Check for updates

New Phytologist

Endogenous viral elements are targeted by RNA silencing pathways in banana

Pierre-Olivier Duroy[1,2] (iD), Jonathan Seguin[3], Sébastien Ravel[1,2] (iD), Rajeswaran Rajendran[3] (iD), Nathalie Laboureau[1,2] (iD), Frédéric Salmon[4,5], Jean-Marie Delos[4,5], Mikhail Pooggin[1,2] (iD), Marie-Line Iskra-Caruana[6] (iD) and Matthieu Chabannes[7,8] (iD)

[1]CIRAD, UMR PHIM, Montpellier, F-34398, France; [2]PHIM Plant Health Institute, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France; [3]Department of Plant Physiology, Botanical Institute, Zürich-Basel Plant Science Center, University of Basel, Basel, Switzerland; [4]CIRAD, UMR AGAP Institut, Capesterre-Belle-Eau, Guadeloupe, F-97130, France; [5]UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, Capesterre-Belle-Eau, Guadeloupe, France; [6]CIRAD, DGD-RS, Montpellier, 34398, France; [7]CIRAD, UMR AGAP Institut, Montpellier, F-34398, France; [8]UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France

Summary

- Endogenous banana streak virus (eBSV) integrants derived from three distinct species, present in *Musa balbisiana* (B) but not *Musa acuminata* (A) banana genomes are able to reconstitute functional episomal viruses causing banana streak disease in interspecific triploid AAB banana hybrids but not in the diploid (BB) parent line, which harbours identical eBSV loci. Here, we investigated the regulation of these eBSV.

- In-depth characterization of siRNAs, transcripts and methylation derived from eBSV using Illumina and bisulfite sequencing were carried out on eBSV-free *Musa acuminata* AAA plants and BB or AAB banana plants with eBSV.

- eBSV loci produce low-abundance transcripts covering most of the viral sequence and generate predominantly 24-nt siRNAs. siRNA accumulation is restricted to duplicated and inverted viral sequences present in eBSV. Both siRNA-accumulating and nonaccumulating sequences of eBSV in BB plants are heavily methylated in all three CG, CHG and CHH contexts.

- Our data suggest that eBSVs are controlled at the epigenetic level in BB diploids. This regulation not only prevents their awakening and systemic infection of the plant but is also probably involved in the inherent resistance of the BB plants to mealybug-transmitted viral infection. These findings are thus of relevance to other plant resources hosting integrated viruses.

Introduction

The burst of genome sequence availability arising from new sequencing technologies has uncovered innumerable viral sequences integrated into the genomes of both eukaryotes and prokaryotes. These so-called endogenous viral elements (EVE), as well as transposable elements (TE), were acquired during evolution by horizontal gene transfer, and the neo-function of these sequences has been instrumental in helping organisms adapt successfully to changing environmental stresses and conditions. However, TE and EVE can either serve as strong constitutive promoters to neighbouring plant genes or inactivate gene products by insertional mutagenesis (Jakowitsch *et al.*, 1999; Gilbert & Feschotte, 2018). To counter these potentially negative effects, plants have developed mechanisms to control TE and EVE integration, keeping them transcriptionally silent (Ito, 2013).

Transcriptional silencing of TEs, which is established and maintained by the RNA-dependent DNA methylation (RdDM) pathway, has been fairly well studied (reviewed in Matzke & Mosher, 2014). Plants also use an RNA interference mechanism – conserved in eukaryotes – to inhibit TE activity and virus replication via transcriptional gene silencing (TGS) or post-transcriptional gene silencing (PTGS) mechanisms (reviewed in Baulcombe, 2022; Vaucheret & Voinnet, 2024). Cytoplasmic PTGS is the major pathway to control RNA viruses as they replicate in the cytoplasm (Deleris *et al.*, 2006; Qu *et al.*, 2008), where the TGS pathway plays a very minor role (Raja *et al.*, 2008); however, plants employ components of the nuclear TGS pathway against episomal DNA viruses, as illustrated for the geminivirus *Cabbage leaf curl virus* (CaLCuV) (Akbergenov *et al.*, 2006; Raja *et al.*, 2014) and the pararetrovirus *Cauliflower mosaic virus* (CaMV) in *Arabidopsis thaliana* (Blevins *et al.*, 2011).

Besides LTR retrotransposons of the two large families Metaviridae and Pseudoviridae (ssRNA-RT viruses that constitute large proportions of plant genomes), DNA viruses of the families Geminiviridae and Caulimoviridae, and some ssRNA viruses, are

also found in plant genomes (Chiba *et al.*, 2011; da Fonseca *et al.*, 2016). Amongst DNA viruses, members of the Caulimoviridae are the most abundant; their integrants are named endogenous pararetroviruses (EPRVs). Integrated copies of Caulimoviridae have been identified in the nuclear genomes of many agriculturally and horticulturally important plant species, including very recently cotton (Aboughanem-Sabanadzovic *et al.*, 2023) and raspberry (Ho *et al.*, 2024) (for reviews, see Chabannes & Iskra-Caruana, 2013; Vassilieff *et al.*, 2023). Besides being spread horizontally between individuals and across species by mechanical and vector-mediated transmission, members of the Caulimoviridae can also be transmitted vertically via integrated sequences fixed in host genomes. Indeed, EVEs of the Caulimoviridae family have been identified in genomes of almost every vascular plant examined, including primitive taxa such as clubmosses, ferns and gymnosperms (Diop *et al.*, 2018; Gong & Han, 2018). Although most such integrations are relics of ancient infection events, existing as either partial or nonfunctional viral sequences with null mutations (Diop *et al.*, 2018; Gong & Han, 2018; Chabannes *et al.*, 2021), some are active and cause diseases. To date, the formation of virus particles from EPRVs has been demonstrated clearly only for *Nicotiana edwardsonii* (Lockhart *et al.*, 2000), petunia (Richert-Poggeler *et al.*, 2003) and some interspecific banana hybrids (Gayral *et al.*, 2008; Chabannes *et al.*, 2013).

In banana, such infective EPRVs represent a major constraint for breeding programmes and thus are of considerable economic importance. Although banana EPRVs have been characterized extensively (Gayral *et al.*, 2008; Chabannes *et al.*, 2013), their regulation remains largely unexplored (Ricciuti *et al.*, 2021). Most cultivated varieties of banana world-wide have resulted from intra- or interspecific crosses of the seedy diploid subspecies of *Musa acuminata* Colla (A genome) and *Musa balbisiana* Colla (B genome). All known banana plants bearing the B genome naturally harbour infective EPRVs (Gayral *et al.*, 2010; Duroy *et al.*, 2016). These viral integrants are named endogenous banana streak viruses (eBSV) and correspond to at least three BSV species (Obino l'Ewai virus – BSOLV; Goldfinger virus – BSGFV; Imové virus – BSIMV) distributed world-wide that share 53–58% identity at the nucleic acid level. Interestingly, the seedy *M. balbisiana* diploid genitor Pisang Klutuk Wulung (PKW) is naturally resistant to both BSOLV, BSIMV and BSGFV infections mediated by mealybug vectors and eBSV-mediated infections (Lheureux, 2002), although PKW plants have been subjected to intensive *in vitro* culture, which is known to activate eBSVs in interspecific hybrids containing these integrants.

In this study, we characterized the regulation of eBSVs in PKW and its self-pollinated segregating progeny to uncover the underlying mechanisms and investigate the impact of *M. balbisiana* genome (B) ploidy on this regulation. We also analysed four interspecific AAB hybrids resulting from a cross between PKW and the tetraploid AAAA banana IDN-T 110, which lacks any eBSV, to explore why infective eBSVs are able to release particles in most interspecific hybrids (Dallot *et al.*, 2001; Lheureux *et al.*, 2003; Cote *et al.*, 2010).

