the plant journal

SEB Society for Experimental Biology

The Plant Journal (2014) **80,** 823–833 doi: 10.1111/tpj.12673

Endogenous pararetrovirus sequences associated with 24 nt small RNAs at the centromeres of *Fritillaria imperialis* L. (Liliaceae), a species with a giant genome

Hannes Becher[1], Lu Ma[1], Laura J. Kelly[1,2], Ales Kovarik[3], Ilia J. Leitch[2] and Andrew R. Leitch[1,*]

[1]*School of Biological and Chemical Sciences, Queen Mary University of London, London E1 4NS, UK,*

[2]*Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AB, UK, and*

[3]*Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno CZ–61265, Czech Republic*

Received 20 June 2014; revised 10 September 2014; accepted 11 September 2014; published online 17 September 2014.

*For correspondence (e-mail a.r.leitch@qmul.ac.uk).

SUMMARY

Endogenous pararetroviral sequences are the most commonly found virus sequences integrated into angiosperm genomes. We describe an endogenous pararetrovirus (EPRV) repeat in *Fritillaria imperialis*, a species that is under study as a result of its exceptionally large genome (1C = 42 096 Mbp, approximately 240 times bigger than *Arabidopsis thaliana*). The repeat (FriEPRV) was identified from Illumina reads using the RepeatExplorer pipeline, and exists in a complex genomic organization at the centromere of most, or all, chromosomes. The repeat was reconstructed into three consensus sequences that formed three interconnected loops, one of which carries sequence motifs expected of an EPRV (including the *gag* and *pol* domains). FriEPRV shows sequence similarity to members of the *Caulimoviridae* pararetrovirus family, with phylogenetic analysis indicating a close relationship to *Petuvirus*. It is possible that no complete EPRV sequence exists, although our data suggest an abundance that exceeds the genome size of Arabidopsis. Analysis of single nucleotide polymorphisms revealed elevated levels of C→T and G→A transitions, consistent with deamination of methylated cytosine. Bisulphite sequencing revealed high levels of methylation at CG and CHG motifs (up to 100%), and 15–20% methylation, on average, at CHH motifs. FriEPRV's centromeric location may suggest targeted insertion, perhaps associated with meiotic drive. We observed an abundance of 24 nt small RNAs that specifically target FriEPRV, potentially providing a signature of RNA-dependent DNA methylation. Such signatures of epigenetic regulation suggest that the huge genome of *F. imperialis* has not arisen as a consequence of a catastrophic breakdown in the regulation of repeat amplification.

Keywords: pararetrovirus, *Fritillaria imperialis*, centromere, RNA interference, cytosine methylation, RdDM, giant genome.

INTRODUCTION

Viral sequences from two virus families, the *Caulimoviridae* (pararetroviruses) and the *Geminiviridae*, are known to have integrated into the nuclear genomes of angiosperms. These endogenous viral sequences are typically present in a few hundred or thousand copies, contributing to the repetitive genome content. They show varying levels of degradation and rearrangements, and, in most cases, are functionally defective. Nevertheless, there are a few examples of endogenous pararetroviral sequences (EPRVs) belonging to *Caulimoviridae* for which complete copies exist, or for which recombination may generate complete copies. Under certain circumstances (e.g. stress, hybridization), these EPRVs may become circularized and released from the genome and become infective (Teycheney and Geering, 2011; Chabannes and Iskra-Caruana, 2013).

Repetitive DNA amplification processes and the efficiency of repeat removal are generally accepted to account for genome size differences in angiosperms of the same ploidy, for example in cotton (*Gossypium hirsutum*; Hawkins *et al.*, 2009), *Arabidopsis lyrata* (Hu *et al.*, 2011) and *Nicotiana tabacum* (Renny-Byfield *et al.*, 2011). In recent years, it has become increasingly clear that much of the repetitive DNA in the genome is epigenetically regulated via pathways that are mediated by small interfering RNAs (siRNAs), leading to DNA methylation, histone modification and chromatin remodelling (Simon and Meyers, 2011; Kim and Zilberman, 2014; Matzke and Mosher, 2014). Our

understanding of these pathways is primarily derived from studies on species with small genome sizes (e.g. *Arabidopsis thaliana*, 1C = 157 Mbp) (Bennett *et al.*, 2003), in which epigenetic regulation silences transposable elements and thus hampers their proliferation. However, angiosperms have an exceptionally large (2400-fold) range in genome size (Leitch and Leitch, 2012, 2013), and the question arises as to whether epigenetic pathways also operate in species with such large genomes, as failure of epigenetic regulation may potentially lead to genome enlargement (Kelly and Leitch, 2011).

With the advent of high-throughput next-generation sequencing approaches, species with larger genomes have recently become amenable to study, and are now the focus of research into the origin and evolution of genome obesity (Kelly and Leitch, 2011; Kelly *et al.*, 2012; Nystedt *et al.*, 2013). Even using low-coverage sequencing data, it is possible to reliably analyse the repetitive content of large genomes as shown for pea (*Pisum sativum*) by Macas *et al.* (2007). Here we use low-coverage genomic DNA sequence data from Illumina's next-generation sequencing platform to characterize an EPRV in *Fritillaria imperialis* (FriEPRV), a species with an exceptionally large genome (1C = 42 096 Mbp or more, if B chromosomes are present) (Leitch *et al.*, 2007). In addition, we also characterize the small RNA (smRNA) and cytosine methylation status of the *F. imperialis* genome, to obtain insights into the regulation of FriEPRV and its evolution.

RESULTS

Characterization of a pararetrovirus-like sequence in *F. imperialis*

We used RepeatExplorer to characterize the repetitive DNA content of the genome of *F. imperialis* (see Experimental procedures, Novák *et al.*, 2010, 2013). The EPRVs identified, hereafter called FriEPRV, were in the 15th largest repeat cluster in terms of the number of reads it contains. The cluster comprises 7727 reads (0.386%) of the 2 000 842 analysed, the latter representing a genome coverage of 0.475% in *F. imperialis*. The 7727 reads were assembled into 109 contigs, 14 of which comprised more than 30 reads (Table 1). These 14 contigs are AT-rich and in total contain 7119 reads, corresponding to 92.1% of all reads in the cluster. All contigs of FriEPRV are contained in Supporting Data file S1. The cluster graph of FriEPRV consists of three distinct loops (loops 1–3, Figure 1a), each of which comprises a set of overlapping or similar sequence reads. The region at which the loops connect occurs because there are similar sequences in the three loops. In addition, multiple Illumina read pairs (insert size 300–500 bp) span the loops, providing evidence that the loops are physically connected (Figure 1b).

