillaris* is likely a more permissive host for EPRVs. LTR-retroelements are localised near centromeres; about 6500 full length elements were found in the *PinfS6* assembly while 4500 were in *PaxiN*. Apart from rDNA, microsatellites and telomeric sequences, no highly abundant tandem repeats were identified in the assembly or raw reads. Repeat cluster analysis indicates that LTR-retroelements are associated with simple sequence repeats and low complexity DNA families and that repeats within *Petunia* are very diverse, with none having amplified to form a major proportion of the genome. The repeat landscape of *Petunia* is different from other species of *Solanaceae*, in particular the x=12 crown group including *Solanum* and *Nicotiana*, with a relative low proportion (60-65%) of total repeats for a genome size of 1.4Gb, x=7, and a high degree of genome plasticity.

KEYWORDS

Petunia vein clearing virus, pararetrovirus, retroelement, *Copia* superfamily, *Gypsy* superfamily, telomere, tandem repeat, repeat explorer, K-mer analysis, genome organisation, chromosome size, fluorescent *in situ* hybridization

INTRODUCTION

The genome sizes reported for 20 species within the genus *Petunia* are similar with a range of 1.30 to 1.57 pg 1C for diploid species with a chromosome number of 2n=14 (Mishiba et al., 2000) corresponding to on average 1.4 Gb. The *Petunia* genome therefore is considerably larger than tomato and potato (900Mb and 844Mb respectively, Tomato Genome Consortium, 2012), but not as large as the hot pepper genome (3,480Mb; Kim et al., 2014) which contains a large proportion of repetitive sequences, in particular long terminal repeat (LTR) retroelements. DNA transposons have been described in *Petunia* (Gerats, 2009) and are analysed in Supplementary Note 3. In contrast, there are few reports of LTR retroelements including the *Gypsy* and *Copia* superfamilies. Matsubara et al. (2005) described the full length rTph1 element that shares features with the C*opia*-superfamily, and recently Kriedt et al. (2013) studied *Petunia* species relationships with the RNase H – 3'LTR region of eight families of *Ty1/Copia* –Tork clade elements that are related to the tobacco Tnt1. Both *Ty3*/*Gypsy* and *Ty1*/*Copia* reverse transcriptase domains were described from several petunia species by Richert-Pöggeler and Schwarzacher (2009). Tandemly repeated satellite DNA families in *Petunia* have not been widely reported.

Petunia (Richert-Pöggeler et al., 2003), similar to other members of *Solanaceae* (Hansen et al., 2005; Geering et al., 2014), has integrated pararetrovirus sequences. A spontaneous outbreak of vein-clearing disease in the hybrid species *P. hybrida* could be traced back to activation of *Petunia vein clearing virus* (PVCV) genomes inserted into the host chromosomes in a tandem array (Richert-Pöggeler et al., 2003). PVCV belongs to the genus *Petuvirus* within the family of *Caulimoviridae* (King et al., 2012). This virus family is also referred to as plant pararetroviruses since it uses reverse transcription for genome amplification. In contrast to retroviruses, their genome consists of a circular double stranded DNA with single strand gaps (Hohn et al., 2008). Furthermore, integration into the host genome is not obligatory for the replication cycle of parartroviruses. Comparing the genome sequences of *P. axillaris* N (*PaxiN*) and *P. inflata* S6 (*PinfS6*) allows the study of the genomic context to determine any effect on diversity and evolution of *Caulimoviridae*.

RESULTS

Petunia vein clearing virus (PVCV) insertions

Integrated sequences of PVCV were found in the genomic scaffolds and on chromosomes of both petunia species analyzed. In both genome assemblies, several arrays of almost complete genome length and degenerated PVCV sequences were found (Tables 1 and 2, Figure 1 and details below). Data obtained from BlastN searches using the whole PVCV genome of 7206 bp were filtered for alignments larger than 500 nt in length. For *PaxiN,* 30 sequences were selected that ranged in size from 542 to 2848 nt in length and 80-99% identity, whereas *PinfS6,* 9 sequences with sizes of 563-635 nt and 78-80% identity were found. Fluorescent *in situ* hybridization with PVCV-specific probes comprising the complete viral genome created a stronger signal in *P. axillaris* than in *P. inflata*, both at centromeric regions of chromosomes III and VI for *P. axillaris* and chromosome IV for *P. inflata* (Figures 2 and 3). The bioinformatic and cytogenetic data indicate that PVCV abundance and perseverance is markedly higher in *P. axillaris* compared to *P. inflata.*

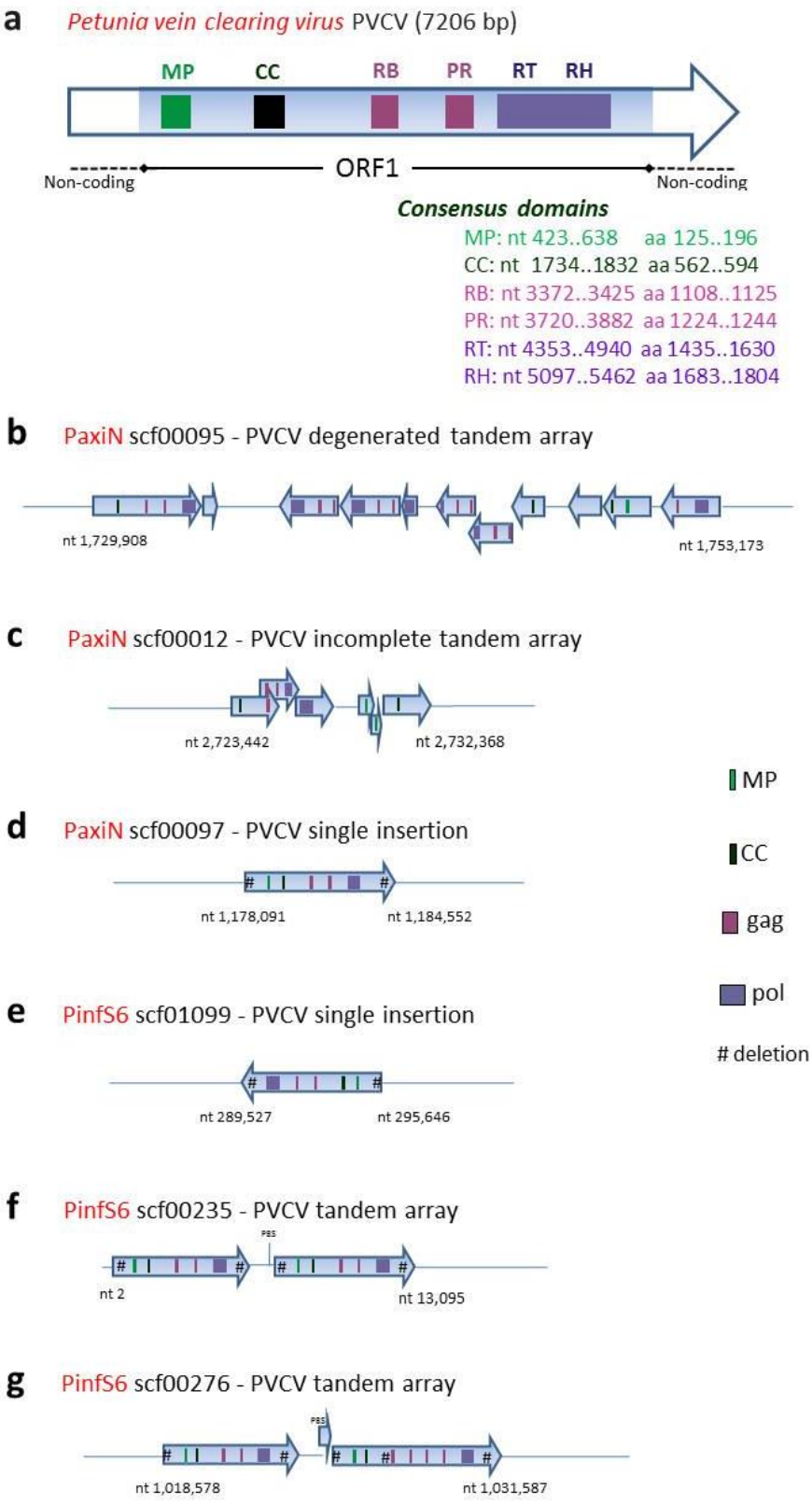


Figure 1: Full length PVCV genomes and PVCV integrations in *P. axillaris* N and *P. inflata* S6 genome assembly scaffolds. Integrations show head-to-tail and inverted orientations, deletions, duplications and overlapping protein motifs. For protein and regulatory DNA element boundaries see Tables 1 and 2.

MP: movement protein, CC: coiled coil domain, gag: group-specific antigen with RB (RNA binding Zn-finger motif of the capsid protein) and PR (protease), pol: polymerase with RT (reverse transcriptase) and RH (RNaseH); PBS: primer binding site comprising first 14nt of PVCV genome. Not shown are TSS (transcription start site) and Pro (Promoter) within the 3'-non coding region. Note that the single insertion depicted in d) for *PaxiN* shows 74% amino acid identity with PVCV compared to the single insertion depicted in e) for *PinfS6* displaying 56% amino acid identity.

The behavior and mode of integration seem to be similar for both species: the PVCV elements are present as single insertions as well as small tandem arrays. The latter display deletions as well as rearrangements and thus are not able to generate full length infectious viral RNA molecules via direct transcription as reported for *P. hybrida* W138 (Richert-Pöggeler et al., 2003). The investigated petunia species differ in preservation of integrated viral sequences. Those in *PinfS6* are subjected to a higher degree of degradation. At a stringent level of more than 2000aa and an id>70% we found only one PVCV copy in *P. inflata* compared to four copies in *PaxiN*.

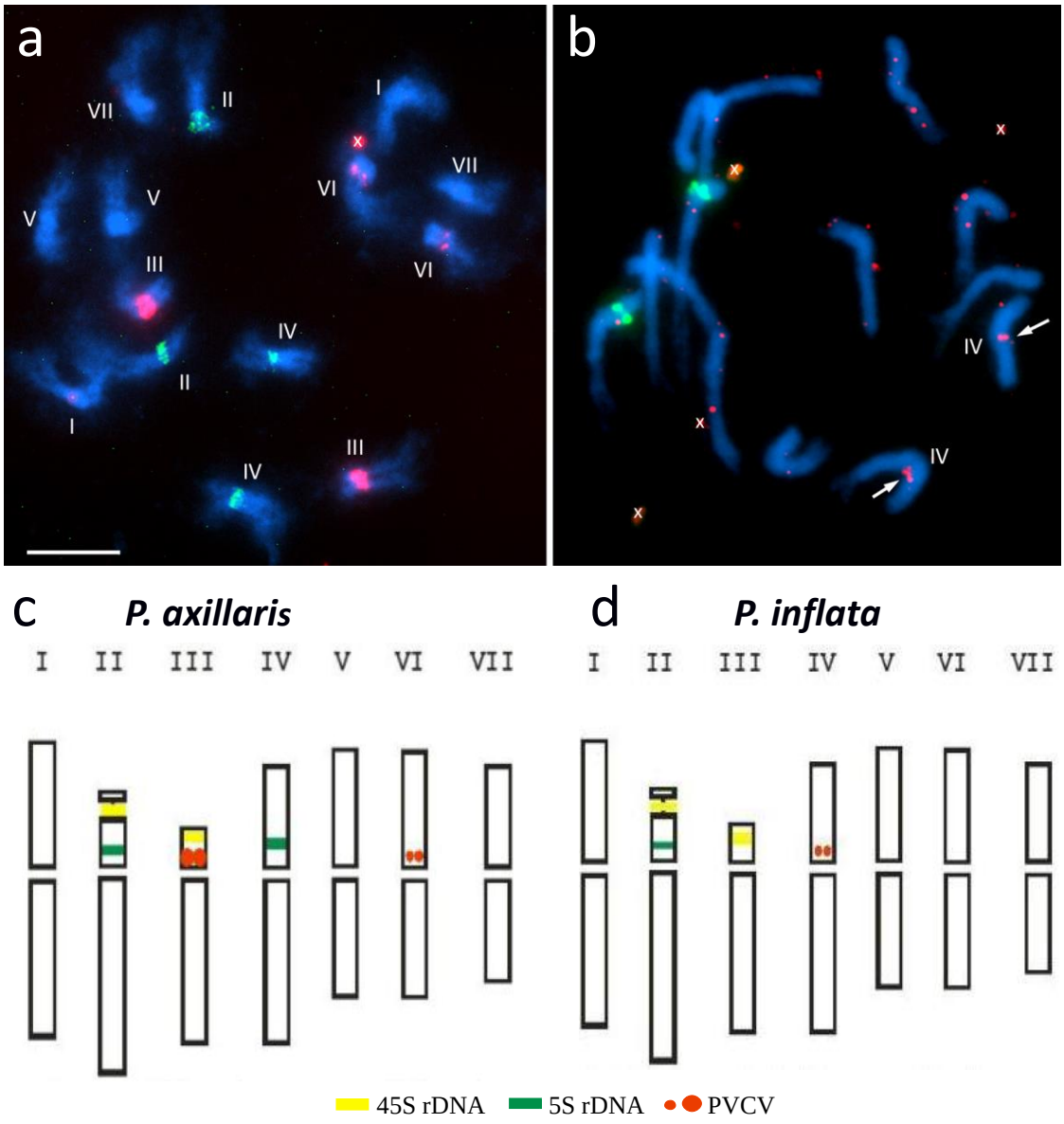


Figure 2: PVCV distribution on metaphase chromosomes *P. axillaris* and *P. inflata*
a,b: Fluorescent *in situ* hybridization using PVCV probe (Richert-Pöggeler et al., 2003; depicted in red) and 5S-rDNA (depicted in green) counterstained with DAPI (blue) **a:** *P. axillaris* chromosome identifications are indicated; PVCV signal is found at the centromeres of chromosomes III and VI (this spread was reprobed with *gypsy*-related retroelement junction fragment 4-18, see Figure 3). **b:** *P. inflata* chromosome IV is identified and the location of PVCV is indicated by arrows while remaining dots are background signal visible due to enhancement of weak signal. x denotes stain precipitates. Bar = 10 µm
c, d: Idiograms showing PVCV, 5S and 45S rDNA.