Materials and Methods

Banana plants and viruses

Banana (*Musa* L.) plants were maintained at CIRAD (Montpellier) in a tropical glasshouse by vegetative propagation (i.e. growing of suckers) in 12-h daylight with luminosity not exceeding 400 w m^{-2}, 75% relative humidity and 26°C : 24°C, day : night.

Pisang Klutuk Wulung is a seedy *Musa balbisiana* diploid plant (BB genome) introduced at CIRAD Montpellier in 2000 from the open-field collection at Neufchâteau station (Guadeloupe). Self-pollinated PKW plants were produced at the CIRAD station (Guadeloupe) in 2012 and introduced at Montpellier after embryo rescue. Plants were propagated as described above.

Banana streak Obino l'Ewai Virus (BSOLV – Genbank accession no.: AJ002234.1), Banana streak Goldfinger Virus (BSGFV – Genbank accession no.: AY493509.1) and Banana streak Imové Virus (BSIMV – Genbank accession no.: HQ659760.1)-infected *Musa acuminata* Cavendish plants, constituting BSV controls, are as described (Rajeswaran *et al.*, 2014). The three different BSV species diverge by at least 20% at the nucleic level in the reverse transcriptase region as defined by the International Committee on Taxonomy of Viruses (ICTV).

The four allotriploid AAB hybrids used in this study were obtained from interspecific hybridization between a virus-free autotetraploid *M. acuminata* male parent (IDN110 Tetra, AAAA) and the diploid PKW female parent. Plants were checked regularly for BSV infection by immune-capture PCR (IC-PCR) (Le Provost *et al.*, 2006). We selected one BSV-free hybrid plant as a control and three hybrid plants self-infected with BSOLV, BSIMV or BSGFV from respective eBSV. The three infected hybrid plants all show the typical streak symptoms on the leaves associated with banana streak disease (Iskra Caruana *et al.*, 2019).

A summary of the plants and viral strains used in this work is illustrated in Fig. 1.

Total nucleic acid preparation

The same banana leaf was used for preparation of total RNA and total DNA as described by Rajeswaran *et al.* (2014). Total RNA for Illumina sequencing analyses was extracted as follows.

Two grams of banana leaf tissue ground in liquid nitrogen was added to 10 ml of extraction buffer (100 mM Tris, 500 mM NaCl, 25 mM EDTA, 1.5% SDS, 2% PVP and 0.7% 2-mercaptoethanol). The mixture was vortexed, incubated at room temperature for 10 min and centrifuged at 1500 *g* for 15 min to pellet the debris. The one-third volume of precooled sodium acetate (pH 6.0) was added to the supernatant, mixed, incubated on ice for 30 min and then centrifuged at 11 000 *g* for 15 min at 4°C. The supernatant was treated with an equal volume of phenol/chloroform/isoamyl alcohol and centrifuged at 10 000 rpm for 10 min at 4°C. The supernatant was extracted with an equal volume of chloroform/isoamyl alcohol and centrifuged as above. RNA in the aqueous phase was precipitated by addition of two to three volumes of cold ethanol, incubated at −70°C for 30 min and
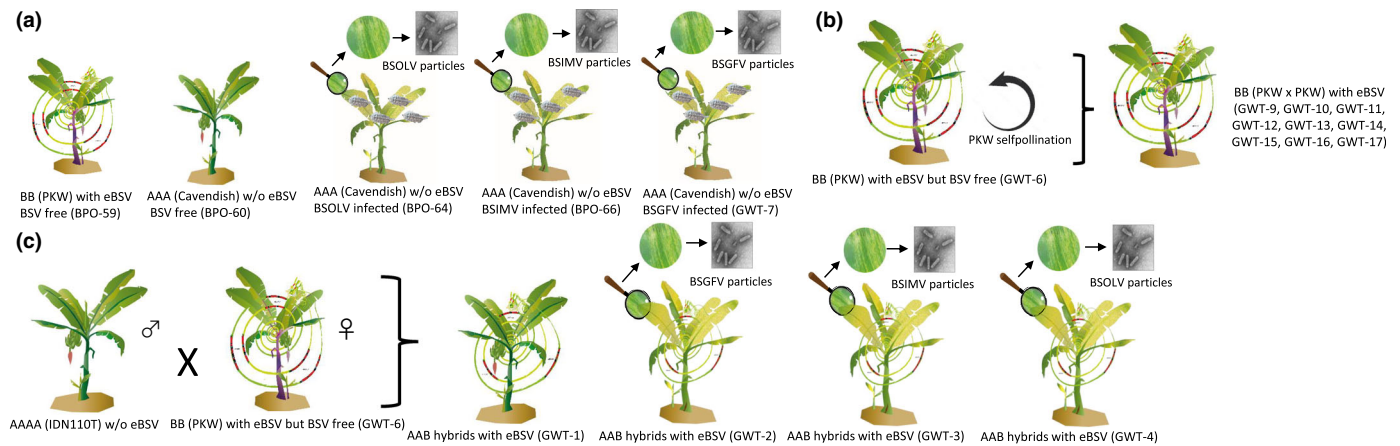
Fig. 1 Schematic representation of the banana plant and viral strains used to characterize the regulation of endogenous banana streak virus (eBSV) in banana. All the plants described here have been deep-sequenced for small RNA (sRNA). The resulting samples' names (BPO or GWT) have been used throughout the manuscript to simplify its reading. (a) Characterization of eBSV-free Cavendish plants infected with different BSV species and PKW, a BSV-free diploid *Musa balbisiana* plant harbouring eBSV. *Musa acuminata* dwarf Cavendish plants (triploid AAA genome without endogenous banana streak virus sequences) (BPO-60) were inoculated by mealybugs (*Planococcus citri*) fed on banana plants infected with single BSV species each as described in Rajeswaran *et al*. (2014) (BPO-64 (Banana streak Obino l'Ewai virus (BSOLV)-infected Cavendish plant), BPO-66 (Banana streak Imové virus (BSIMV)-infected Cavendish plant) and GWT-7 (Banana streak Goldfinger virus (BSGFV)-infected Cavendish plant). The resulting BSV-infected dwarf Cavendish collection was maintained at CIRAD (Montpellier) in a tropical glasshouse by vegetative propagation (i.e. growing of infected suckers). PKW (diploid *Musa balbisiana* (BB) plant – BPO-59) is homozygous for endogenous banana streak Imové sequences (eBSIMV), is heterozygous for endogenous banana streak Goldfinger sequences (eBSGFV) with eBSGFV-7 (GF7) infective and eBSGFV-9 (GF9) noninfective alleles and for endogenous banana streak Obino l'Ewai sequences (eBSOLV) with eBSOLV-1 (OL1) infective and eBSOLV-2 (OL2) noninfective alleles (Chabannes *et al*., 2013). eBSVs are shown as thick red lines in the *M. balbisiana* (B) genome, which is represented in a circular fashion. The three eBSV species are integrated at different loci. (b) Characterization of PKW and different self-pollinated diploid offsprings containing different eBSV allelic combinations. The diploid *M. balbisiana* PKW (GWT-6) was self-pollinated to segregate out the different eBSV alleles. PKW's eBSV genotype is as described in (a). Amongst the different offsprings, nine self-pollinated PKW plants were further characterized. eBSV genotypes for the PKW self-pollinated progeny are: GWT-9 (OL1 + OL2; GF7 + GF9); GWT-10 (OL2 + OL2; GF9 + GF9); GWT-11 (OL2 + OL2; GF7 + GF9); GWT-12 (OL1 + OL1; GF9 + GF9); GWT-13 (OL2 + OL2; GF7 + GF7); GWT-14 (OL1 + OL1; GF7 + GF9); GWT-15 (OL1 + OL2; GF9 + GF9); GWT-16 (OL1 + OL2; GF7 + GF7); GWT-17 (OL1 + OL1; GF7 + GF7) where OL1 stands for eBSOLV-1 infective allele, OL2 for eBSOLV-2 noninfective allele, GF7 for eBSGFV-7 infective allele and GF9 for eBSGFV-9 noninfective allele. All plants are homozygous for eBSIMV. (c) Characterization of different interspecific triploid offsprings generated by the cross between PKW and a tetraploid (AAAA) *M. acuminata* plant. The four allotriploid AAB hybrids used in this study were obtained from interspecific cross between a virus-free autotetraploid *M. acuminata* male parent (IDN110 Tetra, AAAA) and the diploid *M. balbisiana* PKW female parent (GWT-6). One episomal BSV-free hybrid plant as a control (GWT-1) and three hybrid plants self-infected with BSGFV (GWT-2), BSIMV (GWT-3) or BSOLV (GWT-4) following corresponding eBSV counterpart activation were selected for the analysis. The three infected hybrid plants show the typical streak symptoms on the leaves associated with the banana streak disease. The nature of the BSV species responsible for the infection of each of the three hybrids was determined by immune-capture PCR (IC-PCR) (Chabannes *et al*., 2021). PKW's eBSV genotype is as described in (a). eBSV genotypes for the AAB interspecific hybrids are as follows: GWT-1 (OL2; GF9); GWT-2 (OL2; GF7); GWT-3 (OL2; GF9); and GWT-4 (OL1; GF9); all have one copy of eBSIMV. eBSVs are shown as thick red lines as described above.