Table 1 Characterization of contigs and reads obtained in FriEPRV: contigs in bold were not included in the loop consensus sequences

Consensus sequence (length in bp)	Contig name	Number of reads	Contig length in bp	Read depth	GC content in %
Loop 1 (5939)	8	90	250	36.00	42.1
	18	2035	2072	98.21	34.9
	29	1110	1174	94.55	37.6
	39	269	424	63.44	34.0
	80	1690	1948	86.76	35.7
	90	**63**	**253**	**24.90**	**34.4**
	92	79	210	37.62	38.1
	94	629	877	71.72	40.6
Loop 2 (1322)	7	44	286	15.38	43.4
	38	311	969	32.09	41.7
	78	195	647	30.14	41.4
Loop 3 (1216)	11	203	543	37.38	46.2
	102	**34**	**248**	**13.71**	**44.4**
	105	367	932	39.38	42.6

The 14 contigs reconstructed in FriEPRV (Table 1) were further assembled using Geneious (http://geneious.com/) to generate consensus sequences for loops 1–3 (Figure 2 and Table 1). Read depth analysis across each consensus sequence revealed a read depth of approximately 70–130-fold for loop 1, approximately 20–40-fold for loop 2 and approximately 40–60-fold for loop 3 (Figure 2). Thus loop 1 of FriEPRV has approximately twice the read depth of the other loops. Dot-plot analyses of the consensus sequences revealed a minisatellite at the 5′ and 3′ ends of each consensus (Figure S1), probably causing RepeatExplorer to link reads into three loops (Figure 1). Before downstream analysis, this satellite was removed from all consensus sequences except the 5′ end of loop 1. Use of Tandem Repeats Finder (Benson, 1999) showed that the minisatellite has a 30 bp consensus motif (AAGGGGG TTTTGATGCTCTAATACCACTCG)$_n$.

For the consensus sequences of loops 2 and 3, BLASTn and BLASTx analysis revealed no significant matches against National Center for Biotechnology Information databases (using an *e*–value threshold of $<10^{-5}$). In contrast, BLASTx analysis of the consensus sequence of loop 1 revealed a match against a predicted polyprotein domain in *Cicer arietinum*. There were also matches against a *Citrus* endogenous pararetrovirus, a hypothetical protein in *Eutrema salsugineum*, petunia vein clearing virus (PVCV) and a predicted protein from *Populus trichocarpa*. BLASTx also detected conserved domains characteristic of retrovirus-like sequences. All domains identified are illustrated in Figure 2. The consensus sequence of loop 1 also contains a large open reading frame (ORF) of 5361 bp, which is terminated at the 3′ end by a double stop codon (TAATAA). FriEPRV's gen-

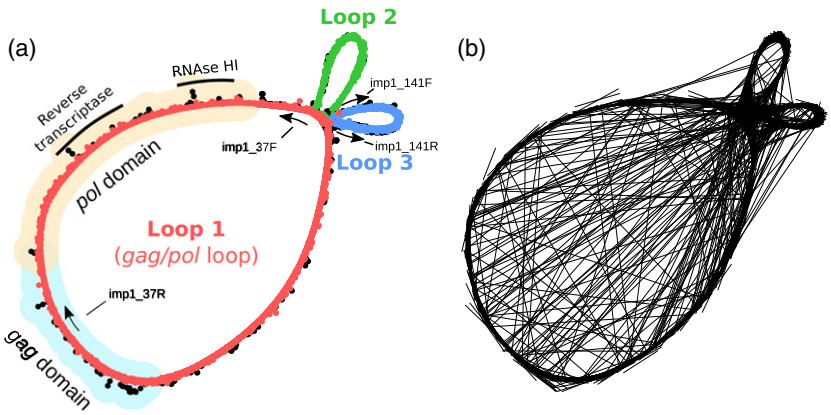

Figure 1. Graphical 2D projection of the structure of FriEPRV generated using RepeatExplorer (Novák *et al.*, 2010).
(a) Each node (dot) represents one of the 7727 Illumina sequence reads. The placement of the nodes reflects sequence similarity, with sequences that are most alike being placed closest together. The sequences have been shaded according to their presence in one of three loops: large loop 1 (red) and minor loops 2 and 3 (blue and green). Black dots represent reads that are not incorporated in the contigs mentioned in Table 1. The protein-coding domains were annotated in RepeatExplorer with reference to RepeatMasker and Repbase (Jurka *et al.*, 2005) and the conserved domain database (Marchler-Bauer *et al.*, 2011), and by MEGABLAST and BLASTx analysis. The labelled arrows indicate the position of primers used for PCR (see Table S1 for details).
(b) Illumina read pairs are connected by lines, indicating physical connection of the loops.

ome proportion of 0.386% equates to 162 Mbp (see Appendix S1 for details of the confidence interval of this estimate). As repeat clustering of 0.475% of the genome of *F. imperialis* results in an approximately 100-fold coverage for loop 1 (*gag*/*pol* loop, 5939 bp), we assume the presence of approximately 21 000 copy equivalents of loop 1 of FriEPRV. Appendix S2 provides a more detailed explanation of how the copy number estimates were derived.

Previously, Llorens *et al.* (2009) analysed the phylogenetic relationships between members of the pararetrovirus family *Caulimoviridae*. Given the sequence similarities that we found, we aligned the consensus sequence of FriEPRV loop 1 to representative *Caulimoviridae* sequences from Llorens *et al.* (2009). Phylogenetic analysis of amino acid sequences from *Caulimoviridae* revealed a strongly supported relationship between FriEPRV and PVCV (Figure S2). Dot-plot analysis of amino acid sequences of FriEPRV loop 1 against PVCV also revealed high levels of conservation in regions with sequence similarity to the *gag* and *pol* polyprotein genes, which also occur in LTR retrotransposons and related elements, as in other members of the *Caulimoviridae* (Figure 3). *pol* usually encodes protease, reverse transcriptase and ribonuclease. In transposable elements, it also encodes integrase (Llorens *et al.*, 2009). Further sequence similarities were found between the N–terminal regions of PVCV and FriEPRV (movement protein, Figure 3), although the DNA sequences diverge between the *gag*/*pol* and N–terminal regions (corresponding to amino acids 300–800 in Figure 3), even when considering potential frame shifts.

Small RNAs

The smRNA reads were mapped to the consensus sequences of the three FriEPRV loops (Figure 2, black bars). Of 4609 mapped smRNA reads, the majority (67%) were found to belong to the 24 nt size category (Figure S3). Peaks of high smRNA read abundance (>100 reads) were found at particular positions along the three consensus sequences. Six peaks were observed on loop 1, one of which was in the *pol* domain. One of the highest peaks corresponds to the 30 bp minisatellite (Figure 2, asterisk). None of the smRNA reads that mapped to FriEPRV mapped to any other repeat cluster identified by RepeatExplorer in the full *F. imperialis* dataset of 2 000 842 reads.

Cytosine methylation

After mapping Illumina reads from sodium bisulfite-treated genomic DNA to each of the three consensus sequences of FriEPRV, we calculated the proportion of cytosines at each location that were methylated in the genomic DNA (see Experimental procedures). We observed that, in the context of CG and CHG (where H represents A, T or C), most cytosine residues were methylated, with medians of approximately 86 and 90% of cytosines at any particular position (Figure 4a). In contrast, cytosines in the context of CHH were significantly less methylated, with methylation being found at only approximately 11% of cytosines at any particular position in the consensus sequences. Nevertheless, some cytosines in the CHH context were methylated in most mapped reads (up to 88%) (see Figure 4a). We compared the distribution of methylated cytosines in the context of CHH (1641 sites) against their map position in