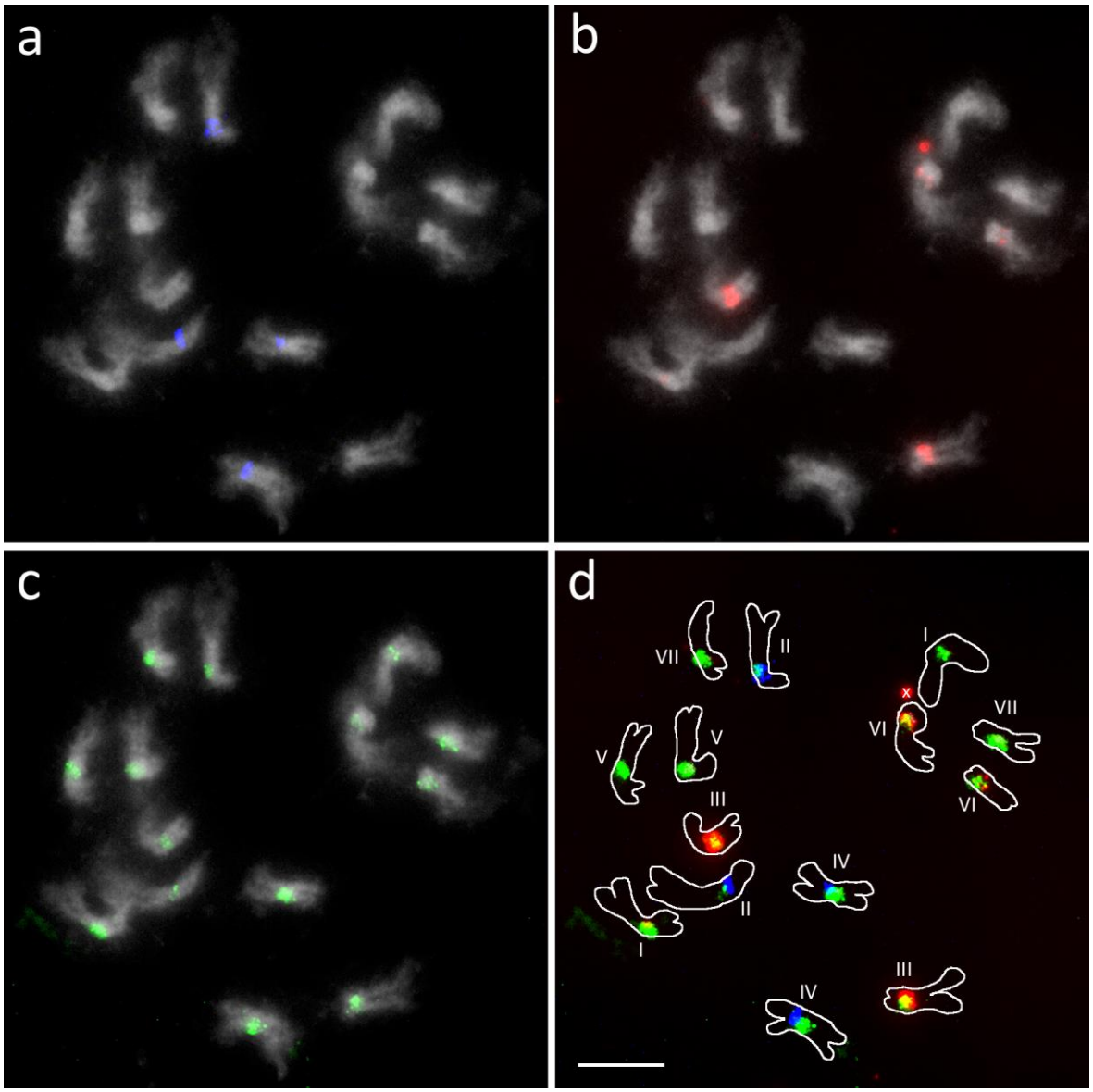


Figure 3: **Fluorescent *in situ* hybridization to metaphase chromosomes of *P. axillaris.***
a, b, c: Chromosomes were counterstained with DAPI (shown in grey), and hybridization sites are shown in blue for the 5S-rDNA (**a**) and red for PVCV (**b**) from the first hybridization (see Figure 2a) and in green for the *Gypsy* super family retroelement junction fragment 4-18 used to reprobe the spread (**c**).
d:The overlay is a drawing of the chromosomes with overlayed and enhanced FISH signal (overlapping signal of red and green appears yellow; original see main text Figure 2b). Chromosomes are identified by their size and 5S rDNA signal. Two large and two small PVCV sites close to the centromeres of chromosomes III and VI are present while all centromeres show dispersed signal with the *Gypsy*-related retroelement probe. X denotes stain precipitate. Bar = 10 µm.

PVCV sequence insertions in *PaxiN*

Highly preserved stretches of the N-terminal half of PVCV ORF1 with 99% aa identity (Table 1) comprising the MP consensus domain were identified in scf03256 and in scf01628. The latter scaffold contained at its 5'-end, nt 7239 -7835, highly preserved (96% identity) regulatory sequences out of the PVCV untranslated region (nt 6591 -7193) including the promoter, transcription start site as well as a polyA signal (Noreen et al., 2007). Both scaffolds contained also other parts of ORF1 that were less conserved (45 to 87% aa identity).

Table 1: Identified PVCV sequences within *PaxiN* scaffolds annotated in the assembly Peaxi162.

Scaffold (% identity)	Scaffold length (bp)	PVCV location in scf	PVCV-ORF1 (2179aa)	Orientation, frame	FISH detection by PVCV probes[1)]
***PaxiN* scaffolds with single PVCV sequences, filter: alignment length > 1500aa, id > 70%**					
scf00097 (74)	2.428.612	1.178.091 - 1.184.552	11-2176	forward	PVCV-L, M, R
scf00254 (74)	1.519.427	751.256 - 744.795	11-2176	reverse	PVCV-L, M, R
scf00447 (72)	1.217.748	1.146.839 - 1.152.364	318-2176	forward	PVCV-L, M
scf00560 (74)	869.919	238.998 - 245.459	11-2176	forward	PVCV-L, M, R
scf00674 (74)	473.906	265.554 - 271.907	11-2141	reverse	PVCV-L, M, R
scf00911 (74)	507.819	34.116 – 29.446	598-2176	reverse	PVCV-M
***PaxiN* scaffolds with highly preserved PVCV sequences, filter: alignment length > 380aa, id = 99%**					
scf03256 (99)	12295	552 - 1964	1-470	forward	PVCV-R
scf01628 (99)	9910	313-1460	50-432	forward	PVCV-R
***PaxiN* scf00012 with multiple PVCV sequences, filter: alignment length > 100aa, id > 40%**					
scf00012 (43%)	3.935.541	2.723.442 - 2.725.583	417-1139	forward, +3	?
scf00012 (64%)	3.935.541	2.725.433 - 2.727.046	1088-1631	forward, +2	PVCV-M
scf00012 (50%)	3.935.541	2.727.046 - 2.728.506	1636-2134	forward, +1	?
scf00012 (65%)	3.935.541	2.729.342 -2.729.914	4-198	forward, +2	?
scf00012 (67%)	3.935.541	2.729.863 - 2.730.255	179-309	forward, +1	?
scf00012 (43%)	3.935.541	2.730.362 - 2.732.368	345-1007	forward, +2	PVCV-L, M
***PaxiN* scf00095 with multiple PVCV sequences, filter: alignment length > 100aa, id > 40%**					
scf00095 (53%)	1.774.960	1.729.908 - 1.734.470	287-1811	forward, +3	PVCV-M
scf 00095 (52%)	1.774.960	1.734.563 - 1.734.898	1854-1965	forward, +2	?
scf 00095 (63%)	1.774.960	1.739.779 - 1.737.299	1126-1956	reverse, -1	PVCV-M
scf 00095 (64%)	1.774.960	1.742.275- 1.739.783	1059-1905	reverse, -1	PVCV-M
scf 00095 (71%)	1.774.960	1.742.929 -1.742.255	1571-1795	reverse, -1	PVCV-M
scf 00095 (58%)	1.774.960	1.744.585 - 1.742.936	945-1503	reverse, -1	PVCV-M
scf 00095 (60%)	1.774.960	1.746.506 - 1.744.659	945-1569	reverse, -3	PVCV-M
scf 00095 (44%)	1.774.960	1.747.798 - 1.746.485	497-951	reverse, -1	PVCV-M
scf 00095 (50%)	1.774.960	1.749.980 - 1.748.940	679-1046	reverse, -3	PVCV-M
scf 00095 (51%)	1.774.960	1.752.079 - 1.749.992	4-665	reverse, -1	PVCV-L
scf 00095 (63%)	1.774.960	1.755.647 - 1.753.173	1221-2059	reverse, -3	PVCV-M

1) For FISH stringent washing conditions were applied, therefore sequences with < 63 % of identity probably did not contribute to the observed signal as letters of PVCV-probes in grey indicate. "?": identified sequences are only covered partly (< 180 aa) by probes used.

In total, 49 scaffolds carrying 1 to 16 fragments of PVCV-like sequences, with a majority (73%) containing 1 to 5 PVCV fragments of various sizes were detected in searches using a cut off >100aa aligned, >40% aa identity. In six scaffolds (see Table 1), large continuous stretches of integrated sequences homologous to PVCV (cut off >1500aa aligned, >70% aa identity) comprising ORF1 in forward or reverse orientation were revealed. Besides single blocks, integrated PVCV sequences were found as tandem array-like structures (see scf00012 and scf00095, in Table 1 and Figure 1b and c). Thus the following integration patterns can be distinguished for the annotated sequences within the *P. axillaris N* genome: 1) the insertion spots (scf00095, scf00012) show rearrangement and fragmentation or 2) the integrated PVCV sequences (scf00097, scf00254, scf00447, scf00560, scf00674, scf00911) cover almost the complete ORF1 region. Interestingly, four of the latter copies

comprise the same region of aa 11-2141 in PVCV ORF1 while the other two copies show larger deletion on the N-terminal end of ORF1.

Filtered results obtained for scf00095 (id% > 40 and alignment length >100 aa) identified 11 copies of various ORF1 fragments which can be assembled into a "tail to tail - tail to head" array. The first part of the array contains deletions at the N- and C- terminal ends of ORF1 as well as near the C-terminal end. The latter is accompanied by a frameshift from +3 to +2. The "tail to head" part contains repetitive blocks of PVCV sequences and frameshifts (Table 1). The same filter applied for scf00012 revealed an almost complete ORF1 of PVCV comprising amino acids 4 to 2134 containing rearranged fragments, smaller deletions as well as frameshifts. For detection of PVCV sequences petunia chromosomes were hybridized with a mixture of three virus specific probes (Richert-Pöggeler et al., 2003) named PVCV-L (nt 657-1793), PVCV-M (nt 2235-5321) and PVCV-R (nt 5445-7206 and nt 1-671) using fluorescent *in situ* hybridization (FISH). In Table 1 we indicated detection only for those scaffold sequences that either are completely covered or contain at least 180 aa in common with probes PVCV-L, –M or -R. We suggest that mostly the single insertions detected by PVCV-L, –M and -R probes are responsible for the PVCV FISH signal identified within the pericentromeric region on chromosomes III and VI of *P. axillaris* (Figures 2 and 3, Table 1).

PVCV sequence insertions in *PinfS6*:

There are two near-complete copies of PVCV ORF1 at the 5'-end of scf00235 (56% identity), separated by 1000nt (nt 5903-6975). Within the region separating the two PVCV copies, a domain identical to the PVCV primer binding site (PBS) was located at nt 6907-6920 of the scaffold (Figure 1f). Besides the PBS no homologies to known sequences were identified, but as shown in Table 3 this junction sequence was frequently adjacent to PVCV insertions.

The strongest similarity to PVCV was found for sequences in scf00276 (Figure 1g). It contained a tandem array of PVCV sequences (head to tail) that were separated by 1 kb of sequences with no known homology (nt 1.024.645-1.025.740 in scf00276). The first copy in frame +3 covered most of PVCV ORF1 with a major (144 aa) and minor (3 aa) deletion at the N-terminal and carboxy terminal end respectively and a stop codon towards the carboxy terminal end. The 2nd copy showed deletions at both ends as well as in the centre of ORF1. The sequence with homology to PVCV aa 1007-1247 was inserted leading to a duplication of the gag region. The tandem array separating sequence (nt 1.024.645-1.025.740) was a hybrid of PVCV- and non-PVCV sequences. The latter showed no homology to sequences deposited in GenBank, but stretches of it could be found in scf09521 and scf16590 (see Table 3).

In scf00059, two blocks of fragmented PVCV sequences in forward orientation followed by clustered PVCV sequences comprising aa 4 to 2047 from PVCV ORF1 in reverse orientation were located at its 3'-end. The latter contained frameshifts. In the first quarter of scf00753, a cluster of repetitive PVCV fragments is present, mostly in the same orientation. They are not continuous but separated by non-PVCV sequences. The existence of only one continuous PVCV-sequence with >70% aa identity found in PinS6scf00276 (Table 2, Figure 1g) compared to numerous of those identified in *P. axillaris N* (Table 1) may explain the weak detection of PVCV in the *P. inflata S6* genome (Figure 2b).

Table 2: Identified PVCV sequences within *PinfS6* scaffolds annotated in the assembly Peinf101.