pelleted by spinning at 10 000 rpm for 15 min at 4°C. RNA was dissolved in 500 µl of DEPC-treated water, extracted again with chloroform/isoamyl alcohol and precipitated with ethanol as above. The RNA pellet was washed with 75% ethanol, air-dried, dissolved in DEPC-treated water and stored at −80°C until use.

Total DNA for methylation profile analysis was isolated as follows. Briefly, 100 mg of banana leaf tissues ground in liquid nitrogen was incubated with 500 µl extraction buffer (100 mM Tris, 1.4 M NaCl, 20 mM EDTA, 2% MATAB, 1% PEG 6000 and 0.5% sodium sulfite) preheated at 74°C. The slurry was mixed and incubated at 74°C for 20 min, and 500 µl of chloroform:isoamyl alcohol was added. The contents were mixed and centrifuged at 21 500 *g* for 15 min. DNA in the aqueous phase was precipitated with an equal volume of cold isopropanol and pelleted by centrifugation at 14 000 rpm for 30 min at 4°C.

After washed with 70% ethanol, the DNA pellet was dried in a speed vacuum and dissolved in 100 µl of water.

sRNA blot hybridization analysis

Small RNA (sRNA) blot hybridization analysis was performed as described (Blevins *et al*., 2006) using the short DNA oligonucleotide probes listed in Supporting Information Table S1. For each hybridization shown in Fig. S1, a mixture of 4–5 ^{32}P-labelled DNA oligonucleotides specific for the BSOLV, BSIMV or BSGFV species corresponding to four to five sRNA hotspots identified by PKW sRNA sequencing analysis was used as a probe. One hybridization per BSV species was performed on total RNA of PKW and the eBSV-free AAA banana plant infected with BSOLV described previously (Rajeswaran *et al*., 2014).

DNA methylation analyses

The cytosine methylation status of eBSV loci was determined using the methylation-dependent McrBC restriction enzyme and by bisulfite sequencing.

For McrBC treatment, 300 ng of total DNA from PKW was digested with 10 U McrBC enzyme (New England BioLabs, Ipswich, MA, USA) in a final volume of 50 μl for 3 h at 37°C as recommended by the manufacturer. As a positive control for McrBC, 300 ng of methylated plasmid (with one McrBC site, supplied by the manufacturer) was spiked in 300 ng of PKW total DNA. Control samples without the McrBC enzyme were incubated in parallel under the same conditions as test samples. Following McrBC treatment, 3 μl (18 ng DNA) of the treated and nontreated samples were used for PCR in a 25 μl reaction mix containing 1 U GoTaq DNA polymerase, 1× GoTaq® reaction buffer (Promega), 200 nM of each primer and 100 μM of each dNTP. Thermal cycling conditions were as follows: 1 cycle at 95°C for 5 min, 25 cycles at 95°C for 30 s, 62°C for 30 s, 72°C for 2 min, followed by one cycle at 72°C for 10 min. PCR products were resolved in 1% agarose gel in 1× TBE buffer. Five different primer pairs were used for methylation analysis of the eBSIMV locus, three targeting the inverted repeat regions producing viral siRNAs (F1F3F 5′-TTCGGTATTATT AAAACTCACACCA-3′, Im6R 5′-CCGACTTGAGGATT TTCCAA-3′; Im6R 5′-CCGACTTGAGGATTTTCCAA-3′, Im 7F 5′-CATCAAAACTCACGCCTTCA-3′; F4F5F 5′-T GGACAGCTTCTGGTGTGAG-3′, Im8R 5′-TCCAGAAG AGTCAGCCCAAA-3′) and two targeting nonrepetitive viral sequence regions that do not produce siRNAs (F3F4R 5′-GAA GCTGTCCAAGCCTATATCA-3′, ImF 5′- TGCCAACGAA TACTACATCAAC-3′; ImF 5′-TGCCAACGAATACTACAT CAAC-3′, Im9R 5′-AGACTTCGCGTGTGAATTTTT-3′).

For bisulfite sequencing, 1 μg of genomic DNA from PKW was processed with an EpiTect Bisulfite kit according to the protocol supplied by the manufacturer (Qiagen). Treated DNA was resuspended in a final volume of 20 μl elution buffer (10 mM Tris pH 8.0). The PCR and sequencing primers listed in Table S2 were designed using the nonselective variant option of the BIS-PRIMER software (Kovacova & Janousek, 2012) meaning that selected primers are able to bind both completely and incompletely modified DNA templates with the same efficiency. Bisulfite-treated DNA (50 ng) was used for PCR in a final volume of 50 μl containing 1 U of Pfu Hotstart Polymerase (Agilent), 1× reaction buffer (Agilent), 500 nM of each primer and 100 μM of each dNTP. Thermal cycling conditions, primer names and sequences, and expected amplicon sizes for the different eBSV species and for one coding region external to the eBSV sequences (Zonadhesin-related protein – id:BAG70984.1) used as a negative control are listed in Table S2. The resulting PCR products were gel-purified and then sequenced in bulk using the same primers as in the PCR reaction using an ABI 3730 capillary sequencer (Sanger method). The raw chromatograms are provided as Note S1. Following bisulfite treatment, genomic DNA is often present in single-stranded form due to mismatches following the conversion of some of the cytosines, and primers designed using BIS-PRIMER software preferentially amplify the positive strand of the DNA. Sequencing reads were curated manually before aligning sequences with the positive strand of the reference sequence. Methylated and nonmethylated cytosines in each context were then counted.