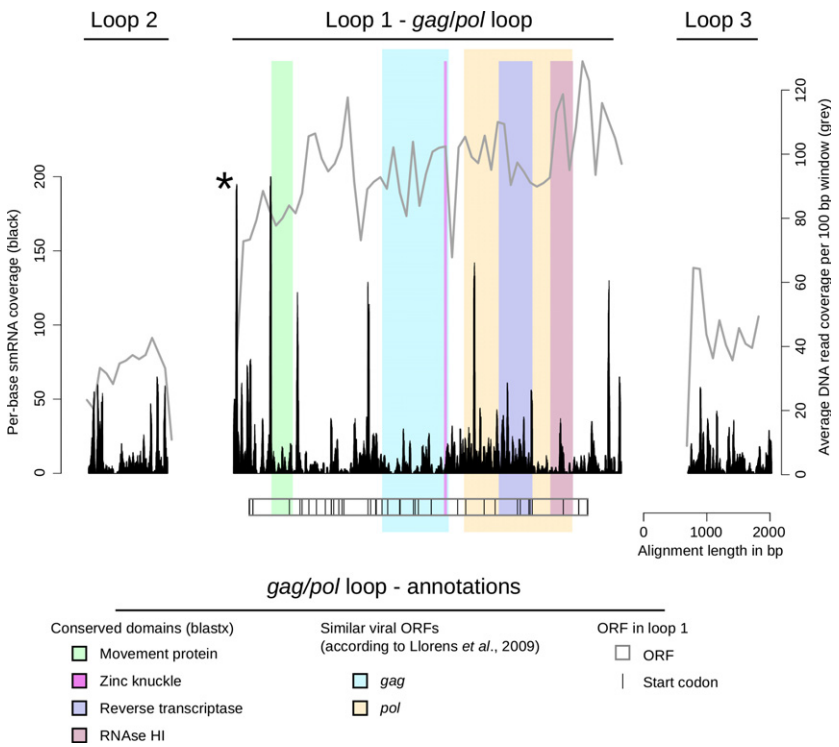

Figure 2. Read depth analysis (per base pair) using Illumina reads of genomic DNA (grey line) and small RNAs (black bars) against the consensus sequences for loops 1–3.
The box under loop 1 indicates the largest open reading frame (ORF); vertical bars represent in-frame start codons. Protein domains were annotated as described by Llorens *et al.* (2009) and BLASTx results. Note that the read depth from genomic DNA for loops 2 and 3 is approximately half that observed in loop 1. Note the multiple peaks of small RNA reads, one of which corresponds to a 30 bp mini-satellite (indicated by an asterisk).

FriEPRV sequences (210 bp windows). Forty cytosines in the context of CHH were predominantly methylated (>70%), these cytosines were scattered apparently at random across loops 1–3. Loop 1 showed comparatively little methylation in the CHH context at its 5′ end (Figure 4b).

An analysis of single nucleotide polymorphisms (SNPs) revealed that C→T and G→A transitions were most abundant (Figure S4), a pattern consistent with deamination of methylated cytosines.

Fluorescence *in situ* and Southern hybridization

PCR primers were designed to amplify the region containing the *pol* domain of loop 1 and the majority of loop 3 (Figure 1a and Table S1). Gel electrophoresis revealed a range of product sizes in each case, probably reflecting heterogeneity amongst natural FriEPRV sequences (shown for loop 1 in Figure 5a). Thus the loop 1 and 3 probes used for fluorescent *in situ* hybridization (FISH) each contained a range of size fragments.

The karyotype of *F. imperialis* comprises 2 $n = 24$ chromosomes, 20 of which are acrocentric, two of which are sub-metacentric and two of which are metacentric (see asterisks in Figure 6a). The *F. imperialis* individual depicted in Figure 6 also contained one additional B chromosome in each metaphase (see arrow in Figure 6a). The Texas Red-labelled loop 1 probe (carrying the *gag*/*pol* domain) hybridized strongly to the centromeric region of the 20 acrocentric chromosomes and the two sub-metacentric chromosomes. In addition, weak labelling was sometimes observed at the centromere of one metacentric chromosome (but is not visible in Figure 6a), although its homologue always appeared unlabelled, even when the image was contrast-enhanced (result not shown). The B chromosome showed a very small and weak hybridization signal in the telomeric region.

The Texas Red-labelled loop 1 probe (red signal) and the Alexa Fluor 488-labelled loop 3 probe (green signal) co-localized at metaphase, giving rise to a yellow hybridization signal (results not shown), a pattern that was also observed for the majority of signals at interphase (Figure 6b). At interphase, some FriEPRV sequences were more decondensed than others, and, some had stronger loop 1 than loop 3 signals, resulting in red rather than yellow florescence (see arrows in Figure 6b).

Southern hybridization, using the probe for loop 1 on *Bst*NI-restricted genomic DNA from a range of *Fritillaria* species (selected to reflect the phylogenetic diversity of the genus) showed only a weak hybridization signal in the lane for *F. imperialis* (Figure 5b). The long exposure times required to reveal the signal reflect the relatively low genome proportion (approximately 0.4%) of FriEPRV. There was one prominent hybridization band at approximately 3 kb, and a number of indistinct minor bands, potentially reflecting multiple copies of similar sequence. None of the other species (listed in Table S2) showed any signal even though the same membrane probed for 18S rDNA revealed signal in all lanes (Figure 5c).

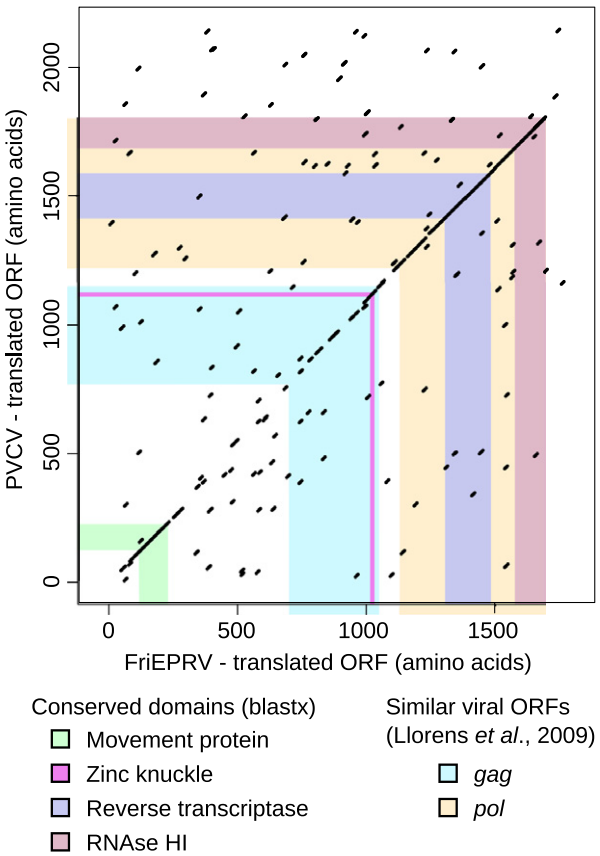

Figure 3. Dot-plot analysis of the translated consensus sequence of FriEPRV loop 1 and petunia vein clearing virus (PVCV).
Note the similarity in the protein-coding domains, particularly in the *pol* domain. Domains are annotated as shown in Figure 2. Note the less conserved region between amino acids 300 and 800.