Scaffold (identity)	Scaffold length scf (bp)	PVCV location in scf	PVCV-ORF1 (2179aa)	Orientation, frame	Comments
PinfS6 scaffolds with single PVCV sequences, filter: alignment length > 1500aa, id > 50%, *alignment length > 800aa, id > 65%*					
scf00251 (55%)	2.742.664	2.734.378 - 2.729.603	4-1591	reverse	
scf 00844 (56%)	223.102	199.961 - 193.842	4-2059	reverse	
scf 01099 (56%)	1.012.215	295.646 - 289.527	4-2059	reverse	
scf 01671 (56%)	432.778	241.933 - 235.814	4-2059	reverse	
***PinfS6* scf00235 with multiple PVCV sequences, filter: alignment length > 1500aa, id > 50%**					
scf 00235 (56%)	1.810.835	6.976 – 13.095	4-2059	forward	tandem array head to tail with filler (s. text)
scf 00235 (56%)	1.810.835	2 – 5.902	78-2059	forward	
scf 00235	1.810.835	5903 - 6975	unknown sequence (s. Table 3) carrying PVCV-PBS nt 6907-69		
***PinfS6* scf00276 with multiple PVCV sequences, filter: alignment length > 800aa, id > 65%**					
scf 00276 (73%)	1.557.018	1.018.578 - 1.024.634	145-2176	forward, +3	tandem array head to tail with filler sequences (s. text)
scf 00276	1.557.018	1.024.645 - 1.025.264	unknown sequence (s. Table 3)		
scf 00276 (75%)	1.557.018	1.025.265 - 1.025.740	PVCV nt 1-476 (PBS, degenerated N terminal part of ORF1		
scf 00276 (68%)	1.557.018	1.025.741 - 1.028.308	145-1003	forward, +2	
scf 00276 (76%)	1.557.018	1.028.309 - 1.029.025	1007-1247	forward, +2	
scf 00276 (78%)	1.557.018	1.029.026 - 1.031.587	1248-2102	forward, +2	
***PinfS6* scf00059 with multiple PVCV sequences, filter: id>50%, alignment length >100aa**					
scf 00059 (61%)	1.578.608	64.320 - 65.480	4-390	forward, +3	
scf 00059 (55%)	1.578.608	66.366 - 67.226	1088-1379	forward, +3	
scf 00059 (51%)	1.578.608	76.324 - 74.510	1433-2047	reverse, -2	
scf 00059 (51%)	1.578.608	77.739 - 76.684	970-1333	reverse, -3	
scf 00059 (46%)	1.578.608	80.153 - 77.736	170-970	reverse, -1	
00059 (68%)	1.578.608	80.641 - 80.156	4-166	reverse, -2	
***PinfS6* scf00753 with multiple PVCV sequences, filter: id>60%, alignment length >100aa**					
scf 00753 (70%)	2.084.733	418.612 - 419.550	4-317	forward, +1	
scf 00753 (72%)	2.084.733	444.046 - 443.360	23-251	reverse, -3	
scf 00753 (73%)	2.084.733	447.339 – 446.515	1316-1587	reverse, -1	
scf 00753 (62%)	2.084.733	450.381 – 450.016	1196-1317	reverse, -1	
scf 00753 (67%)	2.084.733	452.331 – 453.290	4-324	forward, +3	
scf 00753 (69%)	2.084.733	458.072 – 459.013	7-317	forward, +2	
scf 00753 (70%)	2.084.733	472.568 – 473.221	1357-1574	forward, +2	
scf 00753 (61%)	2.084.733	475.644 – 475.988	4-119	forward, +3	
scf 00753 (74%)	2.084.733	475.958 – 476.356	110-242	forward, +2	
scf 00753 (70%)	2.084.733	477.668 – 478.750	1357-1717	forward, +2	
scf 00753 (71%)	2.084.733	480.792 – 481.346	4-189	forward, +3	
scf 00753 (65%)	2.084.733	483.902 – 485.188	1130 - 1561	forward, +2	
scf 00753 (68%)	2.084.733	486.922 – 487.746	4-279	forward, +1	
scf 00753 (71%)	2.084.733	502.531 – 502.031	113-279	reverse, -3	
scf 00753 (61%)	2.084.733	502.856 – 502.530	4-113	reverse, -2	
scf 00753 (63%)	2.084.733	517.270 – 516.731	4-184	reverse, -3	

Table 3: Scaffolds with sequences associated with PVCVs

scf number (identity)	scf Length (bp)	location in scf	orientation, association with PVCV sequences (Ps)
PVCV-sequences containing scaffolds with similar sequences to *PinfS6* scf00235_nt5903-6975 including PVCV-PBS at nt 6907-6920 [1)]			
scf00059 (92%)	1.578.608	68.896 – 67.916	reverse, between Ps
scf00235 (100%)	1.810.835	5.903 – 6.975	forward, between Ps
scf00235 (99%)	1.810.835	13.095 – 13.871	forward, no association with Ps
scf00753 (92%)	2.084.733	456.955 – 457.923	forward, upstream Ps with gap
scf00753 (94%)	2.084.733	479.737 – 480.702	forward, upstream Ps
scf00753 (94%)	2.084.733	492.613 – 493.385	forward, between repetitive Ps with gaps
scf00753 (93%)	2.084.733	438.095 – 437.356	reverse, between Ps with gaps
scf00753 (95%)	2.084.733	444.927 – 444.189	reverse, downstream with gap Ps
scf00753 (92%)	2.084.733	503.928 – 503.327	reverse, downstream with gap Ps
scf00844 (99%)	223.102	193.842 – 192.768	reverse, upstream Ps
scf01099 (100%)	1.012.215	289.527 - 288.522	reverse, upstream Ps
scf01671 (99%)	432.778	235.814 - 234.820	reverse, upstream Ps
Scaffolds with fragments similar to parts of sequence *PinfS6* scf 00276, nt 1.024.645 - 1.025.264			
scf09521 (92%)	17538	8.085 – 8.285	forward
scf09521 (87%)	17538	8.318 – 8.510	forward
scf16590 (95%)	13872	11.193 – 11.038	reverse

1) This sequence shows no homology to known sequences and was found in all PVCV containing scaffolds mentioned in Table 2, with the exception of scf00276.

Similarity analysis of selected integrated PVCV sequences within *PaxiN* and *PinfS6* genomes

Similarity analysis of aligned amino acid sequences from the nearly full-length PVCV insertions identified within the petunia genomes revealed two major clusters for chromosomal PVCV sequences (Figure 4). The tree topology supports the hypothesis that after speciation there had been at least two separate invasion events of petunia genomes by episomal PVCV sequences (Staginnus and Richert-Pöggeler, 2006). At an earlier event only *P. inflata* has been affected which is also illustrated in a lower value of sequence identity (Table 2). A second round of invasion included both *P. inflata* and *P. axillaris* genomes. This clade also includes episomal PVCV, and it probably has contributed to the tandem array structure of inducible endogenous PVCV existing after genome hybridization in *P. hybrida* (Richert-Pöggeler et al., 2003). Here the almost full-length single copies of PVCV show a slightly higher degree of preservation for the *P. axillaris N* genome. Evolution within the genome resulted in deletions, amino acid exchanges and ORF disruption by a stop codon in case of *PaxiN* scf00447 and scf00674. In the *P. inflata* genome context, detrimental forces seem to act more strongly on foreign DNA. The PVCV continuous copy *PinfS6* scf00276 carries not only a stop codon but also an amino acid exchange (*aspartic acid D* to aspartate N) within the reverse transcriptase consensus domain "VYID*D*VLL" common to a broad range of retroelements (Richert-Pöggeler and Shepherd, 1997). Adjacent to the full-length copy fragmented and rearranged PVCV sequences are located in *PinfS6* scf00276. Most likely they originate from the same integration event since the fragments show similar values for amino acid identity and PVCV sequences within the tandem array structure were most conserved.

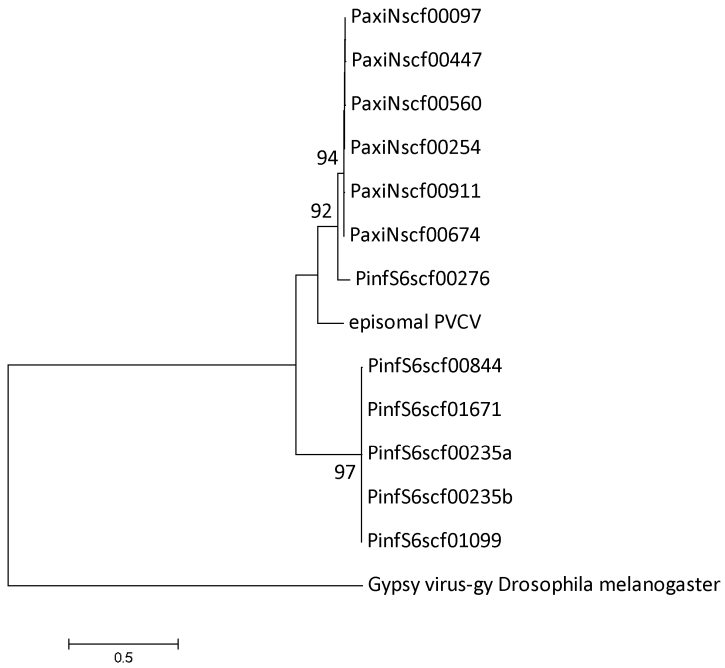
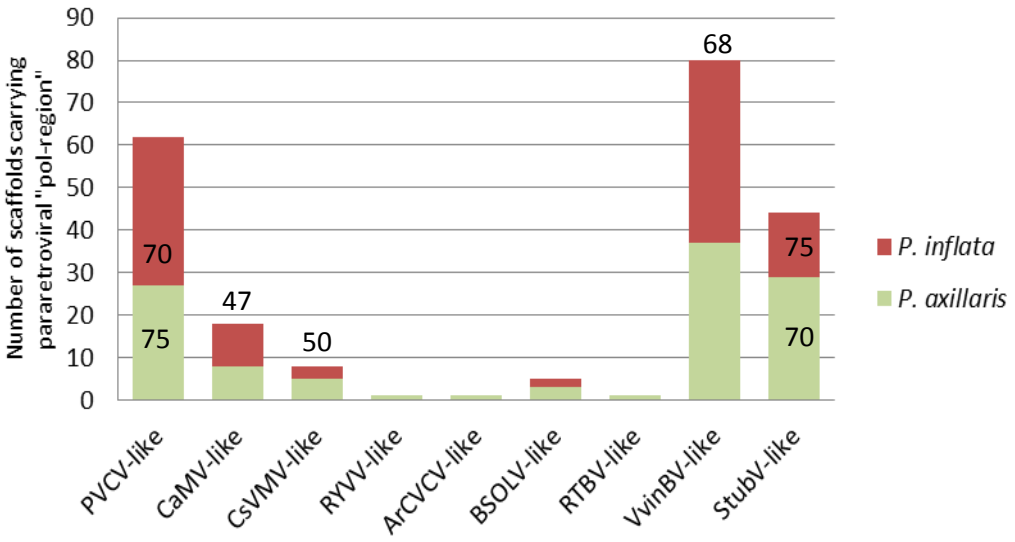


Figure 4: **Similarity analysis of PVCV ORF1 insertions within *P. axillaris N* and *P. inflata S6* genomes.**
The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model (Jones et al., 1992). The tree with the highest log likelihood (-7930.8490) is shown. The percentage of trees, cut off > 70, in which the associated taxa clustered together, is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically: when the number of common sites was < 100, or less than a quarter of the total number of sites, the maximum parsimony method was used; otherwise BIONJ method with MCL distance matrix was used. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 14 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 923 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 (Tamura et al., 2011). The *Gypsy virus –gy* of *Drosophila melanogaster* (DmeGypV-gy) was used as the outgroup.

Comparison of EPRV diversity in *PaxiN* and *PinfS6*

The pol region comprising the reverse transcriptase and RNase H is the most conserved motif of EPRVs. To identify additional PVCV insertions and EPRV diversity, scaffolds were searched for *Caulimoviridae* pol-like domains using the PVCV pol-region comprising aa 1425 to 1804 and *Caulimoviridae* sequences with homology to the PVCV pol-region as identified in BLASTp searches. These include pol-like regions with homology to bacilliform pararetroviruses illustrated by *Banana streak OL virus* and *Rice tungro bacilliform virus* and sequences from the genera *Caulimo*/*Cavemovirus* both displaying isometric particle morphology and a putative pararetrovirus with unknown morphology, *Aristotelia chilensis vein clearing virus*. Sequences were investigated further if the alignment length was >200 amino acids, or if the identity was >60% in case of PVCV and the two selected florendoviruses (Geering et al., 2014) or >45% for all other *Caulimoviridae.*

a



b

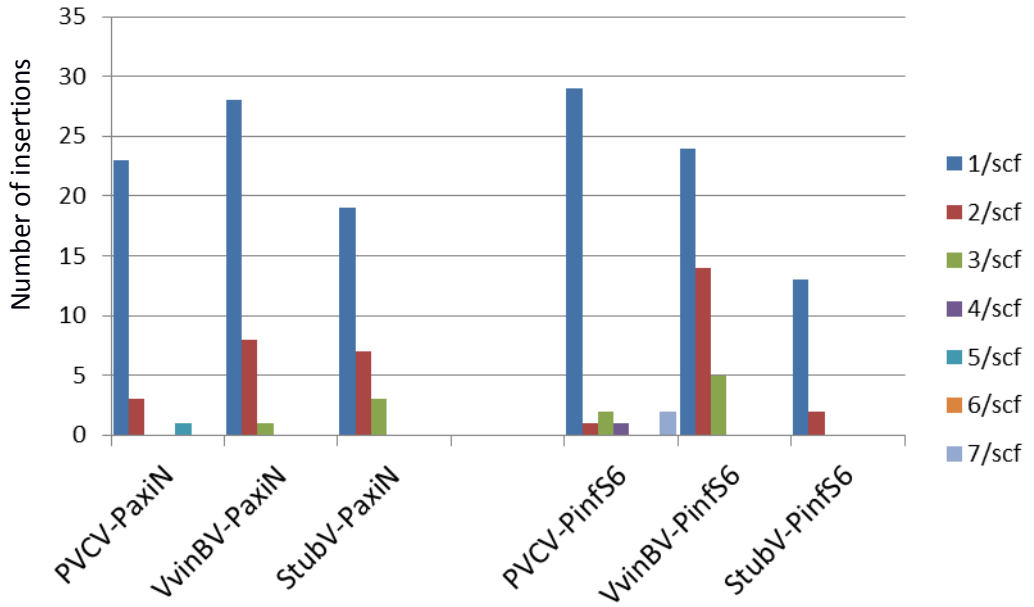


Figure 5: Distribution of EPRV pol-like sequences in *PaxiN* and *PinfS6* assemblies.

a. Numbers connected with columns indicate average amino acid identity in % of the identified pol region compared to the corresponding EPRV.

b. Insertion numbers per scaffold (scf) of pol-like petu- and florendovirus sequences PVCV: *Petunia vein clearing (Petu-) virus;* CaMV: *Cauliflower mosaic (Caulimo-) virus;* CsVMV: *Cassava vein mosaic (Cavemo-) virus;* RYVV: *Rose yellow vein virus* (unassigned genus)*;* ArCVCV: *Aristotelia chilensis vein clearing virus (unassigned genus).* BSOLV: *Banana streak OL (Badna-) virus*; RTBV: *Rice tungro bacilliform (Tungro-) virus;* VvinBV: *Vitis vinifera B (Florendo-) virus isolate -compAsc*1 and StubV: *Solanum tuberosum (Florendo-) virus isolate* -scSt1. The corresponding known episomal viruses displaying different particle morphologies: isometric shape for PVCV, CaMV, CsVMV and RYVV, bacilliform shape for BSOLV and RTBV. For ArCVCV, VvinBV and StubV no information on episomal virus or particle morphology is available.