Illumina deep-sequencing and bioinformatics analysis

Small RNA sequencing RNA quantity was estimated by Qubit and quality checked by loading 100 ng on a Bioanalyzer Nano chip (Agilent 2100). For sRNA deep-sequencing, cDNA libraries of the 18–30-nt RNA fraction of total RNA samples were prepared at two different time points that are 2 yr apart using TRU-SEQ KITS v.5 for samples BPO or v.3 for samples GWT (see Methods S1) and sequenced in separate lanes of Hi-Seq 2000 (BPO-59-66, GWT-1 to −7) or Hi-Seq 2500 (GWT-9 to GWT-17). The resulting samples' names (BPO or GWT) have been used throughout the manuscript to simplify its reading. To analyse mapping × samples × references, we developed a snakemake workflow (Mölder et al., 2021) to perform all steps in parallel on the CIRAD Cluster. After trimming adaptor sequences with ATROPOS v.1.1.25 (Didion et al., 2017), datasets of reads were mapped to reference genome sequences (as detailed in Methods S1) using a Burrows-Wheeler Alignment Tool (BWA v.0.5.9) (Li & Durbin, 2010) with zero or up to two mismatches to each reference sequence. SAMTOOLs v.1.9 was then used to filter and produce the merge bam files. The same read was allowed to map to up to 10 different positions of the reference and was assigned randomly to one of the repeated positions. The bioinformatics analysis of the mapped reads is summarized in Figs 2–4, 6–8 (see later); Dataset S1; Figs S2–S4. To account for the circular BSV genome, the first 50 bases of the viral sequence were added to its 3′-end. For each reference genome/sequence and each sRNA size class (20–25 nt), we counted total number of reads, reads in forward and reverse orientation, and reads starting with A, C, G and T (Dataset S1). The single-base resolution maps of 20-, 21-, 22-, 23-, 24- and 25-nt viral sRNAs (Figs 3, 4, 6, 8, S3, S4; Dataset S2) were generated by a homemade Python script based on the MISIS tool (Seguin et al., 2014, 2016) that includes data normalization. Numbers of viral sRNA reads were normalized per million of total 20–25-nt reads in each library. In the maps, for each position on the sequence (starting from the 5′ end of the reference sequence), the number of matches starting at this position in forward (first base) and reverse (last base) orientation for each read length is given. Note that reads mapped to the last 50 bases of the extended viral sequence were added to reads mapped to the first 50 bases.

Total RNA sequencing Libraries were prepared from 500 ng of DNase-treated total RNA with the Stranded Total RNA kit from Illumina coupled with the Ribo-Zero ribosomal RNA depletion kit. Libraries were sequenced using the Illumina HiSeq2500 platform (Fasteris) and the HiSeq SBS kit v.4, which produced 125-nt single reads.

Reads from each library were mapped to the same three BSV species and the same PKW BAC clones as those described above for
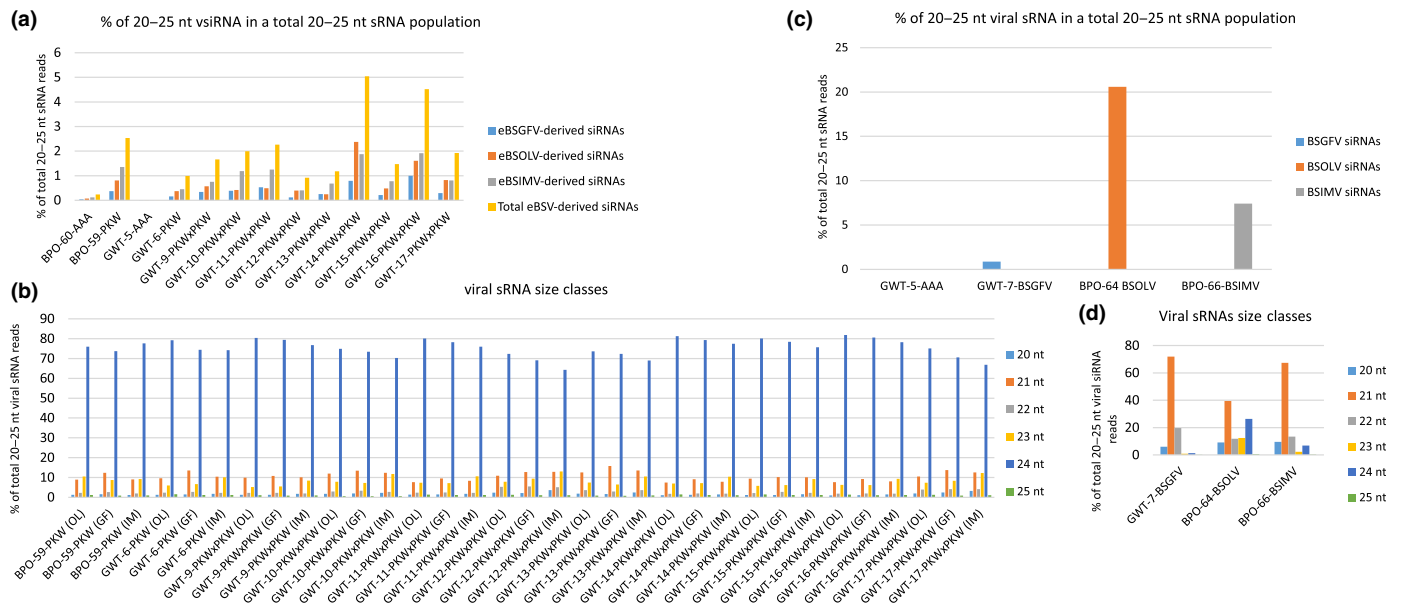
Fig. 2 Characterization of viral small RNAs (sRNAs) in the seedy diploid Pisang klutuk Wulung (PKW) plant, PKW self-pollinated progeny, triploid *Musa acuminata* plants infected with three distinct BSV species and healthy control plants. The graphs show percentages of 20–25-nt endogenous banana streak virus (eBSV)- (a) or BSV- (c) derived sRNAs in the pool of total (host and viral) 20–25-nt reads mapped to the respective virus genomes (eBSOLV, eBSGFV, eBSIMV, BSOLV, BSGFV or BSIMV) with zero mismatches (a, c) and the percentages of each size class of 20–25-nt viral sRNA reads mapped to each of the BSV species genomes with zero mismatches (b, d). In (b), the percentages of each size class of 20–25-nt eBSV-derived sRNA reads are represented for each of the BSV species (BSOLV (OL), BSGFV (GF) and BSIMV (IM)) that have eBSV counterparts present in the banana genome. The eBSV genotypes of plants for Panels 2A and 2B are as follows: PKW (BPO59 or GWT-6) is homozygous for eBSIMV, is heterozygous for eBSGFV with eBSGFV-7 (GF7) infective and eBSGFV-9 (GF9) noninfective alleles and for eBSOLV with eBSOLV-1 (OL1) infective and eBSOLV-2 (OL2) noninfective alleles (Gayral et al., 2008; Chabannes et al., 2013). eBSV genotypes for PKW self-pollinated progeny are as follows: GWT-9 (OL1 + OL2; GF7 + GF9); GWT-10 (OL2 + OL2; GF9 + GF9); GWT-11 (OL2 + OL2; GF7 + GF9); GWT-12 (OL1 + OL1; GF9 + GF9); GWT-13 (OL2 + OL2; GF7 + GF7); GWT-14 (OL1 + OL1; GF7 + GF9); GWT-15 (OL1 + OL2; GF9 + GF9); GWT-16 (OL1 + OL2; GF7 + GF7); GWT-17 (OL1 + OL1; GF7 + GF7). PKW and PKW × PKW plants are all homozygous for eBSIMV. In (c, d), the percentage of 20–25-nt viral sRNAs are represented for a BSV-free Cavendish plant (GWT-5), a BSGFV-infected (GWT-7), BSOLV-infected (BPO-64) or BSIMV-infected Cavendish plant (BPO-66). All of them are deprived of eBSV.

sRNA analysis using the same pipeline and software to the ready-made differences: zero to six mismatches to each reference were allowed and, to account for the circular BSV genome, the first 126 bases of the viral sequence were added to its 3′-end. Reads mapped to the last 126 bases of the extended viral sequence were added to reads mapped to the first 126 bases. The same read was allowed to map to up to 10 different positions of the reference and was assigned randomly to one of the repeated positions. The numbers of viral RNA reads were normalized per million of total reads in each library. The bioinformatics analysis of the mapped reads is summarized in Fig. 5 and detailed in Datasets S3 and S4.