DISCUSSION

FriEPRV: an endogenous pararetroviral sequence

Our data are indicative of the relatively recent insertion and amplification of an endogenous pararetroviral sequence in *Fritillaria* that we have called FriEPRV. Evidence for this includes the findings that (i) a Southern hybridization signal was detected only in *F. imperialis*, (ii) there is conservation of coding sequences in a dot-plot analysis with PVCV, and (iii) there is an absence of stop codons in the ORF of FriEPRV. The consensus sequence of FriEPRV loop 1 reveals domains with similarity to the *gag* and *pol* domains of retroviruses and LTR retrotransposons (Figure 2). BLAST analysis, sequence reconstruction and phylogenetic analysis all indicate that FriEPRV loop 1 is an EPRV sequence, which, in our analysis (Figure S2), is most closely related to PVCV, which belongs to the genus *Petuvirus* within the *Caulimoviridae*. Like PVCV, FriEPRV exhibits one large polyprotein ORF containing *gag* and *pol* domains (Noreen *et al.*, 2007). Dot-plot analysis comparing the ORF of FriEPRV loop 1 with that of PVCV (Figure 3) revealed considerable sequence similarity at the amino acid level, with the exception of the region between amino acids 300 and 800. In LTR retrotransposons, which are closely related to pararetroviruses, this region contains an integrase domain that is not functional in pararetroviruses (Chabannes *et al.*, 2013).

Plant pararetroviruses are the most abundant viral sequences integrated into plant genomes, and have previously been reported in species belonging to Asteraceae, Asparagaceae, Bromeliaceae, Musaceae, Poaceae, Rosaceae, Rutaceae, Salicaceae, Solanaceae and Vitaceae

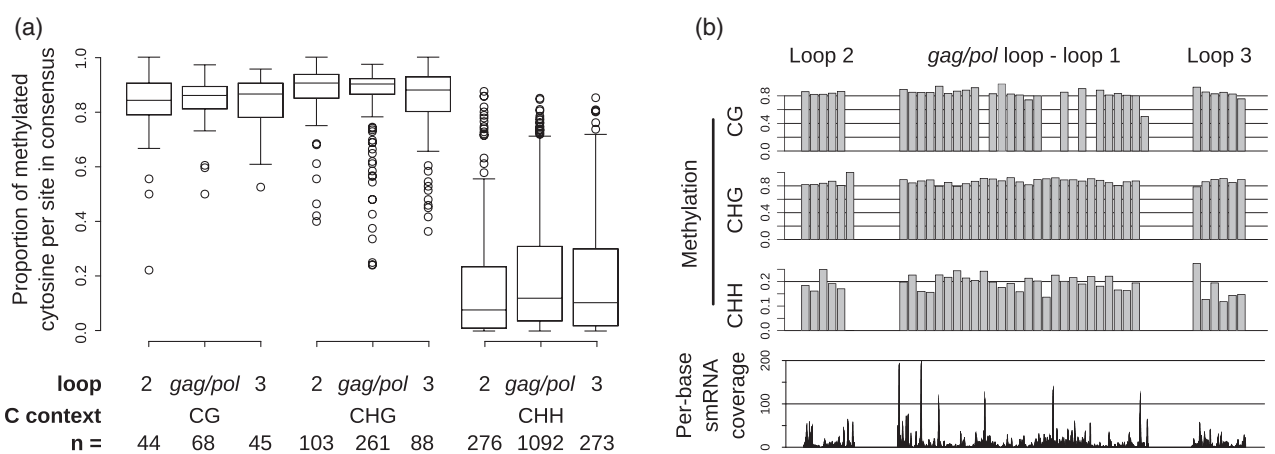

Figure 4. Analysis of the frequency and distribution of methylated cytosines in the consensus sequences of loops 1–3 of FriEPRV.
(a) The numbers (*n*) of CG, CHG and CHH motifs in the consensus sequences of loops 1–3 are shown at the bottom. For each cytosine in the consensus, the proportion of reads that are modified (and hence unmethylated) or unaltered (and hence methylated) by sodium bisulfite treatment was recorded and is indicated as the methylation proportion at each cytosine. Only cytosines with more than ninefold bisulfite read coverage were considered. The plot shows the range of methylation ratios in the context of CG, CHG and CHH motifs. For each box plot, the lower, middle and upper horizontal lines represent the first, second and third quartiles, respectively. The maximum whisker length is 1.5 times the interquartile range. Results for cytosines outside this range are indicated as circles.
(b) The mean level of methylation along the sequences of loops 1–3 is shown in windows of 210 bp for all cytosine contexts (CG, CHG and CHH). In loop 1, there are some regions without cytosine in CG context, hence the missing bars. Per-base coverage of smRNA mapped to the loop consensus sequences is shown at the bottom for comparison.

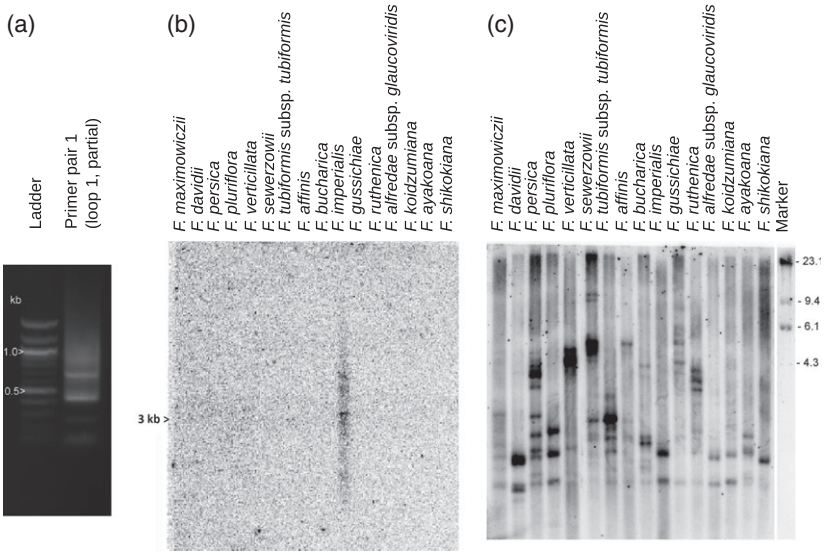

Figure 5. Occurrence of FriEPRV.
(a) Gel electrophoresis analysis of PCR products using primer pair 1 (see Table S1) against loop 1 of FriEPRV for *Fritillaria imperialis*. Note the smear indicating a range of sizes of products.
(b) The products in (a) were used as a probe for Southern hybridization of *Bst*NI-digested genomic DNA from a range of *Fritillaria* species. Only *F. imperialis* shows three hybridization bands in addition to a smear.
(c) Re-hybridization of the same membrane using the 18S probe. In each species, the probe hybridized to bands of variable size, reflecting length polymorphisms in the intergenic spacer.