The two petunia genomes differ slightly in content and diversity of pol-like regions homologous to endogenous pararetroviruses (EPRVs, Figure 5a). In most cases, only one conserved motif was found on a scaffold (Figure 5b). The most abundant pol-like consensus sequences in both genomes, belonged to Petu- and Florendoviruses. Those showed also a high degree of conservation illustrated by 68-75% of sequence identity (Figure 5a). Multiple insertions of three different EPRVs with homology to *Petu-*, *Badna-* and *Florendoviruses* were only found in *PinfS6* scf00909. PVCV-like sequences with 4 separated insertions within the first quarter of *PinfS6* scf00909 were accompanied by 2 StubV-like and 1 BSOLV-like sequences closer to the 3' end of the scaffold. In *PinfS6* scf00073 the two *Florendoviruses* were found at opposite ends of the scaffold with one insertion each. In *PinfS6* scf07983 a StubV-like pol-region was found closer to the middle of the scaffold whereas two VvinBV-like insertions were positioned at the 5' end and 3' end respectively. Only *PaxiN*_scf00380 harbored a combination of the two investigated *Florendoviruses* at its 3'end consisting of two StubV-like pol-region and one VvinBV-like element. The observed higher variability regarding numbers and conservation among *Florendoviruses* in *PinfS6* compared to *PaxiN* (Figure 5) might indicate various time points of invasion.

Other EPRV-like sequences were more degenerate with respect to the full-length virus sequence, showing an average of 49% identity. Multiple occurrences of the same or distinct elements on single scaffolds occurred less frequently, and the motifs were separated by several kb, and did occur in tandem repeats. Examples of linked elements include scf00027 of *PaxiN* with both PVCV-like and CaMV-like domains; scf00654 with both PVCV- and RTBV-like sequences; scf00514 of *Pinf6* with both CaMV-like domains and CsVMV-like pol domains.

Retrotransposons

LTR-STRUCT (McCarthy et al. 2003) was used for *de novo* retroelement searches in the assembled scaffolds. In *P. inflata* S6, a total of 595 RT (reverse transcriptase) active site types, 914 PBS (primer binding site) types, and 996 PPT (polypurine tract) 5'-end types were found from a total of 7354 LTR retrotransposons of which 4147 had RT domains. In *PaxiN*, 4573 LTR retrotransposons were found of which 1,850 included an RT region. These ranged between 1,183bp and 24,737 bp and had LTRs of 76bp to 5344bp. They were classified into *Ty3/Gypsy* (*Metaviridae*) superfamily and *Ty1/Copia* (*Pseudoviridae*)-superfamily elements by their gene order; *Gypsy* elements having the order RT-RH-INT (integrase) while in *Copia* elements the order is INT-RT-RH (see Hansen and Heslop-Harrison, 2004). 52% of the retroelements were categorized as *Gypsy* superfamily, and most of the remaining were *Copia s*uperfamily elements. Twelve selected elements were annotated in Geneious with a local database of motifs taken from the LTR-STRUCT analysis and from Hansen and Heslop-Harrison (2004). The most common retroelement families including *Athila* and *Cyclops Gypsy*-like elements and *BARE Copia*-like elements were found, but some showed various insertions, deletions and inversions (see examples in Figures 6 and 7).

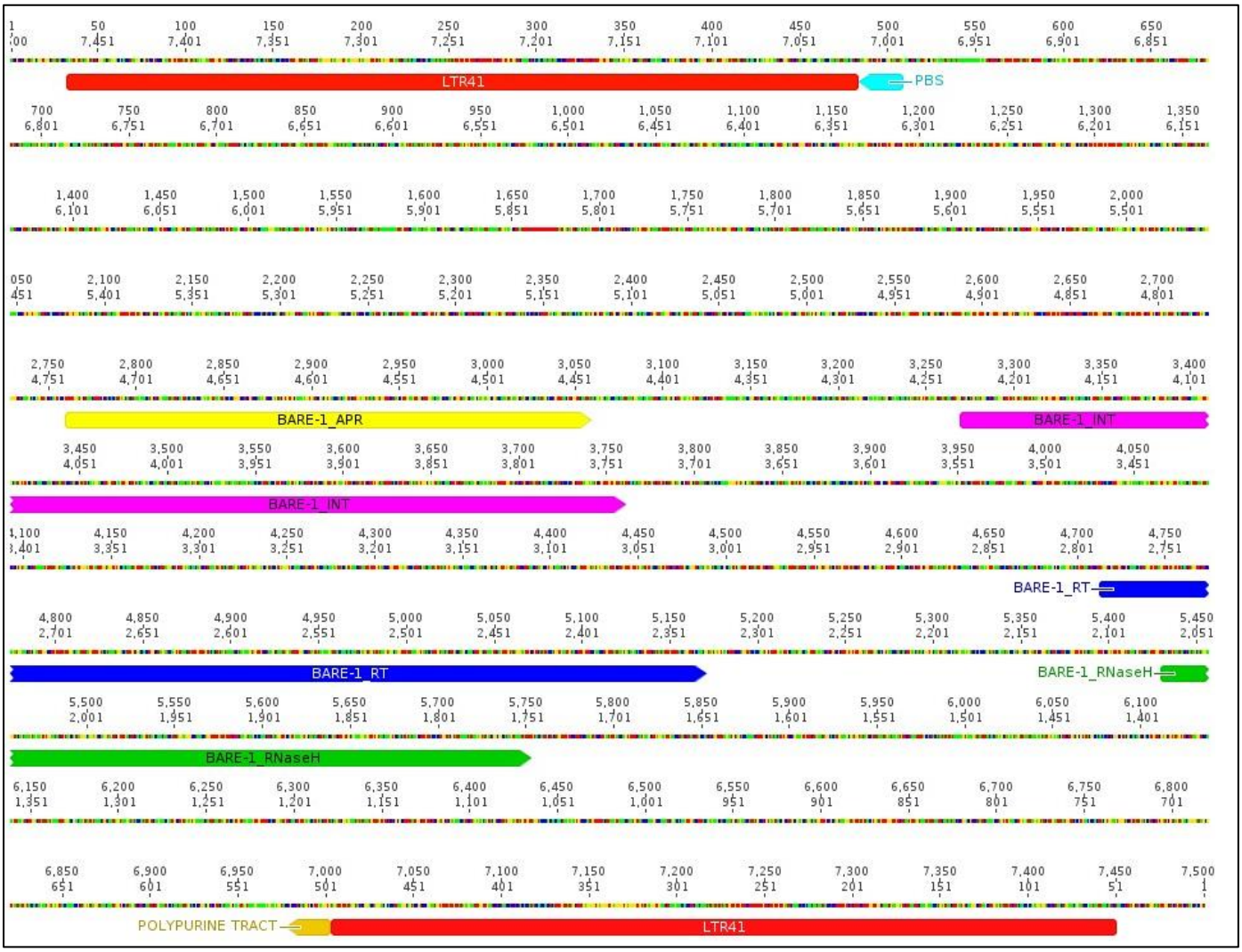


Figure 6: Annotation of the *Copia BARE*-like LTR retroelement 86. It was identified by LTR_STRUC analysis of *PaxiN* assembly Peaxi162 with Ps20000_RT19_B1_L41_86, is 7418 bp long and contains LTR 41 (red) at each end (447bp and 451bp with 96.9% homology), protein binding site (PBS, cyan), the aspartic protein gene (BARE-1_APR, yellow), integrase (BARE-1_INT; magenta), a functional RT region (BARE-1_RT, blue) with the putative active sites SYDDVLF and YVDDILM, RNaseH (BARE-1_RNaseH, green), and a polypurine tract (orange).

Figure 7: Annotation of the *Cyclops Gypsy*-like element 414. It was identified by LTR_STRUC analysis of *PaxiN* assembly Peaxi162 with Ps15485_RT49_B65_L129_414; it is 13252 bp long and contains LTR 129 (red) at each end (1307bp and 1306bp with 96.9% homology), primer binding site (PBS, cyan), integrase (Athila_int and overlapping split Cyclops_INT, magenta), RNaseH (Cyclops_RNaseH, green), a functional RT region (Cyclops_NT, blue) with the putative active sites FLDDLLF and WLDDGII, that is inverted between nt5,708 and nt 11,150, an aspartic protease motif (Cyclops_APR, yellow) and a polypurine tract (orange)

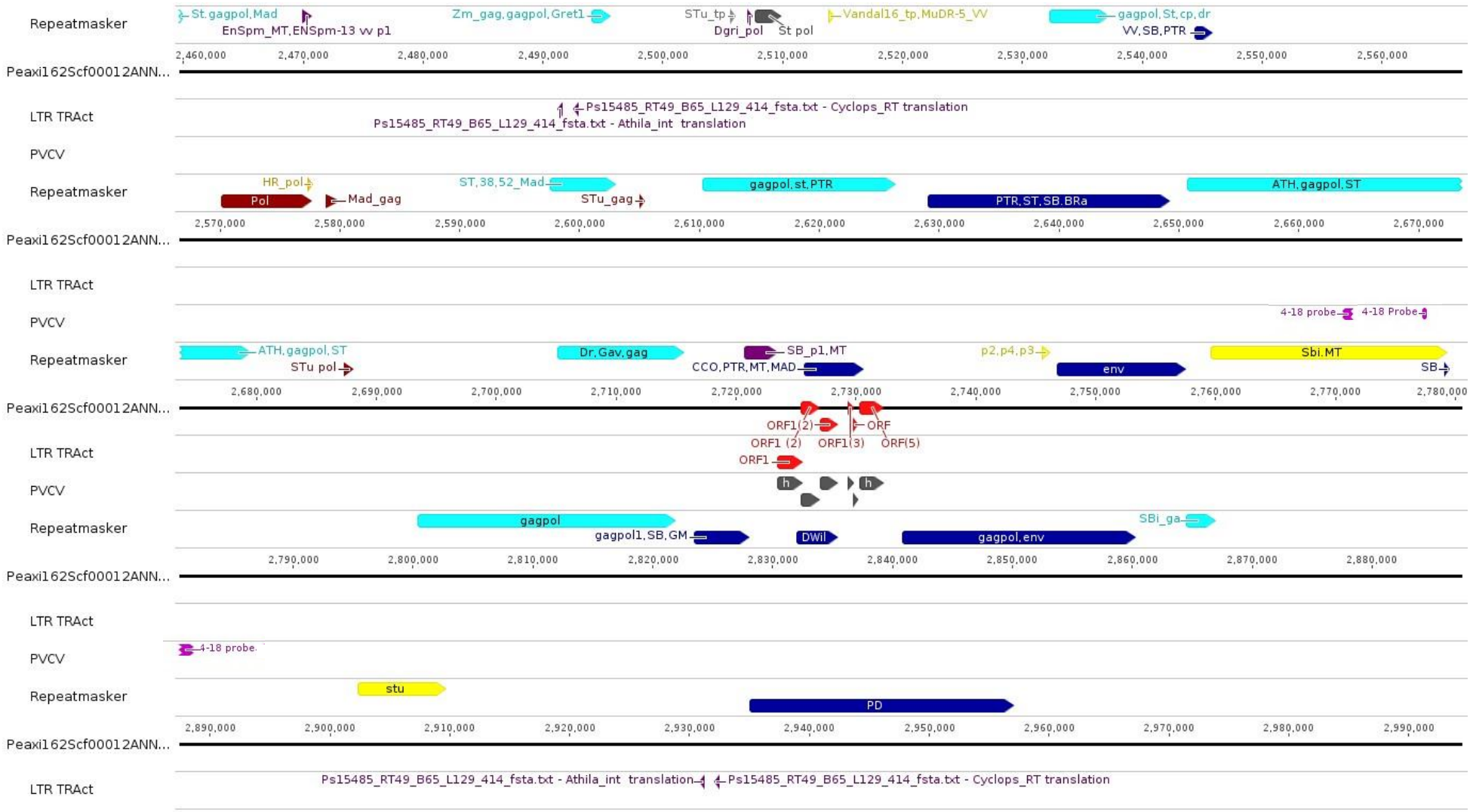


Figure 8
Annotation of *PaxiN* scf00012 from nt 2,460,000 to 2,990,000 using Geneious software with LTR-STRUC analysis (LTR TRAct, PVCV ORFs in red), Repeatmasker, manual alignment with reference transposon sequences from Hansen and Heslop-Harrison (2004), PVCV (grey) and the *Gypsy* super family retroelement junction fragment 4-18 (purple), identification from Tables 1 and 4

Retroelement and endogenous PVCV relation

During the analysis of lambda clones obtained from screening a genomic DNA library of *P. hybrida* (Richert-Pöggeler et al., 2003) it was noted that *Metaviridae* (LTR-*Gypsy* superfamily) sequences were adjacent to integrated PVCV sequences. One such sequence, the 1.2 kb *Gypsy* superfamily retroelement junction fragment 4-18 (1,233bp;nt 4483-5715 of GenBank AY333912, *P. hybrida* lambda clone 4, Richert-Pöggeler et al., 2003) was used for FISH experiments to *P. axillaris* chromosomes (Figure 3c and d). A strong signal is visible at the centromeres of all chromosomes and next to or interspersed with both the strong PVCV signal on chromosome III and the weaker PVCV signal on chromosome VI. To further analyse the surrounding sequences of PVCV, scf00012 of *PaxiN* was annotated (Figure 8). Many gag-pol regions indicative of LTR retroelements were found including some in the immediate vicinity of the PVCV tandem array.