eBSV genotyping

The eBSV allelic composition of the four AAB (IDN110 × PKW) hybrids and PKW self-pollinated plants was assessed by PCR using the procedure and primer sets developed previously (Gayral et al., 2008; Chabannes et al., 2013). Briefly, all PCRs were performed with 20 ng of genomic DNA, 20 mM Tris–HCl (pH 8.4), 50 mM KCl, 100 mM concentration of each dNTP, 1.5 mM MgCl2, 10 pmol of each primer and 1 U of GoTaq DNA polymerase (Promega) in a total reaction volume of 25 μl. PCR conditions were as follows: 1 cycle at 94°C for 5 min, followed by 35 cycles at 94°C for 30 s, 60°C for 30 s for all

eBSGFV, eBSIMV and eBSOLV primers with the exception of primers' couples Dif-OL (F)/(R) (HaeIII), Dif-OL (F)/(R) (AhdI), Marker1-BSOLV2 (F)/(R) and Marker2-BSOLV1 (F)/(R) that have a 65°C annealing temperature (Chabannes et al., 2013), 72°C for 1 min per kb and one final extension at 72°C for 10 min. The sequence of the primers and the size of the PCR product are given in Chabannes et al. (2013). PCR products were visualized under UV light after migration of 10 μl of PCR products on a 1.5% agarose gel in 0.5× TBE (45 mM Tris-borate, 1 mM EDTA, (pH 8)) stained with ethidium bromide. The housekeeping actin gene was amplified with the following primers Actin1F (5′-TCCTTTCGCTCTATGCCAGT-3′), Actin1R (5′-GCCCAT CGGGAAGTTCATAG-3′) and used as genomic DNA/PCR positive control. The PCR conditions are identical to those described above with an annealing temperature of 58°C.

Immunocapture-PCR detection of encapsidated episomal BSV

BSVs were detected by immunocapture-PCR (IC-PCR) using specific BSOLV, BSGFV and BSIMV primers and a polyclonal antiserum raised against a cocktail of purified BSV species and SCBV species as described previously (Le Provost et al., 2006). To avoid contamination by plant genomic DNA,
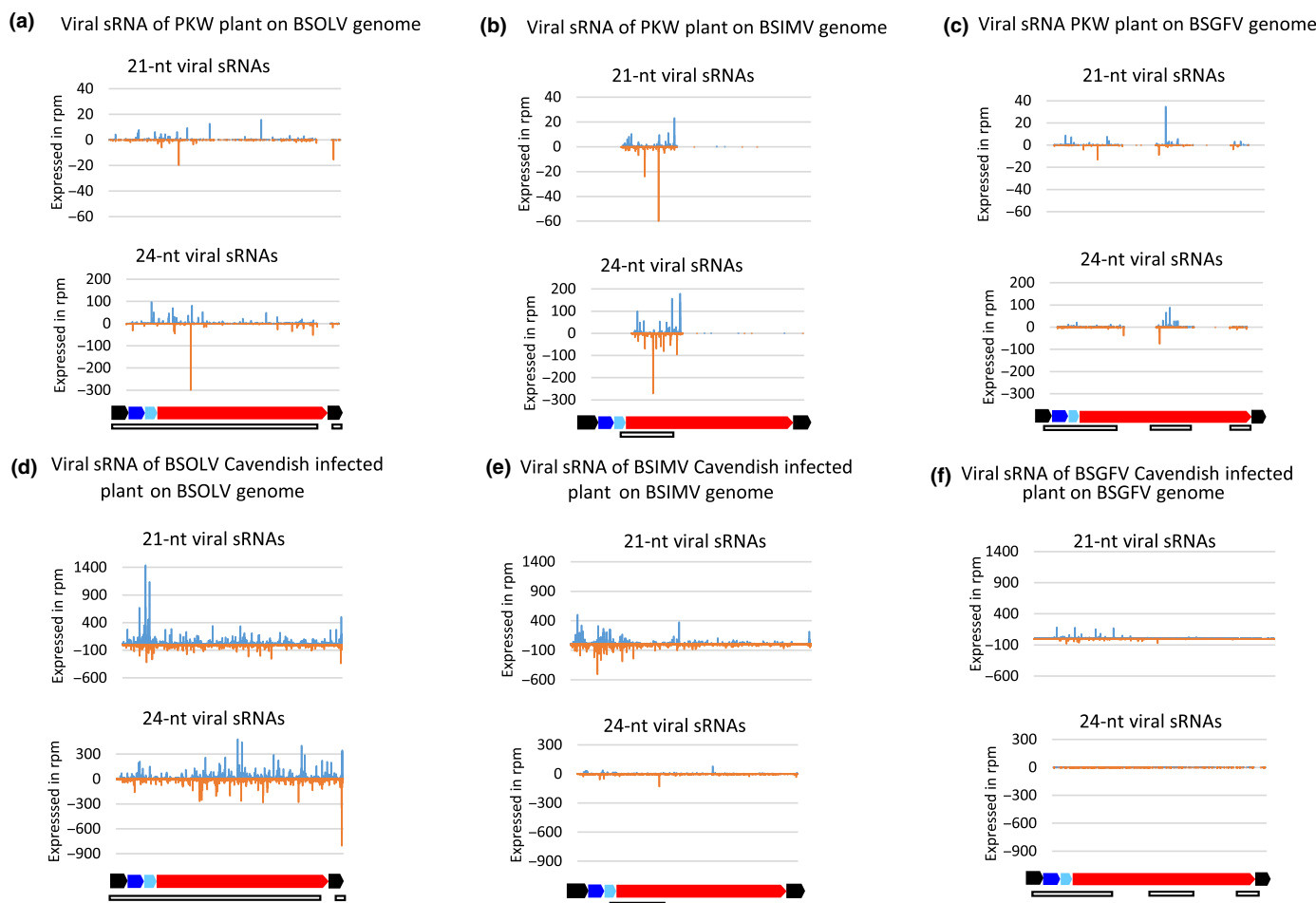
Fig. 3 Single-nucleotide resolution maps of 21-nt and 24-nt endogenous banana streak virus (eBSV)-derived small RNAs (sRNAs) from the seedy diploid *Musa balbisiana* PKW plant and BSV-derived sRNAs from triploid *Musa acuminata* plants infected with either BSOLV, BSGFV or BSIMV species on respective BSV genomes. The graphs plot the number of 21-nt and 24-nt viral sRNA reads with zero mismatches at each nucleotide position of the GenBank genome sequences of BSOLV (a, d), BSIMV (b, e) and BSGFV (c, f) extracted from PKW (a–c) or triploid *M. acuminata* BSV-infected plants (d–f). Bars above the axis represent sense reads starting at each respective position; those below the axis represent antisense reads ending at the respective position (for the map details, see Supporting Information Dataset S2). The genome organization of each BSV species is shown schematically below the graph, with the intergenic region, ORF1, ORF2 and ORF3 in black, dark blue, light blue and in red, respectively. The empty boxes underneath represent the regions of the episomal virus that are inverted/repeated in the corresponding integration (eBSV) in PKW plant. Data are expressed in reads per million (rpm).

samples were treated with RNase-Free DNAse I (Promega). DNAse mix (3 μl 10× buffer (400 mM Tris–HCl (pH 8.0 at 25°C), 100 mM MgSO₄, 10 mM CaCl₂); 3 μl DNAse I (1 U μl⁻¹) and 24 μl water) was added to coated tubes and incubated for 1 h at 37°C. The supernatant was removed and the tubes were washed once with water. DNAse I was inactivated by incubation at 95°C for 10 min (Chabannes *et al.*, 2021).