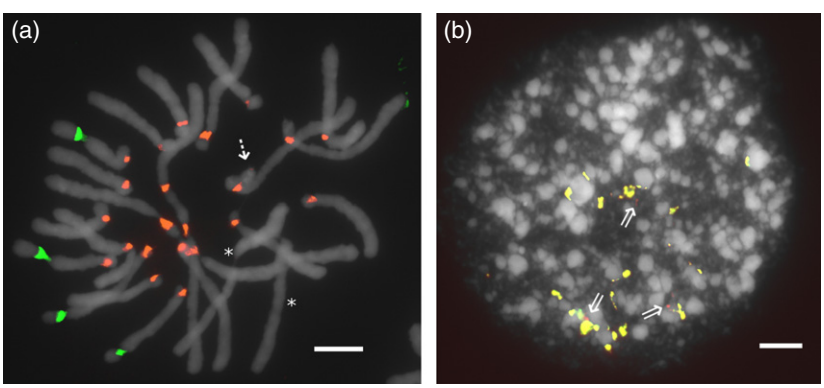

Figure 6. FISH analysis in root tip cells of *F. imperialis*. DNA was stained with 4′,6–diamidino-2–phenylindole (white) to detect the EPRV sequence FriEPRV.
(a) Metaphase chromosomes probed with 18S rDNA (green fluorescence) and loop 1 probe (red fluorescence). Note the centromeric location of the FriEPRV sequences, as indicated by the occurrence of a constriction in many of the signals, on acrocentric and sub-metacentric chromosomes. There are small signals on the B chromosome (arrow), and an absence of signal on the two metacentric chromosomes (asterisks).
(b) Interphase nucleus probed with loop 3 probe (green fluorescence) and loop 1 probe (red). Note that, in most cases, the signals are co-localized (giving a yellow fluorescence signal), although there are a few signals that carry only loop 1 label (arrows). Note also the variable condensation state of FriEPRV at interphase. Scale bars = 10 μm.

(Teycheney and Geering, 2011; Chabannes and Iskra-Caruana, 2013). We are unaware of other reports of an endogenous pararetrovirus in a species belonging to Liliaceae, or indeed in the whole of the monocot order Liliales comprising approximately 1300 species.

Organization of the pararetrovirus sequence FriEPRV

Functional EPRVs may occur as complete elements (which occur in tandem arrays in *Petunia hybrida*) or as incomplete, fragmented and rearranged sequences either at a single locus, as in *Musa balbisiana*, or at several genomic loci, as in *Nicotiana* × *edwardsonii* (Chabannes and Iskra-Caruana, 2013). Despite such differences in genome organization, inter-specific hybridization, stress or wounding may result in release of full-length circularized infectious viral genomes, potentially involving complex patterns of recombination (Chabannes and Iskra-Caruana, 2013). In the present study, RepeatExplorer reconstructed FriEPRV as a complete viral genome (loop 1). Although our analysis utilizes short reads and covers only a fraction of the *F. imperialis* genome, there is evidence indicating the genome organization of FriEPRV: (i) loop 1 is approximately twice as long as the two other loops, (ii) all three loops share similar terminal sequences, and (iii) the loops are physically connected as evidenced by the presence of Illumina read pairs with mates in two loops (see Figure 1b). These results agree with a fundamental organization of the type $(\text{loop 1, 2, 1, 3})_n$. Potentially loops 2 and 3 provide spacer sequences between FriEPRV copies, or contain promoter domains for their transcription. At present, we do not know

whether a complete copy of FriEPRV exists in the genome. Nevertheless, the large range of fragment sizes generated by PCR using primer pair 1 to amplify loop 1 FriEPRV sequences (Figure 5a), and the range of restriction fragment sizes that hybridized with the loop 1 probe following Southern hybridization (Figure 5b), suggest either a complex organization of FriEPRV in the *F. imperialis* genome, including truncated variants, and/or much sequence divergence between the repeats. Nevertheless, given that it takes only one complete EPRV element, or recombination of fragmented elements, to generate active viral genomes (Chabannes and Iskra-Caruana, 2013; Chabannes *et al.*, 2013), it is possible that the FriEPRV in *F. imperialis* may have the potential to be infectious, even if all individual copies are rearranged.

We predict that the equivalent of approximately 21 000 copies of FriEPRV are present in the genome of *F. imperialis* (see Appendix S2). While abundant ERPVs have been reported for *Nicotiana tabacum* (Jakowitsch *et al.*, 1999), *Solanum lycopersicum* (Staginnus *et al.*, 2007) and *Capsicum annuum* (Kim *et al.*, 2014), *F. imperialis* exhibits the highest copy number reported for an ERPV to date. Together, FriEPRV sequences are estimated to account for approximately 162 Mbp of the *F. imperialis* genome, which is comparable to the genome size of *Arabidopsis thaliana* (Bennett *et al.*, 2003). However in the context of *F. imperialis*, this copy number represents a genome proportion of only approximately 0.4%.

Chromosomal distribution of FriEPRV

Fluorescent *in situ* hybridization (FISH) revealed that sequences corresponding to loops 1 and 3 of FriEPRV were predominantly located at the morphological centromeres of the A chromosomes. Additional but very weak signals were also seen on B chromosomes (Figure 6a). Although EPRV sequences at centromeric/pericentromeric sites have been reported previously in *Solanum lycopersicum* (tomato), *Solanum habrochaites* (Staginnus *et al.*, 2007) and *Solanum tuberosum* (potato, Hansen *et al.*, 2005), they are not restricted to such regions in these species. The only other example in which EPRV sequences were found to be restricted to the centromeres was for PVCV in *Petunia hybrida* (Richert-Pöggeler *et al.*, 2003), and it is perhaps notable that PVCV is phylogenetically the closest known EPRV sequence to FriEPRV (Figure S2).

What explains FriEPRV's localization at the centromeres? It is unlikely there is selection against widespread integration of the sequence, except perhaps in genic domains, given the enormous genome size of *F. imperialis* (1C = 42 096 Mbp). Previously, Richert-Pöggeler and Shepherd (1997) annotated the region between the movement protein and the RNA-binding domain in PVCV as integrase. This region has diverged between FriEPRV and PVCV (Figure 3), and other pararetroviruses neither possess functional integrases nor do they require integration for replication (Chabannes *et al.*, 2013). However, the centromeric location of both PVCV and FriEPRV reveals accumulation of the repeats at the centromeres, perhaps even targeted integration, as reported for other transposable elements (Neumann *et al.*, 2011).

Malik and Henikoff (2009) suggested a model for centromeric repeat proliferation involving meiotic drive: the non-Mendelian segregation of chromosomes bearing certain sequences during meiosis leading to their preferential inheritance (Kanizay and Dawe, 2009; Kanizay *et al.*, 2013). If the presence of FriEPRV sequences at centromeres causes (or at some point during its evolution has caused) meiotic drive, it is likely that FriEPRV will spread (or have spread), as chromosomes with more centromeric FriEPRV sequences will be preferentially inherited. This has been reported for heterochromatin in maize (*Zea mays*) and centromeric repeats in other grasses (Kanizay and Dawe, 2009; Kanizay *et al.*, 2013). It also agrees with the theory of 'boom–burst' cycles in the evolution of centromere sequences (Zhang *et al.*, 2014).

Epigenetic silencing in the context of a giant genome

Our knowledge about epigenetic silencing and its implications for genome composition is mainly based on the analysis of species with small genomes. For example, in Arabidopsis, epigenetic silencing of repetitive DNA is regulated by two pathways: (i) the self-reinforcing maintenance methylation pathway involving the KRYPTONITE (KYP) family of histone methyl transferases and chromomethylases, and (ii) the sRNA-directed DNA methylation pathway. KRYPTONITE/chromomethylase maintenance methylation pathways are chiefly active in the centres of long transposable elements (TEs) with compact chromatin structures, in which KRYPTONITE proteins dimethylate lysine 9 of histone 3 (H3K9me2) in the presence of DNA methylation at CHG and CHH motifs. Chromomethylase proteins in turn recognize H3K9me2 sites and methylate DNA, keeping the central parts of TEs silenced independent of their actual sequence. In contrast, the sequences at the edges of TEs are mainly silenced by sRNA-directed DNA methylation, which is sequence-specific and does not depend on H3K9me2 and chromatin compaction. Here, sequence specificity matters, because sequences neighbouring silenced TEs may need to be transcriptionally active and hence retain a more open euchromatin conformation (Kim and Zilberman, 2014).