Because of the distribution of fragment 4-18 (Figure 3c and d), scaffolds of *PaxiN* assembly at least 500kb in length were searched by BlastN to identify sequences related to fragment 4-18 that were at least 150 bp long. Within scaffolds scfs0000-scfs0999, 635 scaffolds contained at least one copy, but only 7% of scaffolds had more than a total of 1% of sequences homologous to 4-18 and only 9 clones contained 3-6% (see Figure 9A). Longer scaffolds on the whole contained less 4-18 related sequences, while not all but some shorter scaffolds have more (see Figure 9B). Possibly, this is an under-representation and unassembled shorter reads need to be checked. Results for scaffolds scf00012, scf00095 and scf00097 that also contain PVCV are given in Table 4. Interestingly only in scf00012, a larger number (58, 0.9%) of 4-18 related sequences were found, three copies in the vicinity of PVCV.

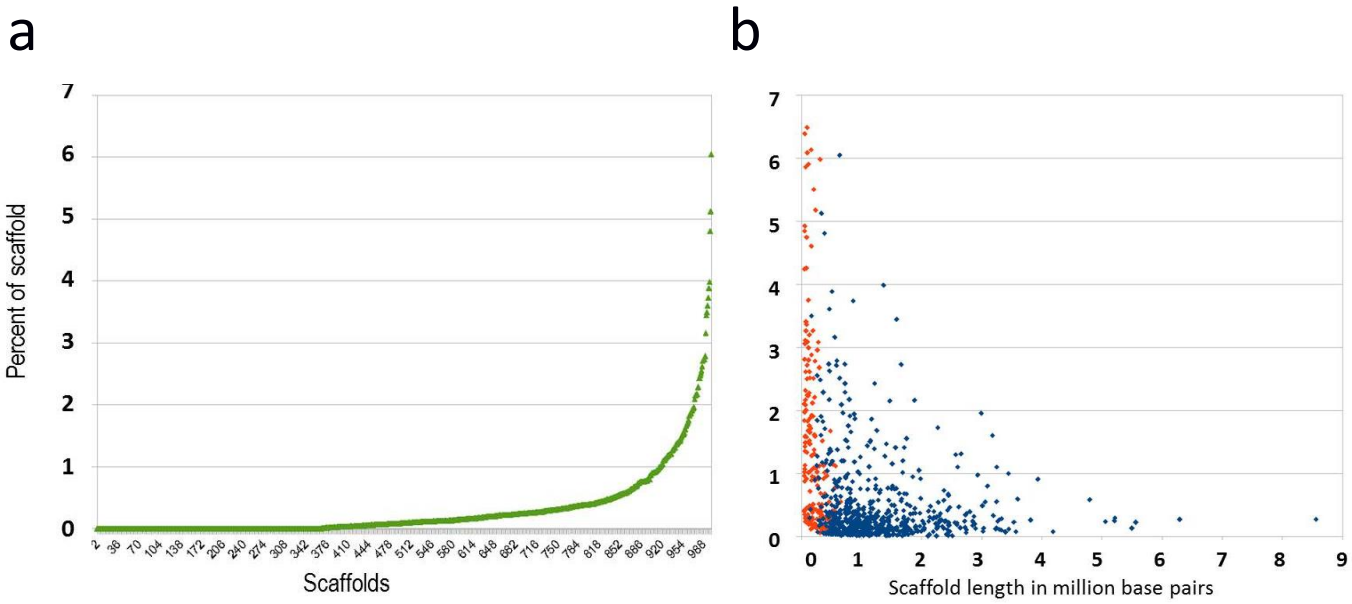


Figure 9: Percentage of sequences longer than 150bp and homologous to *Gypsy* gag-pol 1.2kb fragment 4-18 in *P. axillaris* N scaffolds.

a. Arranged by increasing frequency in scfs 0-999

b. Plotted against scaffold length of all scaffolds with a length greater than 500kb; scaffolds without 4-18 sequences are not plotted; scaffolds 0-999 blue, remaining red

Table 4: *Gypsy* superfamily retroelement junction fragement 4-18 in scf00012, scf00095 and scf00097 of *PaxiN* assembly Peaxi162. These also contain PVCV sequences. Hits longer than 150 bp are listed with start and end for hit (scf) and query (4-18 fragment) as well as total number of hits, length and proportion in scaffold highlighted in yellow. Those sequences in the vicinity of PVCV (±200kb) are highlighted in orange.

Scaffold	caffold length	Sequence length	% scf	Hit end (nt)	Hit start (nt)	Query nd (nt)	Query art (nt)	Pairwise identity	E value
scf00012		187		51,828	51,642	1219	1405	84.8%	1.73E-48
scf00012		522		88,675	88,154	465	937	67.5%	8.97E-46
scf00012		670		154,346	153,677	3	663	73.2%	1.16E-107
scf00012		649		159,814	160,462	6	612	66.0%	1.02E-38
scf00012		455		174,278	174,716	1324	1778	78.0%	3.56E-95
scf00012		409		180,259	180,650	1370	1778	78.9%	4.06E-88
scf00012		1113		230,971	232,083	3	1105	85.0%	0
scf00012		938		389,698	390,635	6	889	65.2%	1.97E-60
scf00012		700		399,287	399,986	6	663	66.5%	2.57E-46
scf00012		205		424,250	424,046	1410	1614	83.3%	7.36E-47
scf00012		353		430,032	429,681	1426	1778	76.0%	1.17E-50
scf00012		342		430,554	430,222	1324	1665	76.1%	2.74E-52
scf00012		1120		502,183	503,298	1	1120	85.4%	0
scf00012		992		535,999	535,008	6	907	63.9%	4.96E-49
scf00012		700		596,817	596,118	6	663	65.9%	3.57E-38
scf00012		1110		665,804	666,913	1	1105	85.2%	0
scf00012		695		682,086	681,392	6	663	66.9%	4.96E-49
scf00012		385		716,677	717,061	6	372	71.2%	2.11E-47
scf00012		491		724,892	725,382	495	936	65.6%	4.96E-30
scf00012		1149		738,774	737,626	1	1120	79.2%	0
scf00012		977		760,183	761,159	6	923	63.1%	4.96E-30
scf00012		415		774,848	774,464	1364	1778	74.6%	1.97E-60
scf00012		564		791,730	791,230	1215	1778	69.2%	5.29E-55
scf00012		1105		869,720	870,819	1	1105	86.3%	0
scf00012		984		915,597	916,580	6	936	67.3%	4.64E-81
scf00012		237		954,226	953,991	1191	1427	84.8%	3.81E-63
scf00012		913		957,558	956,646	21	897	70.6%	2.57E-65
scf00012		434		961,128	960,709	1219	1652	80.5%	6.44E-92
scf00012		698		962,392	961,695	6	663	65.7%	1.02E-38
scf00012		954		985,715	984,762	6	907	63.5%	1.42E-30
scf00012		608		990,652	991,224	1171	1778	72.3%	6.88E-79
scf00012		199		1,010,491	1,010,688	465	663	75.1%	2.57E-27
scf00012		976		1,022,877	1,021,902	6	935	67.9%	1.33E-100
scf00012		589		1,082,067	1,082,655	37	616	75.4%	6.43E-111
scf00012		266		1,083,013	1,083,278	674	937	71.9%	4.65E-24
scf00012		345		1,157,369	1,157,025	1435	1778	75.7%	4.07E-50
scf00012		265		1,157,894	1,158,158	674	937	72.1%	1.09E-25
scf00012		1036		1,177,630	1,176,595	6	937	64.1%	4.96E-49
scf00012		265		1,179,759	1,179,495	674	937	73.1%	4.96E-30
scf00012		234		1,180,756	1,180,523	6	230	71.8%	2.57E-27
scf00012		388		1,183,735	1,183,348	6	375	68.4%	1.42E-30
scf00012		229		1,206,353	1,206,126	3	231	72.5%	3.13E-26
scf00012		229		1,218,306	1,218,079	3	231	73.9%	6.05E-29
scf00012		487		1,245,074	1,245,539	1169	1655	68.5%	2.75E-33
scf00012		458		1,257,141	1,257,598	1	449	83.4%	2.4E-135

Table 4 cont

Scaffold	Scaffold length	Sequence length	% scf	Hit end	Hit start	Query end	Query start	Pairwise identity	E value
scf00012		296		1,281,496	1,281,783	1483	1778	75.7%	5.66E-42
scf00012		408		1,288,213	1,288,613	1371	1778	72.7%	4.96E-49
scf00012		1332		1,298,230	1,299,509	442	1773	73.2%	0
scf00012		810		1,327,235	1,326,430	311	1120	85.1%	0
scf00012		288		1,328,193	1,327,906	1	288	89.6%	2.1E-104
scf00012		1134		1,372,898	1,371,765	1	1120	84.8%	0
scf00012		648		2,665,696	2,666,283	1100	1747	85.8%	0
scf00012		179		2,671,061	2,671,239	1597	1775	99.4%	2.56E-84
scf00012		976		2,881,365	2,882,340	34	937	64.9%	1.42E-49
scf00012		949		3,006,423	3,005,475	11	886	65.1%	5.29E-55
scf00012		691		3,401,387	3,402,077	6	663	67.1%	7.86E-34
scf00012		695		3,479,170	3,478,476	430	1120	85.8%	0
scf00012		347		3,480,607	3,480,261	1	347	89.1%	4.94E-125
scf00012 SUM	3,935,541	35793	0.9095%						
number of hits		58							
scf00095		234		475,877	476,110	231	6	72.3%	7.37E-28
scf00095		195		484,575	484,381	231	37	73.8%	3.82E-25
scf00095		492		1,094,204	1,094,695	937	493	65.9%	1.42E-30
scf00095		226		1,094,981	1,095,206	231	6	74.8%	7.86E-34
scf00095		700		1,483,284	1,483,983	663	6	67.4%	2.74E-52
scf00095		1192		1,491,641	1,492,832	1191	1	82.3%	0
scf00095		231		1,541,528	1,541,758	936	706	73.4%	8.97E-27
scf00095		691		1,542,092	1,542,782	663	6	65.0%	3.82E-25
scf00095		668		1,551,339	1,550,672	663	1	75.4%	4.06E-126
scf00095		265		1,558,575	1,558,839	935	674	70.6%	1.33E-24
scf00095 SUM	1,774,960	4894	0.2757%						
number of hits		10							
scf00097		990		193,085	192,096	923	6	62.7%	1.33E-24
scf00097		640		197,909	198,548	663	39	69.8%	1.73E-67
scf00097		700		258,127	257,428	663	6	66.8%	4.07E-50
scf00097		233		323,849	324,081	906	674	75.1%	1.52E-36
scf00097		473		324,118	324,590	889	465	65.3%	3.13E-26
scf00097		235		324,848	325,082	231	6	71.9%	7.37E-28
scf00097		697		508,542	507,846	663	6	66.7%	2.74E-52
scf00097		265		509,117	508,853	937	674	75.0%	4.35E-37
scf00097		1200		576,557	575,358	1199	1	84.2%	0
scf00097		1199		856,964	858,155	1199	1	79.5%	0
scf00097		179		1,304,062	1,303,891	1773	1595	80.1%	4.96E-30
scf00097		1195		1,308,810	1,307,626	1195	1	86.5%	0
scf00097		179		1,309,276	1,309,099	1773	1595	82.5%	1.25E-37
scf00097		422		1,416,974	1,416,553	1778	1368	75.3%	4.64E-62
scf00097		280		1,489,752	1,490,031	937	674	70.9%	6.05E-29
scf00097		199		1,490,342	1,490,540	663	465	74.8%	3.82E-25
scf00097		195		1,850,949	1,851,143	231	37	74.4%	8.97E-27
scf00097		863		2,225,747	2,226,609	1105	249	85.8%	0
scf00097	2,428,612	10144	0.4177%						
number of hits		19							

K-mer analysis in *P. axillaris* N

K-mer frequencies (measurements of the occurrence of each sequence motif k bases long in raw read data) are independent of assembly algorithm and thus an unbiased method to access the repetitive portion of a genome. Genome repetivity was assessed via 16- and 32-mer frequencies (Figure 10a and b) as in the tomato genome (Tomato Genome Consortium, 2012, Supplementary Figure 42) and were overlaid onto the tomato, potato and sorghum data (Figure 10c). The *P. axillaris* genome is larger than tomato and potato, so it should be expected that the slope indicates larger repetivity, and 16-mers occurring ≥10 times account for 50% of the genome, approximately twice the frequency in the smaller solanaceous genomes. However, interestingly for 16-mers occurring ≥ 30 times, the *Petunia* slope follows sorghum, even though the sorghum genome is only half the size of *Petunia*. It is notable that the top 10-20 most frequent k-mers are composites of AT, AAT, AG, A, C and AAG microsatellites. Larger abundant k-mers were also analysed, and for example some 54-mers and 64-mers have a few thousand repeats in the genomes, but only short tandem arrays with 10-20 copies were found in scaffolds suggesting that the assembly had collapsed some arrays of near-identical repeats.

In an attempt to identify larger tandem repeats in *P. axillaris*, 128-mers (Table 5) that were repeated more than 1001 times (in total 2347 of them) were subjected to a *de novo* assembly. They assembled into contigs with the largest of 1943 bp, 531 bp, 185 bp and 175 bp. The contig of 1943 bp has a total of 8,217 hits in *PaxiN* assembly and was found to be part of a *Gypsy* superfamily LTR-retroelement. There are some small duplications/tandem repeats within this large repeat and it is found as single repeat unit in most of the large scaffolds of *PaxiN* and also in the GenBank accession AY136628 of *P. hybrida*. The second 128-mer contig contains more of a repeat structure and hits to three large *Petunia* sequences in GenBank (AY136628, AB472856 and EF517793) and is present in 530 of the *PaxiN* contigs (0.8%) with high homology of near 100%. Dotplots of pairs of these scaffolds indicated that there is a larger repeated unit, up to 8 kb – with homology to a *Gypsy* superfamily retroelement with RNaseH, RT, INT, LTRs and other domains.

Table 5. Assembly using Geneious assembler of 128-mer identified in *PaxiN* raw reads.