Results

eBSVs in the banana seedy diploid BB PKW and its self-pollinated progeny produce predominantly 24-nt siRNAs from inverted repeats

In the seedy *M. balbisiana* diploid genitor PKW, eBSV loci represent complex structures composed of a succession of

duplicated and/or inverted viral fragments. The di-allelic BSGFV and BSOLV integrations are present on Chromosome 1 with a single infective allele (eBSGFV-7 and eBSOLV-1, respectively) containing sequences comprising the entire viral genome, whereas the BSIMV integration is monoallelic, located in Chromosome 2 with two potentially infective alleles (Gayral *et al.*, 2008; Chabannes *et al.*, 2013). For eBSOLV-2, a deletion of 309 bp occurs at the junction of open reading frame (ORF)3 and the intergenic sequence (IG), thus precluding reconstitution of a functional virus (Chabannes *et al.*, 2013), whereas three null mutations present in ORFs of eBSGFV-9 (but not eBSGFV-7) lead to premature stop codons (Gayral *et al.*, 2008).

To explore whether the banana sRNA-generating silencing machinery targets eBSV loci, we deep-sequenced a fraction of 19- to 30-nt sRNAs from two independent PKW plants and nine plants representing self-pollinated PKW progeny lines.
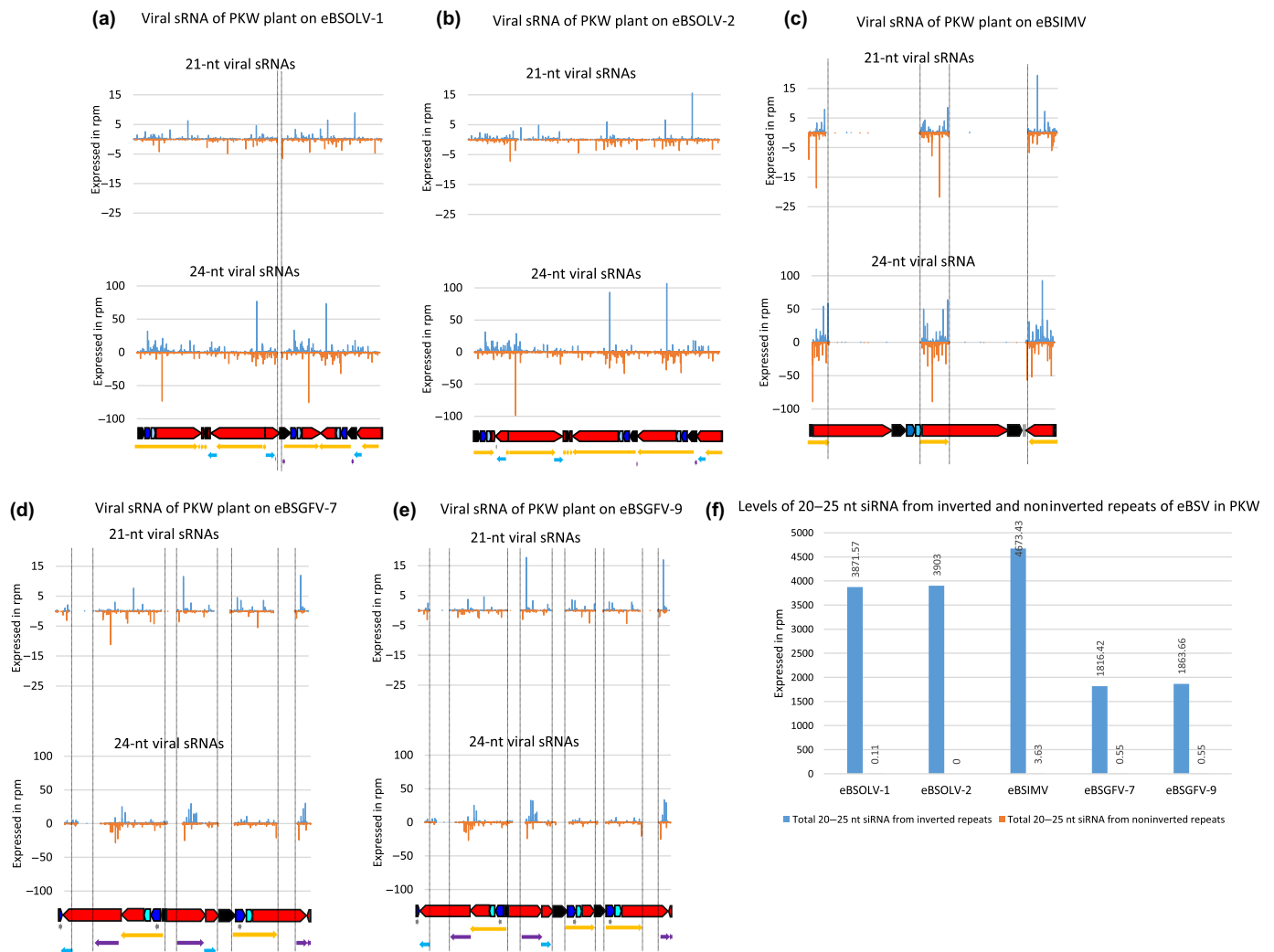
Fig. 4 Single-nucleotide resolution maps of 21-nt and 24-nt endogenous banana streak virus (eBSV)-derived small RNAs (sRNAs) from the seedy diploid *Musa balbisiana* PKW plant on respective eBSV genomes and levels of 20–25-nt eBSV-derived sRNAs from inverted and noninverted repeats. The graphs plot the number of 21-nt and 24-nt PKW eBSV-derived sRNA reads with zero mismatches at each nucleotide position of the GenBank genome sequences of eBSOLV alleles (a, b), eBSIMV (c) and eBSGFV alleles (d, e). Bars above the axis represent sense reads starting at each respective position; those below the axis represent antisense reads ending at the respective position (for map details, see Supporting Information Dataset S2). The genome organization of each eBSV species and alleles is shown schematically below the graph, with the intergenic region, ORF1, ORF2 and ORF3 in black, dark blue, light blue and red, respectively. The coloured arrows underneath represent inverted/repeated regions and vertical bars delineate unique or duplicated but not inverted regions. The levels of 20–25-nt sRNA derived from inverted and noninverted repeats of each of the five eBSV alleles are represented in (f). Data are expressed in reads per million (rpm).

Individuals from self-pollinated PKW progeny were selected on the basis of their eBSV allelic composition following genotyping with molecular markers (Gayral *et al.*, 2008; Chabannes *et al.*, 2013). As eBSIMV is monoallelic in PKW, all offspring have the same parental genotype for this eBSV. For eBSGFV and eBSOLV, the nine plants chosen represent all possible allelic combinations (homozygous or heterozygous integration) for eBSOLV-1, eBSOLV-2, eBSGFV-7 and eBSGFV-9. All selected plants were negative for BSV infection as checked by IC-PCR.

We restricted our bioinformatics analysis to sRNAs of 20–25 nt, the size range known to correspond to functional plant miRNAs and siRNAs. sRNAs were analysed by mapping reads to reference sequences of banana genomes A and B, circular viral genomes (BSOLV, BSGFV and BSIMV) and PKW BAC clones harbouring either eBSOLV-1, eBSOLV-2, eBSIMV, eBSGFV-7 or eBSGFV-9. sRNA mapping and counting results are detailed in Datasets S1 and S2 and summarized in Figs 2–4, 6–8, S2–S4.