For TEs surrounded by euchromatic areas in Arabidopsis, the sRNA-directed DNA methylation pathway often causes characteristic peaks of CHH methylation and smRNA abundance at the edges of TEs, while CHG methylation tends to be lower at the edges of TEs and to increase and then plateau towards the centres (Zemach *et al.*, 2013). Our results for FriEPRV show that it is methylated almost

equally throughout its sequence (Figure 4b), while smRNA mapping shows several peaks across both coding and non-coding domains (Figures 2 and 4b). This pattern resembles the epigenetic landscape observed for *Gypsy* elements in Arabidopsis, which are mainly located in heterochromatic areas (Zemach *et al.*, 2013). However, we were unable to find any association between smRNA peaks and patterns of cytosine methylation.

The presence of SNPs (Figure S4) suggests that the FriEPRV sequences have not evolved in a concerted manner (as some arrays of repetitive sequences do) but have diverged. This is consistent with its comparatively high level of cytosine methylation in the CHG context, which is probably connected to condensed chromatin and hence a reduced frequency of the recombination required for homogenization processes (Peng and Karpen, 2008). Recombination-based processes are not only the basis for concerted evolution (Nei and Rooney, 2005) but also removal of repetitive sequences and hence genome downsizing (Grover and Wendel, 2010). The fact that most of the SNPs are C→T transitions probably reflects the deamination of methylated cytosine, which is indicative of long-term methylation of FriEPRV since its amplification in this species.

Overall, our analysis of the structure, methylation status and epigenetic regulation of FriEPRV clearly indicates that, at least for this genomic sequence in *F. imperialis*, epigenetic regulation is operational. These data thus provide direct evidence that the huge genomes of *F. imperialis* are not caused by a catastrophic breakdown in the epigenetic regulation of repeats leading to their runaway amplification.

EXPERIMENTAL PROCEDURES

Plant material and root preparations

The plant species studied are listed in Table S2. All species were grown at the Royal Botanic Gardens, Kew, UK, in pots on open ground, or were obtained from the personal collection of Laurence Hill (www.fritillariaicones.com). Root tips were collected and placed in a saturated solution of α–bromonaphtalene for 24 h on ice to accumulate cells at metaphase. The roots were then fixed in 3:1 v/v ethanol/glacial acetic acid, and incubated for 3 h at room temperature (approximately 20°C). Finally, roots were transferred to 100% ethanol and stored at −20°C.

DNA/RNA extraction

Total genomic DNA was extracted from silica-dried leaf tissue using a cetyl trimethylammonium bromide method as described by Kovarík *et al.* (2000). DNA samples were resuspended in 1× TE buffer, treated with RNase A and purified using NucleoSpin® Extract II columns (Macherey-Nagel, http://www.mn-net.com/).

Total RNA, including the small RNA fraction (smRNA, <200 nt), was extracted from mature leaf tissue that had been stored in RNAlater® (Qiagen, http://www.qiagen.com/) at −80°C. Extractions were performed using a mirVana™ miRNA isolation kit (Life Technologies, https://www.lifetechnologies.com) according to the manufacturer's instructions.

Illumina sequencing

Paired-end sequencing (2× 100 bp, 300–500 bp insert size) of total genomic DNA, sodium bisulfite-treated genomic DNA (unmethylated cytosine converted to uracil), and the smRNA fraction was performed by the Centre for Genomic Research at the University of Liverpool on the Illumina (http://www.illumina.com/) HiSeq® platform. smRNA libraries were size-selected to retain fragments with inserts of 6–66 bp. Sequencing data were supplied in FASTQ format with adaptors already trimmed.

Bioinformatics

RepeatExplorer. The RepeatExplorer pipeline (http://repeatexplorer.umbr.cas.cz/) clusters next-generation sequencing reads into groups of similar reads, and assembles contigs from these reads. The output may also be presented in graphical form by application of a Fruchterman–Reingold algorithm, which positions reads that are most similar closest together (Novák *et al.*, 2010). RepeatExplorer also annotates cluster regions using RepeatMasker and Repbase (Jurka *et al.*, 2005) and the conserved domain database (Marchler-Bauer *et al.*, 2011).

After removing reads with a Phred score of less than 20 for more than 10% of their bases, 2 000 842 Illumina paired-end reads of genomic DNA from *F. imperialis* (corresponding to 0.475% of the genome) were clustered into repeat families using RepeatExplorer (Novák *et al.*, 2010, 2013), as described previously (Renny-Byfield *et al.*, 2013).

Consensus sequences. To generate consensus sequences for the three looped domains of cluster FriEPRV (see Results), contigs containing ≥30 reads generated by RepeatExplorer (see Table 1) were further assembled manually through similarities in overlapping domains using Geneious® version 5.5.6 (http://www.geneious.com/). Contigs had low coverage at their ends. These low-coverage regions were removed if they showed ambiguities compared to high-coverage regions. Contigs 90 and 102, which contained divergent sequences at low coverage, were removed completely. The remainder were assembled into three consensus sequences corresponding to the three domains of FriEPRV: loop 1, loop 2 and loop 3. To further characterize these consensus sequences, MEGABLAST and BLASTx searches were performed using the National Center for Biotechnology Information online BLAST facility (http://blast.ncbi.nlm.nih.gov/Blast.cgi) with an *e*-value threshold of 10^{-5}.

Annotation of SNPs. We used the CLC Genomics Workbench 6.5.1 (CLCbio, http://www.clcbio.com/) to map all 2 000 842 genomic reads against the loop consensus sequences, and subsequently detected SNPs using CLC's function 'Probabilistic Variant Detection' with default settings.

Methylation analysis. The distribution of cytosine methylation in FriEPRV was analysed through comparison of sequences from sodium bisulfite-treated genomic DNA with consensus sequences from untreated DNA using Bismark (Krueger and Andrews, 2011), which comprises a set of Perl scripts that align bisulfite-modified FASTQ reads from Illumina data to reference sequences using the Bowtie short read aligner (Langmead *et al.*, 2009). Of 39 287 390 reads, approximately 0.1% could be mapped to the loop consensus sequences. For each cytosine in the consensus sequence, we recorded the number of bisulfite-treated reads with a cytosine or thymine at each cytosine position in the consensus, with thymine

arising as a consequence of bisulfite treatment of unmethylated cytosine. In order to avoid artefacts arising from low coverage, we considered only cytosine residues that were more than ninefold covered with reads derived from bisulfite-treated DNA. We excluded cytosine residues involved in SNPs (see Appendix S3). For each remaining cytosine in the consensus, we calculated the percentage of times that it was methylated in genomic DNA. These percentages indicate how many individual copies of FriEPRV are methylated at any particular cytosine.

Phylogenetic analysis. A 5361 nt open reading frame within FriEPRV loop 1 (see Results for details) was identified, translated and added to sequences from the GAGCOATPOL_caulimoviridae dataset (http://gydb.org/index.php/Collection_alignments), which contains genic domain sequences of the *Caulimoviridae* family. The amino acid sequences were aligned using T–Coffee (Notredame *et al.*, 2000), using the program's default parameters at the T–Coffee server (www.tcoffee.org/) (Di Tommaso *et al.*, 2011). Phylogenetic reconstruction using maximum parsimony and bootstrap analysis were performed using PAUP* version 4.0 b10 (Swofford, 2003), as described by Kelly *et al.* (2013); only positions scored as 'good' in the T–Coffee alignment were included in the phylogenetic analyses. Details of alignment scoring are given at www.tcoffee.org/Documentation/t_coffee/t_coffee_tutorial.htm.