	Unused reads	**Contigs >=128 bp**	**Contigs >=1000 bp**
Number	2	10	1
Minimum length (bp)	128	129	1,943
Median length (bp)		143	
Mean length (bp)	128	363	1,943
Max length (bp)	128	1,943	1,943
N50 length (bp)		1,943	
Number of contigs >= N50		1	1
Length sum (bp)	256	3,638	1,943

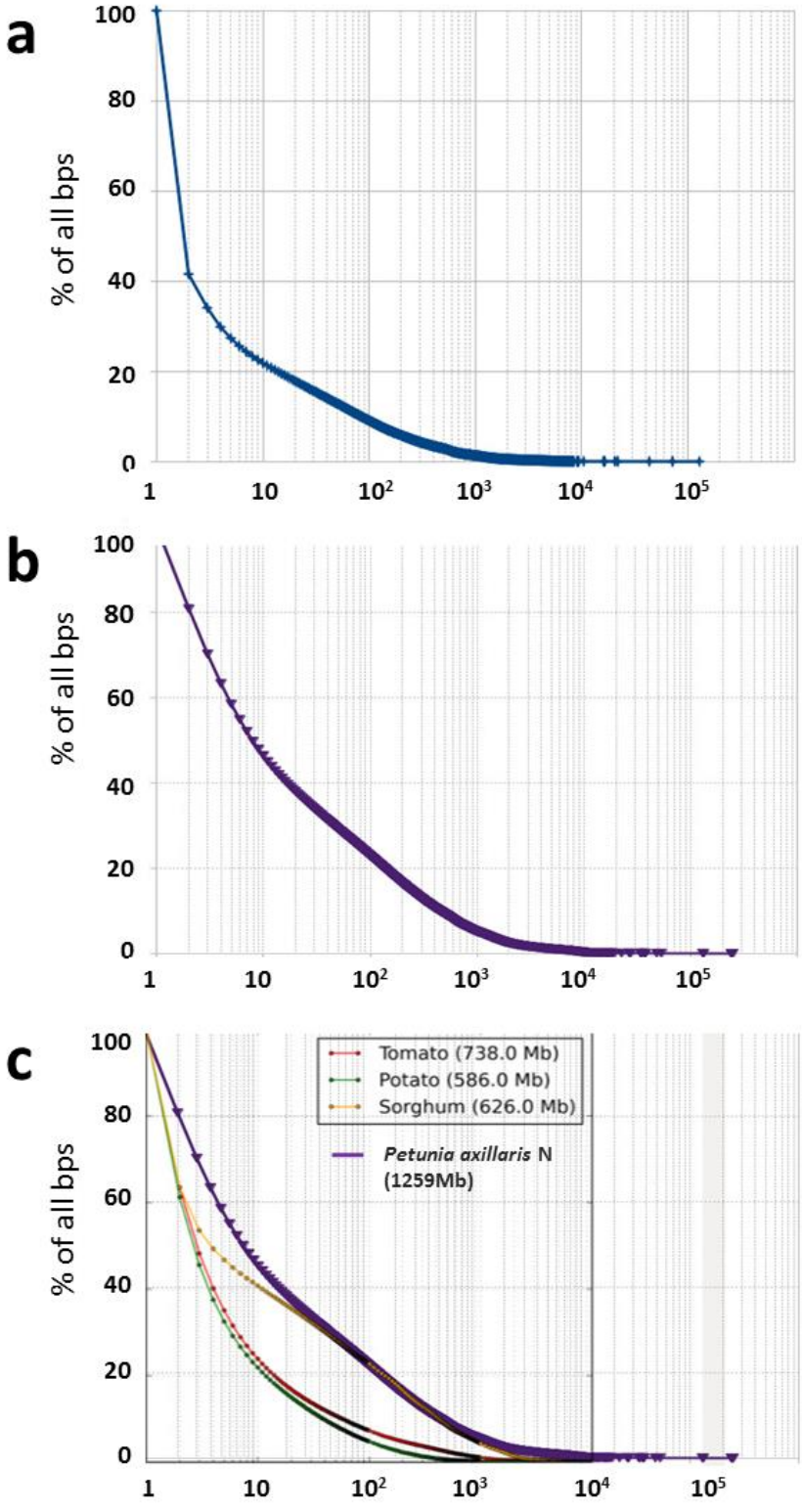


Figure 10: ***P. axillaris N*** **genome repetivity analyzed by k-mer frequency in raw reads**

a. 32-mer frequency.

b. 16-mer frequency

c. 16-mer data overlaid onto graph from tomato (Tomato Genome Consortium, 2012, Supplementary Figure 42). Genome size is given in Mb of assembled genomes.

Telomeres, tandem repeats and larger repeats

A few scaffolds included multiple copies of the plant telomere sequence TTTAGGG; some occurrences appear in intercalary positions. For example, *PaxiN* scf02211 was 12 kb long with telomere motifs at both ends totalling 2.5 kb, while scf47951 with a length of 2.5 kb was mostly composed of degenerate telomere arrays. In scf01230, telomere tracts up to 3 kb long were part of a longer tandem repeat unit of 11,246 bp, while scf03941 had a 180bp tandem repeat next to the telomere tract.

Another tandem repeat was extracted from *PaxiN* scf02038 and is 169 bp long. Many hundreds of smaller scaffolds were composed largely of a 169bp repeat (eg scf36935 is 16.5 copies over 2795 bp; scf01515 has 22 copies over 3778 bp; scf1920 has c.64 copies). Of larger scaffolds, scf01294 (409 kb) ends with 12 copies, scf00744 (498 kb) has 21 internal copies, scf00700 (826 kb) has 41 copies at the end over 7 kb (both orientations). Further, scf00420 has copies at the end, scf00451 an internal array, scf00286 multiple dispersed short arrays, and scf00207, scf00160, scf00128, scf00074 and scf00003 all have multiple copies. However, no scaffolds were found that could be related clearly to centromere structures, nor were any scaffolds candidates for giving the distribution of the fragment 4-18 on chromosomes (Figure 8 and above): the FISH results show a strong signal around the centromeres suggesting large numbers of a repetitive DNA motif. Two scaffolds could be candidates to locate around centromeres (e.g. Figure 11 for *PaxiN* scf00160), but further analysis including FISH and targeted cloning are needed to establish the position and distribution of these sequences, and identify them as centromere related.

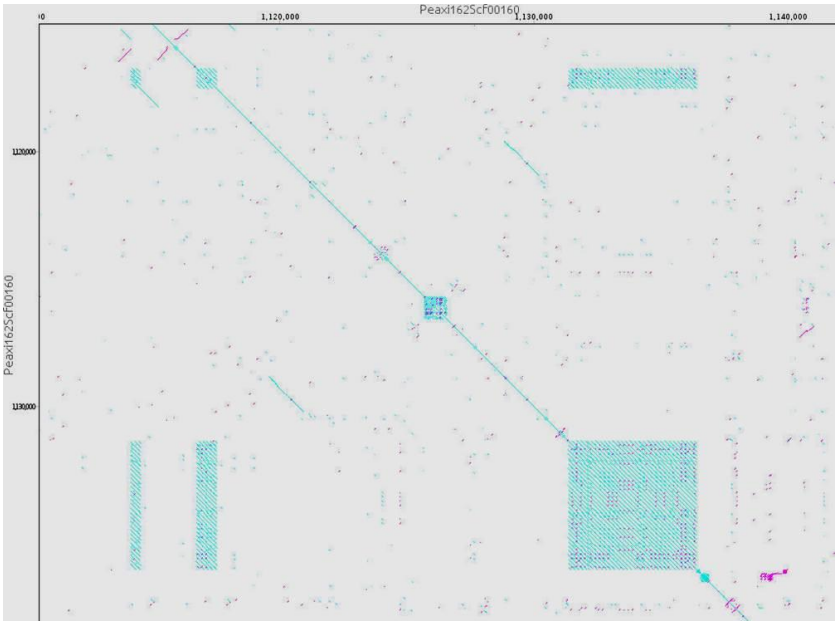


Figure 11: Tandem repeat structure.
Dotplot of *PaxiN* scf00160 nt1,110,000 to 1,140,000 with (c. 160 bp monomer) showing longer and shorter units.

Repeat analysis

As well as the k-mer analysis, genome-wide characterization of repetitive elements can use graph based clustering of DNA sequences from raw reads (Novak et al., 2010; 2013). The program RepeatExplorer was run using unassembled reads. For *P. axillaris N* 3,833,689 reads were selected

from an Illumina run, and for *P. inflata S6* 3,229,900 reads were examined. They resulted in 195,174 clusters using 2,488,367 reads for *P. axillaris,* about 65% of the analysed data pool (Figure 12a) and for *P. inflata*, 198,865 clusters using 2,511,217 reads, about 78% of the analysed data pool (Figure 12B); the remaining were not clustered. No cluster in either genome dominated the analysis, with the top 5 clusters representing 3.6% in *PinfS6* and 3.3% in *PaxiN* of the total and then declining in frequency gradually over several hundred motifs (Figure 12), contrasting with other genomes where a high proportion of repeats are represented by a small number of clusters (e.g. cacao, Sveinsson et al., 2013). RepeatExplorer depicts clusters graphically as connected dots; protein domains of transposable elements are colour coded. The total number of base pairs, reads and genome proportion are calculated; in addition the hits to known repeats present in the repeat masker database are identified. In Figure 13 the most frequent clusters of *PaxiN* and two most frequent clusters of *PinfS6* are shown. The most striking feature is that clusters are made up of a varied composition of sequence types, including LTR and low complexity repeats. The programme identified many further complex clusters, enriched in degenerate LTR-*Gypsy* and LTR *Copia* elements. Tandem repeats, rDNA and simple repeats were included in a few clusters.

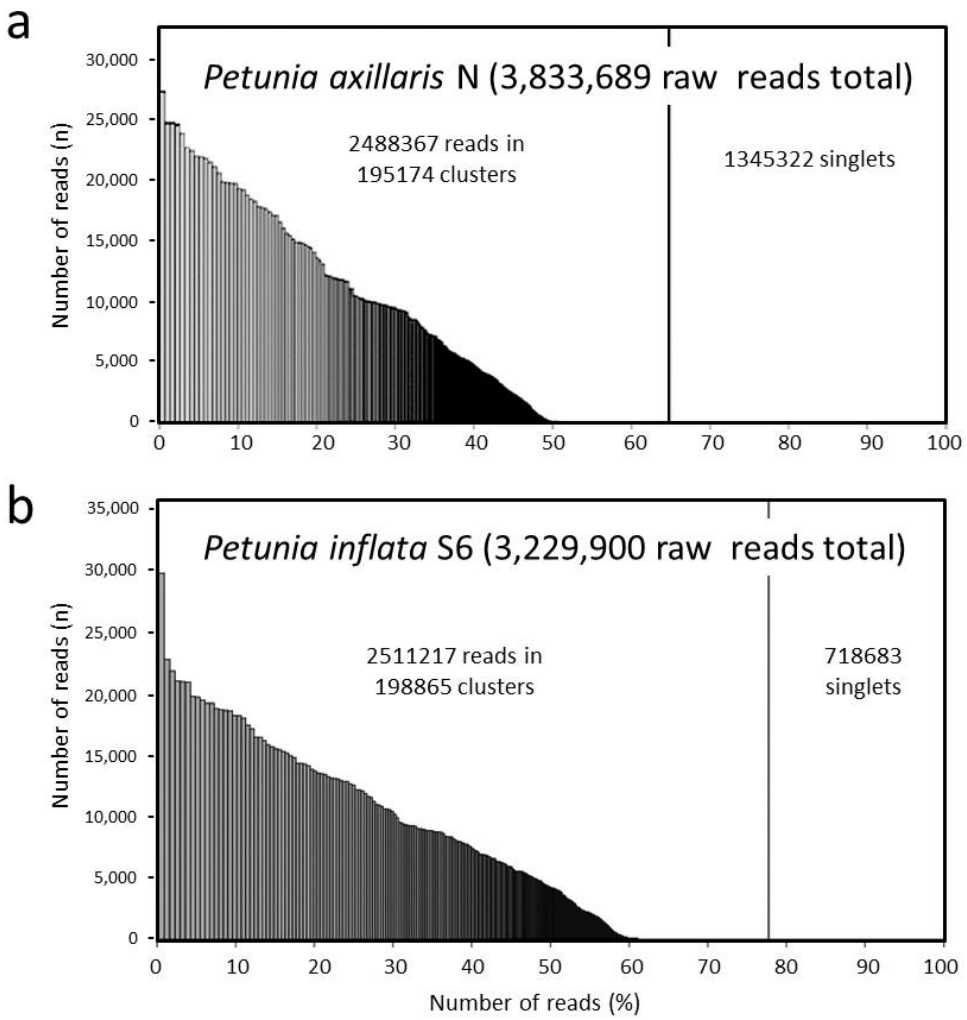


Figure 12: **Summary of Repeat Explorer (Novak et al., 2010, 2013) analysis using 3 million randomly selected raw reads a.** from *P. axillaris* N *and* ***b.*** *P. inflata* S6.

The most common clusters are not greatly more abundant than subsequent clusters. The top 350 clusters (all >0.01% proportion of the genome) represented 65 to 78% of all repeat clusters and include 50% (*P. axillaris* N) and 60% (*P. inflata* S6) of the genome.