In the two PKW plants, all three viral integrations were found to accumulate sRNAs, representing 0.2% or 0.37% (eBSGFV), 0.37% or 0.8% (eBSOLV) and 0.45% or 1.35% (eBSIMV), of the total 20 nt–25-nt sRNAs (Fig. 2a). In the nine self-pollinated PKW plants, total viral sRNA reads from the three eBSV loci combined
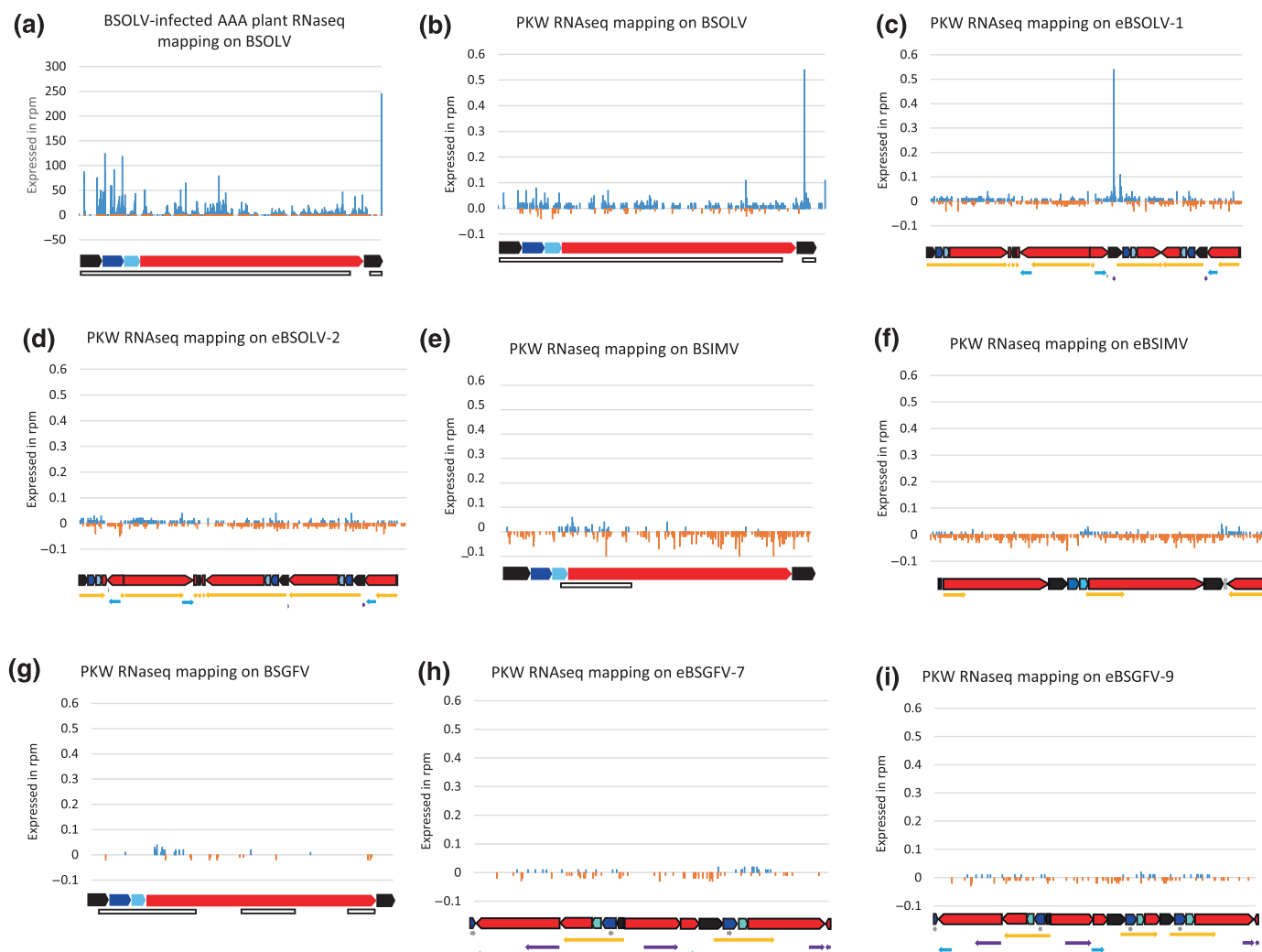
Fig. 5 Single-nucleotide resolution maps of endogenous banana streak virus (eBSV)-derived RNA-Seq data from the seedy diploid *Musa balbisiana* PKW plant and BSOLV-derived RNA-Seq data from a triploid *Musa acuminata* plant infected with BSOLV. The graphs plot the number of BSOLV-derived RNA-Seq reads from a triploid *M. acuminata* BSOLV-infected plant (a) and PKW eBSV-derived RNA-Seq reads with zero mismatches at each nucleotide position of the GenBank genome sequences of BSOLV (b), eBSOLV alleles (c, d), BSIMV (e), eBSIMV (f), BSGFV (g), eBSGFV alleles (h, i). Bars above the axis represent sense reads, those below the axis represent antisense reads, starting and ending at the respective positions (for map details, see Supporting Information Datasets S3 and S4). The genome organization of each eBSV species and alleles is shown schematically below the graph, with the intergenic region, ORF1, ORF2 and ORF3 in black, dark blue, light blue and red, respectively. The empty boxes or coloured arrows underneath BSV and eBSV sequences represent the inverted/repeated regions, respectively. Data are expressed in reads per million (rpm).

vary from 1% to 5% (Fig. 2a). Considering that eBSV sequences of *c.* 53 kb in total length per single B genome represent only 0.01% of the total size of the *M. balbisiana* (B) genome (*c.* 430 Mb), the accumulation of eBSV-derived sRNAs in PKW and its progeny is substantial. The quantities of viral sRNAs also vary from plant to plant but sRNA accumulation is almost always lower for eBSGFV and higher for eBSIMV; accumulation from eBSOLV is intermediate (Fig. 2a). The size class profiles of eBSV loci-derived sRNAs in PKW and its self-pollinated progeny plants are dominated by the 24-nt class, at *c.* 80% of the total 20–25-nt viral reads for each eBSV, followed by the 21-nt (*c.* 10%) and the 23-nt (*c.* 6–8%) classes, with the remaining three classes (20, 22 and 25 nt) being under-represented (Fig. 2b). The presence of 24-nt sRNAs derived from eBSVs was further validated by sRNA blot hybridization

analysis of a PKW plant and a *Musa acuminata* Cavendish (AAA) plant without eBSV as negative control using pools of four to five probes for eBSGFV, eBSOLV and eBSIMV corresponding to viral sRNA hotspot regions detected by Illumina sequencing: the results confirmed the predominant accumulation of 24-nt viral sRNAs in PKW (Fig. S1).