RNA mapping. smRNA reads were quality-filtered using Biopython (http://biopython.org/, only Phred scores >30 allowed), and converted to FASTA format. smRNA reads were mapped to the three loop consensus sequences of FriEPRV using Bowtie (Langmead *et al.*, 2009), allowing only perfect matches over the whole read sequence.

PCR

The primer pairs, designed using PerlPrimer version .1.21 with default settings (downloaded from http://perlprimer.sourceforge.net/download.html), and the PCR conditions are shown in Table S1, and their positions are shown in Figure 1. PCR was performed in 20 μl volumes containing 1 unit of NEB® Taq polymerase (NEB, https://www.neb.com/), 1× NEB® Taq buffer, 200 μM of each nucleotide and approximately 20 ng *F. imperialis* DNA, with the addition of 0.8 μl formamide for 18S PCR.

Fluorescent *in situ* hybridization

Cell spreads. Cell spreads were prepared from fixed root tips as described by Lim *et al.* (2006). Briefly, root tips were lightly digested using 1% v/v pectinase and 2% v/v cellulase in citrate buffer, and spread under a coverslip in 60% v/v acetic acid. Coverslips were then removed after freezing in liquid nitrogen.

Probe labelling. PCR products (see above) were checked using agarose gel electrophoresis, and then purified and eluted in water using a QIAquick PCR purification kit (Qiagen). Purified PCR products were labelled by nick translation in 40 μl volumes containing 3 μg DNA, 0.01 M β–mercaptoethanol, 50 μM of each dATP, dGTP and dCTP, 10 μM dTTP, 50 μM labelled dUTP (Texas Red for the loop 1 probe; Alexa Fluor 488 for the18S and loop 3 probes), 50 mM Tris/HCl, 5 mM $MgCl_2$, 0.005% w/v BSA, 20 units DNA polymerase I (Invitrogen, http://www.lifetechnologies.com/) and 0.2 units DNase I (Thermo Scientific, http://www.thermoscientific.com/). Nick translation was performed at 15°C for 1 h. Products were checked via agarose gel electrophoresis.

FISH. Fluorescent *in situ* hybridization (FISH) was performed as described by Lim *et al.* (2006) with minor modifications. Briefly, slides were rinsed twice for 5 min in 2× SSC (0.3 M sodium chloride and 0.03 M tri-sodium citrate), fixed for 10 min in 4% v/v formaldehyde in 2× SSC, rinsed four times for 3 min in 2× SSC, dehydrated in ethanol of increasing concentrations (70, 95 and 100%) for 2 min each, and finally air-dried. Then 20 μl of hybridization mix containing 50% w/v formamide, 0.3 M sodium chloride, 0.03 M tri-sodium citrate, 10 mM Tris/HCl, 2 mM EDTA, 2.8 μg salmon sperm DNA (for blocking) and 75–150 ng probe were added to each slide. Slides were denatured at 80°C using a Dyad™ DNA engine carrying a PRINS block (Bio–Rad, http://www.bio–rad.com/). After incubation in an airtight plastic box at 37°C overnight, slides were quickly rinsed in 2× SSC at room temperature to remove coverslips, incubated in 2× SSC at 55°C for 20 min, briefly rinsed in 2× SSC again, dehydrated in ethanol as described above, and air-dried. Slides were mounted in a drop of Vector Shield™ mounting medium containing 4′,6–diamidino-2–phenylindole (Vector Laboratories, https://www.vectorlabs.com/), and stored for at least 1 h before microscope analysis.

Images were taken using a DMRA2 epifluorescence microscope (Leica, http://www.leica-microsystems.com/) equipped with an Orca ER™ monochrome camera (Hamamatsu, http://www.hamamatsu.com/). Contrast and brightness enhancement as well as merging of layers were performed using OpenLab™ imaging software (Improvision, Cambridge, UK, http://www.perkinelmer.co.uk/pages/020/cellularimaging/improvision/default.xhtml).

Southern hybridization. Southern hybridization was performed as described previously (Koukalova *et al.*, 2010). Extractions of total genomic DNA were sourced from the DNA bank at the Royal Botanic Gardens, Kew, UK (see Table S2); species of *Fritillaria* were selected to represent the phylogenetic diversity of the genus (Day *et al.*, 2014). DNA was digested with excess *Bst*NI (2× 2 h), and subjected to electrophoresis in agarose gels, using 1–2 μg of DNA/lane. Gels were blotted onto Hybond N+ membrane (GE Healthcare Life Sciences, http://www.gelifesciences.com/). After transfer, Southern blot hybridization was performed in 0.25 M sodium phosphate buffer, pH 7.0, supplemented with 7% w/v SDS at 65°C for 16 h, using α–^{32}P-dCTP-labelled PCR products from primer pair 1 (>10^8 dpm μg^{-1} DNA, Dekaprime kit, Fermentas, http://www.thermoscientificbio.com/fermentas/). The membrane was washed with both 2× SSC, 0.1% w/v SDS (2× 5 min) and 0.2× SSC, 0.1% w/v SDS (15 min each). The membrane was imaged using a Storm phosphorimager (Molecular Dynamics, http://www.gelifesciences.com/). The membrane was subsequently reprobed using α–^{32}P-dCTP-labelled PCR products from primer pair 3 as above.

ACKNOWLEDGEMENTS

We thank Richard Nichols and Steven Dodsworth (both Queen Mary University of London, School of Biological and Chemical Sciences) as well as Mike Fay (Royal Botanic Gardens Kew, Jodrell Laboratory) for helpful comments. We thank the Leonardo da Vinci programme for funding H.B. and the Marie Curie programme for funding L.M.; this work was also supported by the Natural Environment Research Council (grant number NE/G01724/1) and the Czech Science Foundation (P501/13/10057S). This paper includes Illumina data generated by the Centre of Genomic Research, which is based at the University of Liverpool, UK. We thank Mr Laurence Hill (Richmond, Surrey) for plant material.

SUPPORTING INFORMATION

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

Figure S1. Dot plot of the consensus sequences of loops 1–3 before trimming.

Figure S2. Phylogenetic placement of FriEPRV.

Figure S3. Length distribution of small RNAs matching the FriEPRV consensus sequences.

Figure S4. Graph showing the number of SNPs observed relative to the consensus sequence of loop 1.

Table S1. Primers and PCR conditions used.

Table S2. Plant material studied.

Data S1. Contigs and consensus sequences.

Appendix S1. Estimation of the FriEPRV genome proportion (GP).
Appendix S2. Copy number of FriEPRV.
Appendix S3. Cytosine methylation analysis: SNPs in genomic reads.

REFERENCES

Bennett, M.D., Leitch, I.J., Price, H.J. and Johnston, J.S. (2003) Comparisons with *Caenorhabditis* (~100 Mb) and *Drosophila* (~175 Mb) using flow cytometry show genome size in *Arabidopsis* to be ~157 Mb and thus 25% larger than the *Arabidopsis* Genome Initiative estimate of ~125 Mb. *Ann. Bot.* **91**, 547–557.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580.