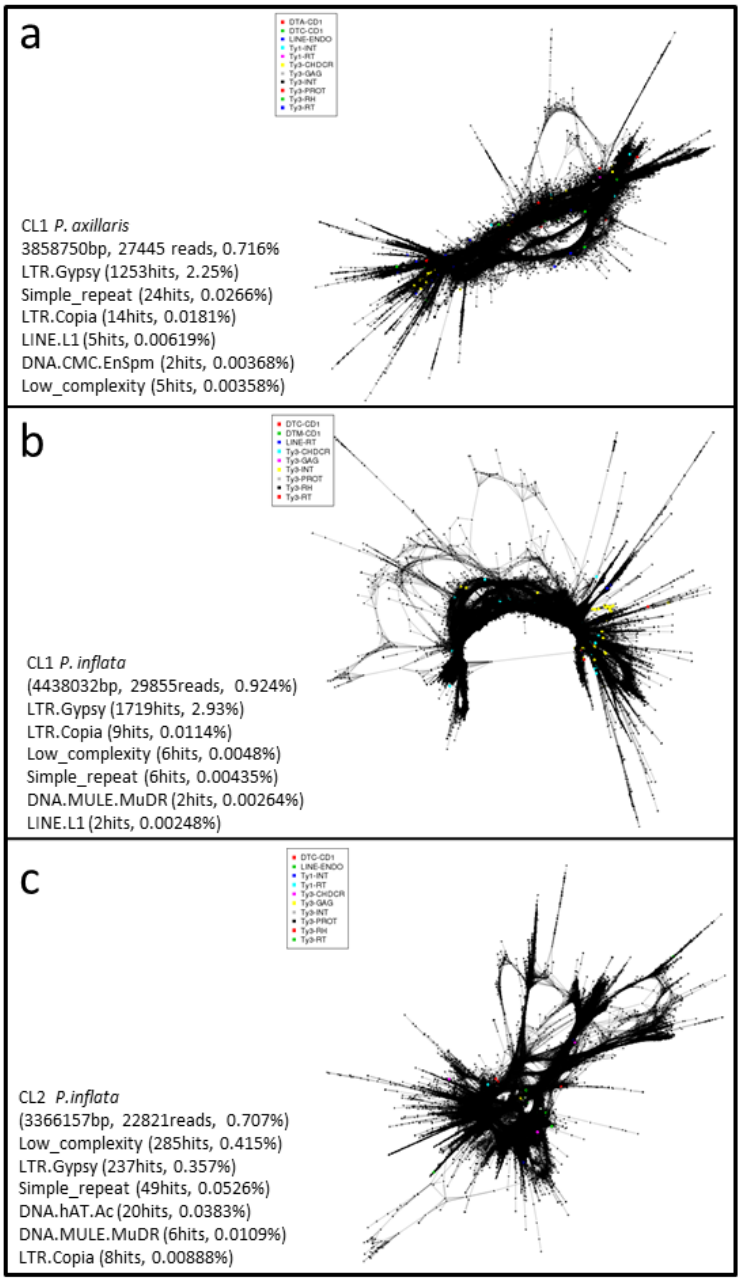


Figure 13: Most frequent repeat clusters CL1 of *P. axillaris* N (A) and CL1 and CL2 of *P. inflate* S6 (B, C). Output from Repeat Explorer (Novak et al., 2010, 2013). The clusters are displayed as composites containing low complexity and degenerate LTR-transposable elements with no abundant retroelement protein motifs identified (shown by the few coloured dots). Each cluster is less than 1% of the genome.

DISCUSSION

Members of Solanaceae are known to be suitable hosts for a number of plant viruses. More than 150 plant viruses have been shown to infect *P. hybrida* (Engelmann and Hamacher, 2008). Plant genome sequencing indicates that several *Caulimoviridae* invaded the genome of their host in the course of viral infection (Tomato Genome Consortium, 2012; Kim et al., 2014; Geering et al., 2014). The identified single insertion sites within the *P. axillaris* N and *P. inflata* S6 genome respectively are suggestive of provirus stages typical for retroviruses and endogenous retroviruses. Integration of retroviruses requires an integrase and is enhanced in actively dividing cells (Young et al., 2013). The high regeneration potential of Solanaceae species like *Petunia* might promote capture of plant pararetroviruses of the family of *Caulimoviridae* that lack an integrase from the *Petunia* genome. De-differentiation of pararetrovirus infected somatic cells as been happening during callus proliferation and plant regeneration might offer gateways for invasion followed by vertical transmission (Hohn et al., 2008). Our analysis provides insight into the elements which have been manifested in the male

and female germline lineages (Schmidt et al., 2012) and we expect that the frequency of pararetroviral invasion in chromosomal DNA of somatic cells is much higher. We have speculated in the past (Richert-Pöggeler and Schwarzacher, 2009; Staginnus and Richert-Pöggeler, 2006) that the close proximity of endogenous pararetroviruses and LTR-retroelements points to a co-evolution of these two similar elements as well as to their integration and silencing mechanisms. The available sequence data of two petunia species prove *Petunia* as an ideal model system to study endogenous pararetrovirus co-evolution within solanaceous hosts as well as potential functions besides being an infectious entity. Furthermore EPRVs can be suitable markers for monitoring dynamic processes during genome hybridisation as happened during generation of *P. hybrida*. Both genomes contain insertions of regulatory sequences (*PaxiN* scf01628, *PinfS6* scf00235 and *PinfS6* scf00276) from the untranslated region of the PVCV genome that await further analysis of their functionality in gene expression and reverse transcription respectively. Recent studies by Mushegian and Elana (2015) discuss molecular functions for the host provided by integration of pararetroviral movement protein (MP)-like sequences. Based on their phylogenetic analyses and known MP functions in macromolecule trafficking they propose a possible role of integrated MP in control of tissue differentiation. Indeed, in *P. axillaris N* integrated MP sequences of scf01628 and scf03256 showed the highest degree of conservation and thus may be active in the host rather than viral context.

The pericentromeric localization and array structure of integrated PVCV sequences is similar in both petunia species. However, the preservation and copy number of these endogenous viral sequences is higher in *P. axillaris* compared to *P. inflata*. That was also true with regard to diversity of EPRVs. Thus *P. axillaris* seems to be a more permissive host for EPRV invasion, so we suggest that there are several factors controlling EPRV invasion and preservation that differ even between related species, and probably even cultivars, including the organization in the genome, presence of miRNA, DNA methylation and histone modifications; such factors may be under evolutionary selection depending on disease pressure and consequences for the different species.

Our analysis indicates that *Petunia* genomes are rich in repetitive DNA and the K-mer analysis of *P. axillaris* N indicates more repeats are present than in tomato and potato (Tomato Genome Consortium, 2012). However, when comparing total genome size, the amount of repeats in both the Pa*xi*N and *Pinf*S6 assemblies with about 60-65% (Table 6) is low for genomes of 1.4Gb. In particular LTR-retroelements are unusually low both in total DNA bps and number full length elements that we were able to identify. We found about 5,000-6,000 full elements with roughly equal numbers of *Gypsy* and *Copia* superfamily retroelements. This number is similar to tomato with a genome assembly of 740Mbp, but a relatively large LTR-retrolement component. It is in contrast to hot pepper (*Capsicum annuum*) with the largest solanaceous genome (more than 3Gb) so far sequenced where large numbers of LTR-retroelements in particular *Gypsy* superfamily elements make up 70% of the repetitive DNA fraction, are responsible for the genome expansion and conversion of euchromatin into heterochromatin (Kim et al., 2014). *Nicotiana* species also have relatively few identified LTR-retroelements (Table 6), and here *Copia* superfamily retroelementss make the difference between *N. tomentosiformis* and *N. sylvestris* (Sierro et al., 2013). Interestingly, *N. tomentosiformis* and *N. sylvestris* can also be distinguished by their EPRV composition (Gregor et al., 2004). The difference to the smaller tomato genome however, is mainly attributed to shorter repeats (listed under 'others' in Table 6) that represent almost 30% of the repetitive DNA in the about 2.4Gb *Nicotiana* genomes. This situation is mirrored in the much small cucumber genome (244Mb) where transposable elements are relatively low in abundance, but satellite tandem repeats make up almost half of the repetitive DNA and are concentrated at centromeres and telomeres as evidenced by FISH (Huang et al., 2009). A

similar situation is found in *Brassica rapa* (Brassica rapa Genome Consortium, 2011) and *Beta vulgaris* (Dohm et al., 2014) where satellites repeats have been found (with only two families abundant, locating at centromeres, in *Brassica*, Harrison and Heslop-Harrison, 1994); however, the larger genome sizes of 500-600 Mbp can be attributed to the LTR-retroelement fractions.

The relative low proportion of repeats, make in turn the gene and low copy sequence space relative larger and would correspond to the lower fragmentation found in *P. axillaris* N after the last triplication event (paleohexaploidisation) in comparison to tomato and potato (see Supplementary Note 5).

In *Petunia*, FISH has indicated a concentration of retro-elements like sequences around the centromeres (this study) as well as their dispersion throughout the chromosomes (Richert-Pöggeler and Schwarzacher, 2009), but this is not reflected in the assembly, as is normal with shotgun sequence data; even the relatively high proportion of PacBio and mate-pair reads apparently spanning the length of retroelements. However, DNA transposons are common and were found at a much higher frequency than in *Nicotiana* and *Solanum* (Table 6 and Supplementary Note 3). Interestingly, the literature does not report tandem satellite repeats in *Petunia* and our repeat searches within the assembly have not found typical 180bp or 340bp repeats that wrap around nucleosomes in a specific manner (see Heslop-Harrison and Schwarzacher, 2013). However, shorter repeats of about 60bp have been found, as well as some longer repeats of 500-1000bp. In addition, many mixed repeat family clusters incorporating retroelements, simple sequence repeats and low complexity repeats were identified by the RepeatExplorer algorithm, but none present a substantial percentage in the genome in contrast to the RepeatExplorer data in cacao (Sveinsson et al., 2013).

It is therefore apparent that the repeat structure of *Petunia* differs from other species of Solanaceae so far analysed in detail and indicates a high degree of genome plasticity. Genome size alone might however not dictate the distribution, type and amount of repetitive elements. It is notable that *Petunia* chromosomes (Table 6), with an average of 200Mb per chromosome (three times that of tomato or potato), are relatively large for the overall genome size that is distributed over 7 rather than the more common 12 pairs of chromosomes in the family that form the related x=12 clade (Saerkinen et al., 2013). This has consequences for chromosomal organisation, recombination and homogenisation events and together with DNA transposon frequency and the presence of EPRVs might have an effect on the overall genome organisation.

The genomic sequence data will be seminal for investigating interactions of the identified reverse transcribing elements of the family of *Caulimoviridae, Metaviridae* and others both *in planta* as well as at the single cell level and in culture. The identified diversity and abundance of the polymerase motif of viral retroelements raises questions about possible functions of reverse transcription in genome maintenance and/or speciation. Genome sequence data reveal existence of petuvirus-like sequences not only in the solanaceous plant family but also in woody plants, for example in the family of *Rutaceae* (Yang et al., 2003; Roy et al., 2014). The combination of horizontal and vertical transmission among multiple members of *Caulimoviridae* probably contributed to the abundancy of EPRVs within angiosperms (reviewed by Teycheney and Geering, 2010, Geering et al., 2014). Whereas information about the contributing genomes is available, participating vectors mediating transfer of the episomal forms still need to be elucidated. The effects of genome hybridization during generation of *P. hybrida* with *P. axillaris* and *P. inflata* as parental crossing partners, on repeat and in particular EPRV evolution, activation and function can now be studied in greater detail. Thus will also contribute to the general understanding of mechanisms involved in lateral DNA transfer.

Table 6: Comparison of repeat content, total genome and chromosome sizes in *Solanaceae* and selected eudicot species.

Species	Repeats	DNA transposons		LTR retroelements and retrotransposons		non-LTR retroelements (SINES, LINES)		others (satellites, unknown, low complexity)		sequence assembled	1C DNA content	Chromo -some number	DNA/chr
	% assembled genome	bp	%	bp	%	bp	%	bp	%	bp	Mbp	n=	Mbp
Cucumis sativus 1)	24.01	2,808,075	1.24	23,622,636	10.43	3,961,988	1.75	25,762,300	10.58	243,500,000	367	7	52
Brassica rapa 2)	44.79	15,518,826	3.2	131,612,046	27.14	15,925,293	3.28	54,174,500	11.17	485,000,000	560	10	56
Beta vulgaris 3)	42.3	19,820,000	3.33	122,670,000	20.59	32,190,000	5.40	77,320,000	12.98	595,744,681	730	9	81
Solanum lycopersicum 4)	68	6,050,581	0.86	459,739,604	61.77	4,089,807	0.55	31,421,760	4.26	737,600,000	900	12	75
Solanum tuberosum 4)	62.20	6,543,927	1.2	311,628,974	54.35	5,796,327	1.16	32,394,740	5.53	585,800,000	844	12	70
Petunia axillaris N 5)	63.08	65,589,038	5.21	508,788,466	40.41	29,284,495	2.33	190,486,700	15.13	1,259,000,000	1380	7	197
Petunia inflata S6 5)	59.22	59,714,447	4.64	475,871,680	36.98	38,215,801	2.97	188,141,800	14.63	1,286,000,000	1430	7	204
Nicotiana tomentosiformis 6)	74.84	22,593,004	1.34	882,169,158[8)]	52.21[8)]	8,078,343	0.48	571,894,844	20.33	1,689,000,000	2360	12	197
Nicotiana sylvestris 6)	71.95	33,621,895	1.51	1,082197,020[9)]	48.65[9)]	9,869,117	0.44	703,763,729	21.34	2,222,000,000	2680	12	223
Capsicum annuum 7)	76.36	165,894,072	5.41	1,780,527,144	58.11	47497259	1.55	345,248,200	11.29	3,058,000,000	3480	12	290
Capsicum chinense 7)	79.55	197,445,015	6.69	1,649,035,494	55.84	69182970	2.34	433,647,200	14.68	2,954,000,000	3140	12	262

1) Huang et al. (2009);
2) Brassica Genome Consortium (2011);
3) Dohm et al. (2014);
4) Tomato Genome Consortium (2012);
5) this study numbers taken from the repeatmasker analysis used for the assembly (see supplementary Note 1);
6) Sierro et al. (2013);
7) Kim et al. (2013)
8) this number contains 666,441,913 bp (39.13%) LTR-retroelements and 220,727,245 bp (13.08%) non-identified retrotransposons,
9) this number contains 851,543,954 bp (38.32%) LTR-retroelements and 230,653,066 bp (10.33%) non-identified retrotransposons

METHODS

Identification of PVCV-like sequences in the Petunia genomes

Nucleotide sequences of PVCV accession U95208.2 as well as amino acid (aa) sequences of PVCV ORF 1 (accession NP_127504.1) were compared with scaffolds of *PaxiN* and *PinfS6*respectively deposited on the Blast Server: http://petuniasp.sgn.cornell.edu/blast/blast.html using Blast settings without filter and tBlastN respectively. Due to various degrees of sequence degradation distinct thresholds were set as indicated in Tables 1 and 2.

Determination of EPRV diversity

Pol regions of selected *Caulimoviridae* were compared with scaffolds of *PaxiN* and *PinfS6* respectively using the Blast server as above. The thresholds were set at > 200 amino acid (aa) alignment length for all elements and >60% aa identity for *Petu*- and *Florendoviruses*, as well as >45% for all other *Caulimoviridae*. The following accessions and pol regions were incorporated in the search:
NP_569141.1 for PVCV, *Petuvirus*, ORF1, aa 1425..1804,
BAO53400.1 for *Cauliflower mosaic virus* (CaMV), *Caulimovirus*, isolate JPNS2, ORF 5, aa 285..671,
Q89703.1 for *Cassava vein mosaic virus* (CsVMV), *Cavemovirus*, ORF 3, aa 237..637,
YP_007761644.1 for *Rose yellow vein virus* (RYVV), unassigned genus, ORF 3, aa 446..837,
AHN13810.1 for *Aristotelia chilensis vein clearing virus* (ArCVCV), unassigned genus, putative ORF, aa 33..412,
AHA62452.1 for *Banana streak OL virus* (BSOLV), *Badnavirus*, ORF 3 partial, aa 1..413,
NC_001914 for *Rice tungro bacilliform virus* (RTBV), ORF3, aa 1225..1612,
Vitis vinifera B virus (VvinBV_compAsc1), *Florendovirus*, Geering et al. 2014, ORF 1, aa 965..1346,
and *Solanum tuberosum virus* (StubV_scSt1), *Florendovirus*, Geering et al. 2014, ORF 1, aa 1433..1816.
Alignment of sequences has been done using ClustalW within the *MEGA* version 5 software package (Tamura et al., 2011). Identified hits were manually edited to remove overlapping hits and the sequence with the higher score was selected.

*DNA similarity analyse*s

The following PVCV sequences were used for alignment using Clustal W in the *MEGA* version 5 (Tamura et al., 2011) with default settings applied: scaffolds of *P. axillaris* with single insertion of PVCV coding sequences (*PaxiN*_00097, *PaxiN*_00254, *PaxiN*_00447, *PaxiN*_00560, *PaxiN*_00674 and *PaxiN*_00911), scaffolds of *P. inflata* with single insertion of PVCV coding sequences (*PinfS6*_00844, *PinfS6*_01099, *P.inf6S*_01671 and *P.inf6S*_00276) and with double insertions (*P.inf6S*_00235a and *PinfS6*_00235b). Episomal PVCV sequences, isolated from *N. glutinosa* (infectious virus; AAK68664) have been also included in the analysis; LTR-retrotransposon from *Drosophila melanogaster* (Accession AAA70219) has been used as outgroup.

Tandem repeat and retroelement analysis

Basic analysis of the assemblies were performed on Ubuntu Linux 13.10, with Geneious version 7.1.4 (and earlier) by Biomatters (Kearse et al., 2012; available from http://www.geneious.com/). K-mer analyses were performed using Jellyfish version 2.1.3 (Marcais and Kingsford, 2011). Other programmes to search for repeats included LTR-STRUC (McCarthy et al., 2003), LTR finder (http://tlife.fudan.edu.cn/ltr_finder) and RepeatExplorer (Novak et al., 2010; 2013).

Fluorescent in situ *hybridization*

Probe labelling, chromosome preparation and *in situ* hybridization followed the procedure of Schwarzacher and Heslop-Harrison (2000). The 5S rDNA probe was a 410 bp fragment from the clone pTa794 (Gerlach and Dyer, 1980) containing the 5S rDNA repeat unit of *Triticum aestivum*. Three viral probes that, in combination, cover most of the sequence of an infectious chromosomal PVCV copy were produced using the SacI subclone l5-7 from lambda clone 5 (Richert-Pöggeler et al., 2003) while fragment 4-18 (1,233bp long) originates from lambda clone 4 (GenBank AY333912) and contains part of a *Gypsy* gag-pol region. Probes were labelled with biotin-11-dUTP (Roche) or digoxigenin-11-dUTP (Roche) by PCR using M13 forward and reverse sequencing primers for cloned sequences or template specific primers for virus (see Richert-Pöggeler et al., 2003) and 4-18 fragement (forward primer #34935, TGG TAG CGA CTT GTA TCG AGC, reverse primer #34936, TCA ACA AGT AAG CCA CGC AGG, nt 4483-5715 **.**

Root tips from young plants, *P.axillaris* and *P. inflata* (both from the University of Nottingham collection received in 2001) were fixed with 96% ethanol:glacial acetic acid (3:1) after treatment with 0.2 M 8-hydroxyquinoline for 3-4 h. Chromosome preparations were made following proteolytic digestion with cellulase and pectinase, treated with RNase and fixed in 4% paraformaldehyde (see Schwarzacher and Heslop-Harrison, 2000)

The probe mixtures contained 100-200ng labelled probes, 50% (v/v) formamide, 20% (w/v) dextran sulphate, 2x SSC, 0.025µg of salmon sperm DNA and 0.125% SDS (sodium dodecyl sulphate) and 0.125mM EDTA (ethylenediamine-tetraacetic acid). Chromosomes and 40µl of probe were denatured together and allowed to hybridize overnight at 37°C. Post-hybridization washes were at 42°C in 20% formamide and 0.1xSSC, giving a stringency of 80±85%. Detection of hybridization sites was carried out with 4µg/ml Fluorescein-conjugated anti-digoxigenin (Roche) and 2µg/ml Alexa 495-conjugated streptavidin (Molecular Probes). Chromosomes were counterstained with 4µg/ml DAPI (4´,6-diamidino-2-phenylindole) and mounted in CitifluorAF. Slides were analysed with a Zeiss Axioplan2. Fluorescent microscope and images captured with an Optronix S97790 cooled CCD camera. Overlays of hybridization signal and DAPI images were prepared with Adobe Photoshop CS4 using only cropping and functions that treat all pixels equally. For some slides, after photographing and noting down the coordinates of the metaphase, the first probing was washed away during a repeat denaturation step and a second probing was carried out.

ACKNOWLEDGEMENTS

We thank and the Botanic Garden University of Leicester for growing petunia plants for root tip collection and George Heslop-Harrison for help with the retroelement annotation. We are grateful for critical reading and helpful comments by J. Schoelz, University of Missouri, USA. Part of the work was supported by EU-Framework V Paradigm.

AUTHOR CONTRIBUTIONS

KRP analysed the PVCV and EPRV in the genome assemblies; PHH and TS analysed the tandem repeats and retroelements; KRP and TS performed and analysed the FISH experiments; all authors contributed to the MS.

REFERENCES

Brassica rapa genome consortium. (2011). The genome of the mesopolyploid crop species *Brassica rapa.* Nature **43**: 1035-1039. doi: 10.1038/ng.919

Dohm, J.C., Minoche, A.E., Holtgrawe, D., Capella-Gutie´rrez, S., Zakrzewski, F., Tafer, H., Rupp, O., Sorensen, T.R., Stracke, R., Reinhardt, R., Goesmann, A., Kraft, T., Schulz, B., Stadler, B.F., Schmidt, T., Gabaldo, T., Lehrach, H., Weisshaar, B. and Himmelbauer, H. (2014). The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature **505**: 546-549. doi: 10.1038/nature12817.

Engelmann, J. and Hamacher, J. (2008). Plant Virus Diseases: Ornamental Plants. In: Mahy, B. W. and van Regenmortel, M.H. V. (Eds.) Desk Encyclopedia of Plant and Fungal Virology Vol. 4, Acad. Press. Elsevier pp. 436-458.

Geering, A.D.W., Maumus, F., Copetti, D., Choisne, N., Zwickl, D.J., Zytnicki, M., McTaggart, A.R., Scalabrin, S., Vezzulli, S., Wing, R.A., Quesneville, H., Teycheney, P.-Y. (2014). Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. Nature Communications 5:5269. doi: 10.1038/ncommms6269.

Gerats, T. (2009). Identification and exploitation of Petunia transposable elements: a brief history. In Petunia: Evolutionary, Developmental and Physiological Genetics, 2nd edition. Eds. Gerats, T. and Strommer, J. Springer, New York. doi: 10.1007/978-0-387-84796-2.

Gregor, W., Mette, M.F., Staginnus, C., Matzke, M.A., Matzke, A.J. (2004). A distinct endogenous pararetrovirus family in Nicotiana tomentosiformis, a diploid progenitor of polyploid tobacco. Plant Physiol. 13: 1191-**119**9.

Hansen, C.N., Harper, G. and Heslop-Harrison, J.S. (2005). Characterization of pararetrovirus-like sequences in the genome of potato (*Solanum tuberosum*). Cytogenetic and Genome Research **10**: 559-565. doi: 10.1159/000084989.

Hansen, C.N. and Heslop-Harrison, J.S. (2004). Sequences and phylogenies of plant pararetroviruses, viruses and transposable elements. Adv. Bot. Res. **41**: 165-193.

Heslop-Harrison, J.S. and Schwarzacher, T. (2013). Nucleosomes and centromeric DNA packaging. Proc. Nat. Acad. Sci. USA.http://www.pnas.org/content/110/50/19974.full.pdf+html http://dx.doi.org/10.1073/pnas.1319945110 -

Hohn, T., Richert-Pöggeler, K.R., Staginnus, C., Harper, G., Schwarzacher, T., Teo, C.H., Teycheney, P.-Y., Iskra-Caruana, M.-L. and Hull, R. (2008). Evolution of integrated plant viruses. In: Roossinck, M.J. (ed.) Plant Virus Evolution. Springer Berlin Heidelberg pp. 53-81.

Huang, S. et al. (2009). The genome of the cucumber, *Cucumis sativus* L. Nature Genetics **41**: 1275-1281. doi: 10.1038/ng.475.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences **8**: 275-282.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. and Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647–1649. doi: 10.1093/bioinformatics/b199.

Kim et al. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nature Genetics **46**: 270-278. doi: 10.1038/ng.2877.

King, A.M.Q., Adams, M.J., Carstens, E.B. and Lefkowitz, E.J. (2012). Virus taxonomy. Ninth Report of the International Committee on Taxonomy of Viruses, Elsevier Academic Press, San Diego.

Kriedt, R.A., Cruz, G.M.Q., Bonatto, S.A. and Freitas, L.B. (2014). Novel transposable elements in Solanaceae: Evolutionary relationships among Tnt1-related sequences in wild petunia species. Plant Mol. Biol. Rep. **32**: 142-152. doi: 10.1007/s11105-013-0626-8.

Marcais, G. And Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics **27**: 764-770. doi: 10.1093/bioinformatics/btr011.

Matsubara, K., Kodama, H., Kokubun, H., Watanabe, H. and Toshio, A. (2005). Two novel transposable elements in a cytochrome P450 gene govern anthocyanin biosynthesis of commercial petunias. Gene **358**: 121-126.

McCarthy, E.M. and McDonald, J.F. (2003). LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics **19**: 362–367.

Mishiba, K.I., Ando, T., Mii, M., Watanabe, H., Kokubun, H., Hashimoto, G. and Marchesi, E. (2000). Nuclear DNA content as an index character discriminating taxa in the genus etunia sensu Jussieu (Solanaceae). Annals of Botany **85**: 665-673. doi: 10.1006/anbo.2000.112.

Mushegian, A.R. and Elena, S.F. (2015). Evolution of plant virus movement proteins from the 30K superfamily and their homolgs integrated in plant genomes. Virology **476**: 304-315.

Noreen, F., Akbergenov, R., Hohn, T., Richert-Pöggeler, K.R. (2007). Distinct expression of endogenous *Petunia vein clearing virus* and the DNA transposon *dTph1* in two *Petunia hybrida* lines is correlated with differences in histone modification and siRNA production. Plant J. **50**: 219-229.

Novak, P., Neumann, P. and Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics **11**: 378. doi:10.1186/1471-2105-11-378.

Novak, P., Neumann, P., Pech, J., Steinhaisl, J. and Macas, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. Bioinformatics **29**: 792-793.

Richert-Pöggeler, K.R., Noreen, F., Schwarzacher, T., Harper, G. and Hohn, T. (2003). Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. EMBO J. **22**: 4836–4845.

Richert-Pöggeler, K.R. and Schwarzacher, T. (2009). Impact of etroelements in shaping the petunia genome in petunia: Evolutionary, Developmental and Physiological Genetics, 2nd edition. Eds. Gerats, T. and Strommer, J. Springer, New York. doi: 10.1007/978-0-387-84796-2.

Richert-Pöggeler KR and Shepherd, R.J. (1997). Petunia vein clearing virus: a plant pararetrovirus with the core sequences for an integrase function. Virology **236**: 137-146.

Roy, A., Shao, J., Schneider, W.L., Hartung, J.S. and Brlansky, R.H. (2014). Population of endogenous pararetrovirus genomes in Carrizo citrange. Genome Announc. **2**: e01063-13. doi:10.1128/genomeA.01063-13.

Saerkinen, T., Bohs, L., Olmstead, R.G. & Knapp, S. (2013) A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. BMC Evolutionary Biology **13:** 214.

Schmidt, A., Schmid, M.W. and Grossniklaus, U. (2012). Analysis of plant germline development by high-throughput RNA profiling: technical advances and new insights. Plant J. **70**: 18-29.

Schwarzacher, T. and Heslop-Harrison, J.S. (2000). Practical in situ hybridization. BIOS (Oxford).

Sierro, N., Battey, J.N.D., Ouadi, S., Bovet, L., Goepfert, S., Bakaher, N., Peitsch, M.C. and Ivanov, N.V. (2013). Reference genomes and transcriptomes of Nicotiana sylvestris and Nicotiana tomentosiformis. Genome Biology **14**: R60.

Staginnus, C. and Richert-Pöggeler, K.R. (2006). Endogenous pararetroviruses: Two-faced travelers in the plant genome. Trends Plant Sci. **11**: 485–491.

Sveinsson, S., Gill, N., Kane, N.C. and Cronk, Q. (2013). Transposon fingerprinting using low coverage whole genome shotgun sequencing in Cacao (Theobroma cacao L.) and related species. BMC Genomics **14**: 502.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution **28**: 2731-2739. doi:10.1093/molbev/msr121.

Teycheney, P.-Y. and Geering, A. (2011). Endogenous viral sequences in plant genomes. In: Caranta, C., Aranda, M.A., Tepfer, M. and Lopez-Moya, J.J. (eds.) Recent Advances in Plant Virology. Caister Academic Press, Norfolk, UK pp. 343-362.

Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**: 635-641. doi: 10.1038/nature11119.

Yang, Z.N., Ye, X.R., Molina, J., Roose, M.L. and Mirkov, T.E. (2003). Sequence analysis of a 282-kilobase region surrounding the citrus tristeza virus resistance gene (*Ctv*) locus in *Poncirus trifoliata* L. Raf. Plant Physiol. **131**: 482–492.

Young, G.R., Stoye, J.P., Kassiotis, G. (2013). Are human endogenous retroviruses pathogenic? An approach to testing the hypothesis. Bioessays **35**: 794–803.