Chabannes, M. and Iskra-Caruana, M.-L. (2013) Endogenous pararetroviruses - a reservoir of virus infection in plants. *Curr. Opin. Virol.* **3**, 615–620.

Chabannes, M., Baurens, F.C., Duroy, P.O., Bocs, S., Vernerey, M.S., Rodier-Goud, M., Barbe, V., Gayral, P. and Iskra-Caruana, M.L. (2013) Three infectious viral species lying in wait in the banana genome. *J. Virol.* **87**, 8624–8637.

Day, P.D., Berger, M., Hill, L., Fay, M.F., Leitch, A.R., Leitch, I.J. and Kelly, L.J. (2014) Evolutionary relationships in the medicinally important genus *Fritillaria* L. (Liliaceae). *Mol. Phylogen. Evol.* **80**, 11–19.

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.-M., Taly, J.-F. and Notredame, C. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* **39**, W13–W17.

Grover, C. and Wendel, J.F. (2010) Recent insights into mechanisms of genome size change in plants. *J. Bot.* **2010**, Article ID 382732.

Hansen, C.N., Harper, G. and Heslop-Harrison, J.S. (2005) Characterisation of pararetrovirus-like sequences in the genome of potato (*Solanum tuberosum*). *Cytogenet. Genome Res.* **110**, 559–565.

Hawkins, J.S., Proulx, S.R., Rapp, R.A. and Wendel, J.F. (2009) Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proc. Natl Acad. Sci. USA*, **106**, 17811–17816.

Hu, T.T., Pattyn, P., Bakker, E.G. *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481.

Jakowitsch, J., Mette, M.F., van der Winden, J., Matzke, M.A. and Matzke, A.J.M. (1999) Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. *Proc. Natl Acad. Sci. USA*, **96**, 13241–13246.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467.

Kanizay, L. and Dawe, R.K. (2009) Centromeres: long intergenic spaces with adaptive features. *Funct. Integr. Genomics*, **9**, 287–292.

Kanizay, L.B., Albert, P.S., Birchler, J.A. and Dawe, R.K. (2013) Intragenomic conflict between the two major knob repeats of maize. *Genetics*, **194**, 81–89.

Kelly, L.J. and Leitch, I.J. (2011) Exploring giant plant genomes with next-generation sequencing technology. *Chromosome Res.* **19**, 939–953.

Kelly, L.J., Leitch, A.R., Fay, M.F., Renny-Byfield, S., Pellicer, J., Macas, J. and Leitch, I.J. (2012) Why size really matters when sequencing plant genomes. *Plant Ecol. Divers.* **5**, 415–425.

Kelly, L.J., Leitch, A.R., Clarkson, J.J., Knapp, S. and Chase, M.W. (2013) Reconstructing the complex evolutionary origin of wild allopolyploid tobaccos (*Nicotiana* section *Suaveolentes*). *Evolution*, **67**, 80–94.

Kim, M.Y. and Zilberman, D. (2014) DNA methylation as a system of plant genomic immunity. *Trends Plant Sci.* **19**, 320–326.

Kim, S., Park, M., Yeom, S.-I. *et al.* (2014) Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat. Genet.* **46**, 270–278.

Koukalova, B., Moraes, A.P., Renny-Byfield, S., Matyasek, R., Leitch, A.R. and Kovarik, A. (2010) Fall and rise of satellite repeats in allopolyploids of *Nicotiana* over c. 5 million years. *New Phytol.* **186**, 148–160.

Kovarík, A., Koukalová, B., Lim, K.Y., Matyásek, R., Lichtenstein, C.P., Leitch, A.R. and Bezdek, M. (2000) Comparative analysis of DNA methylation in tobacco heterochromatic sequences. *Chromosome Res.* **8**, 527–541.

Krueger, F. and Andrews, S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25.

Leitch, A.R. and Leitch, I.J. (2012) Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytol.* **194**, 629–646.

Leitch, I.J. and Leitch, A.R. (2013) Genome size diversity and evolution in land plants. In *Plant Genome Diversity*, **vol 2**, Physical structure, behaviour and evolution of plant genomes (Leitch, I.J., Greilhuber, J., Doležel, J. and Wendel, J.F., eds). Wien: Springer-Verlag, pp. 307–322.

Leitch, I.J., Beaulieu, J.M., Cheung, K., Hanson, L., Lysak, M. and Fay, M.F. (2007) Punctuated genome size evolution in Liliaceae. *J. Evol. Biol.* **20**, 2296–2308.

Lim, K.Y., Kovarik, A., Matyasek, R., Chase, M.W., Knapp, S., McCarthy, E., Clarkson, J.J. and Leitch, A.R. (2006) Comparative genomics and repetitive sequence divergence in the species of diploid *Nicotiana* section *Alatae*. *Plant J.* **48**, 907–919.

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

Macas, J., Neumann, P. and Navrátilová, A. (2007) Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. *BMC Genomics*, **8**, 427

Malik, H.S. and Henikoff, S. (2009) Major evolutionary transitions in centromere complexity. *Cell*, **138**, 1067–1082.

Marchler-Bauer, A., Lu, S., Anderson, J.B. *et al.* (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225–D229.

Matzke, M.A. and Mosher, R.A. (2014) RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15**, 394–408.

Nei, M. and Rooney, A.P. (2005) Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39**, 121–152.

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., Widmer, A., Doležel, J. and Macas, J. (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mobile DNA*, **2**, 4.

Noreen, F., Akbergenov, R., Hohn, T. and Richert-Poggeler, 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.

Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217.

Novák, P., Neumann, P. and Macas, J. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, **11**, 378.

Novák, 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.

Nystedt, B., Street, N.R., Wetterbom, A. *et al.* (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**, 579–584.

Peng, J.C. and Karpen, G.H. (2008) Epigenetic regulation of heterochromatic DNA stability. *Curr. Opin. Genet. Dev.* **18**, 204–211.

Renny-Byfield, S., Chester, M., Kovařík, A. *et al.* (2011) Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. *Mol. Biol. Evol.* **28**, 2843–2854.

Renny-Byfield, S., Kovarik, A., Kelly, L.J., Macas, J., Novak, P., Chase, M.W., Nichols, R.A., Pancholi, M.R., Grandbastien, M.A. and Leitch, A.R. (2013) Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. *Plant J.* **74**, 829–839.

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

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.

Simon, S.A. and Meyers, B.C. (2011) Small RNA-mediated epigenetic modifications in plants. *Curr. Opin. Plant Biol.* **14**, 148–155.

Staginnus, C., Gregor, W., Mette, M.F., Teo, C., Borroto-Fernandez, E., Machado, M.L., Matzke, M. and Schwarzacher, T. (2007) Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. *BMC Plant Biol.* **7**, 24.

Swofford, D.L. (2003) *PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods)* Version 4. Sunderland, MA: Sinauer Associates.

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

Zemach, A., Kim, M.Y., Hsieh, P.H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S.L. and Zilberman, D. (2013) The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell*, **153**, 193–205.

Zhang, H., Koblízková, A., Wang, K. *et al.* (2014) Boom-bust turnovers of megabase-sized centromeric DNA in *Solanum* species: rapid evolution of DNA sequences associated with centromeres. *Plant Cell*, **26**, 1436–1447.