To compare the profiles of eBSV-derived sRNAs with those of the corresponding virus-derived sRNAs, we re-analysed previous sRNA sequencing data for Cavendish AAA plants infected independently with BSOLV, BSGFV or BSIMV (Rajeswaran *et al.*, 2014). Viral sRNAs matching the genomes of BSV species with zero mismatches constitute a significant fraction (1% for BSGFV, 8% for BSIMV and 22% for BSOLV) of the total 20–25-nt sRNAs in BSV-infected *M. acuminata* (AAA) plants
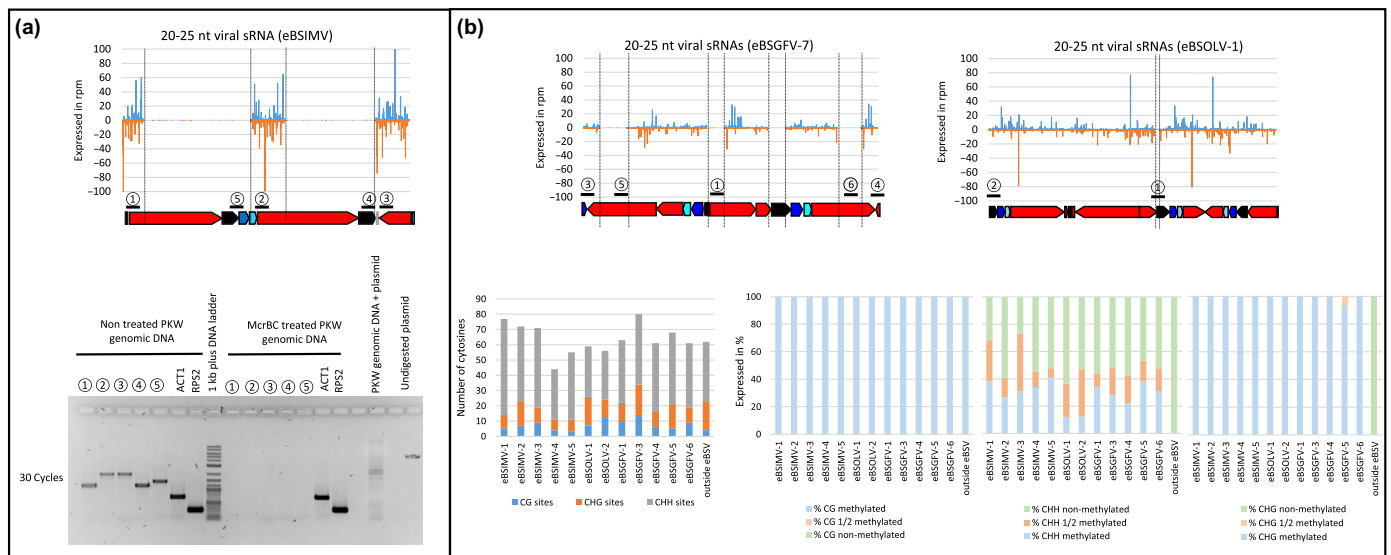
Fig. 6 Analysis of endogenous banana streak virus (eBSV) methylation status in the seedy diploid PKW plant by enzymatic and bisulfite sequencing. (a) Total genomic DNA from the seedy diploid *Musa balbisiana* PKW plant was digested with McrBC before PCR using eBSIMV primers targeting regions accumulating (or not) viral small RNA (sRNA) (upper panel) (see the Materials and Methods section). A 'nontreated' control sample was incubated in parallel in the same buffer but without enzyme. As a positive control for McrBC action, a single site-methylated plasmid DNA was spiked in total DNA from PKW plant releasing two bands for the plasmid and a smear for the genomic PKW DNA. Two housekeeping genes actin 1 (ACT1 – HQ85237) and ribosomal protein S2 (RPS2 – HQ853246) were used as control of McrBC treatment since those genomic regions are unmethylated. (b) Upper panel: Schematic representation of eBSGFV and eBSOLV regions (black bars with numbers) accumulating or not viral sRNA selected for bisulfite methylation analysis following PCR amplification (see the Materials and Methods section). Lower left panel: graph representing the number of cytosines within each PCR product for the three cytosine contexts in PKW. A PCR product 15 kb apart from eBSGFV in a gene coding region (named outside eBSV) was used as a negative control. Lower right panel: graphs illustrating the percentage of methylated, partially methylated or unmethytated cytosines in the three different contexts for every PCR fragment amplified from eBSIMV, eBSOLV and eBSGFV genomic DNA of PKW initially treated with sodium bisulfite.

(Rajeswaran *et al.*, 2014) (Fig. 2c). Conversely, size profiling of viral sRNAs from viruses shows the predominance of the 21-nt class for all three species, followed by the 22-nt class for BSGFV and BSIMV or the 24-nt class for BSOLV (Rajeswaran *et al.*, 2014) (Fig. 2d).

To analyse plant sRNAs, we used an updated version (Martin *et al.*, 2016) of *M. acuminata* Pahang's genome (D'Hont *et al.*, 2012) and the published *M. balbisiana* genome (Wang *et al.*, 2019). The size class profiles of nonviral banana sRNAs produced in PKW and the nine self-pollinated PKW plants are dominated by 21-nt sRNAs, followed by the 24 or the 22-nt classes, and then the 23- and 20-nt classes (Fig. S2). This differs from *M. acuminata* (A) genome-derived sRNAs in triploid AAA plants, which are dominated by 20-, 21- or 22-nt classes, followed by 24-nt and 23-nt classes in AAA plants (Rajeswaran *et al.*, 2014; Fig. S2).

Analysis of single-nucleotide-resolution maps of the three eBSV loci-derived sRNAs from PKW and its self-pollinated progeny mapped on the reference sequences of the three BSVs and the known alleles of eBSOLV (eBSOLV-1 and eBSOLV-2), eBSGFV (eBSGFV-7 and eBSGFV-9) and eBSIMV revealed that viral siRNAs do not cover the entire episomal viral genomes (Fig. 3a–c); they do not target a common viral zone of the BSV genome but are derived almost exclusively from the inverted repeat sequences present in all eBSOLV, eBSGFV and eBSIMV loci (Fig. 4). This is the case for the two major size classes, that is the predominant 24-nt and the minor 21-nt class, although their

hotspot profiles differ within the inverted repeat regions (Fig. 4). In eBSV loci, the inverted repeats represent most of the BSOLV sequences except for the 369 bp region at the end of ORF3 and the beginning of the IG, three regions of the BSGFV sequence and one region of the BSIMV sequence (Figs 3, 4). Confirming earlier findings (Rajeswaran *et al.*, 2014), our re-analysis of single-nucleotide-resolution maps of BSV-derived sRNAs revealed that 20-to-25-nt sRNA species cover the entire circular virus genome without gaps, with characteristic hotspots for the three major size classes on both strands of the virus genome (Fig 3d–f; Dataset S2).

Inverted repeat-derived viral siRNAs are generated from both infective and noninfective eBSV alleles

To determine whether viral siRNAs are produced from infective and/or noninfective eBSVs in PKW (Gayral *et al.*, 2008; Chabannes *et al.*, 2013), we analysed sRNAs from the PKW self-pollinated progeny; both infective and noninfective alleles of both eBSGFV and eBSOLV produce viral siRNAs from the inverted repeats. The presence of either noninfective alleles or the infective alleles in homozygous state (for either eBSOLV or eBSGFV, alone or together) does not appear to influence the amount of viral sRNAs (Fig. 2a).

Analysis of single-nucleotide-resolution maps of total 20–25-nt-long eBSV-derived sRNAs revealed very similar profiles in PKW and its progeny lines, with a predominance of the 24-nt

class, although quantities and hotspot profiles vary slightly between plants (Fig. S3). In all the plants, accumulation of viral siRNAs is restricted nearly exclusively to inverted and repeated regions in the eBSVs. According to Chabannes *et al.* (2013), a noninverted and nonrepeated region in the eBSOLV-1 allele spans positions 6672–7040 of the respective virus, whereas eBSOLV-2 presents noninverted and nonrepeated regions at nucleotide positions 6672–6732 and 7042–7183, and a repeated but not inverted region at position 7184–7264; the sequence between 6733 and 7040 is missing in eBSOLV-2. The mapping of total 20–25-nt sRNAs from homozygous eBSOLV-1 plants and homozygous eBSOLV-2 plants on nucleotides 6672–7264 shows that sRNAs are accumulated only between positions 7040 and 7264 of homozygous eBSOLV-1 plants, thus confirming that viral sRNA accumulation is restricted exclusively to the inverted repeats of eBSOLV sequences (Fig. S4). Such differences in viral siRNA (vsiRNA) patterns in PKW self-pollinated offsprings could not be differentiated for eBSIMV as both alleles are identical. The extra 2.3-kb fragment present in the noninfective eBSGFV-9 allele is repeated but not inverted (Gayral *et al.*, 2008), thus precluding differentiation between the two alleles.

eBSVs are weakly transcribed in PKW

To provide evidence that eBSV loci in PKW are transcribed to create precursors of sRNAs, we performed a transcriptomic analysis using Illumina strand-specific library preparation and sequencing protocols. We sequenced total (polyadenylated and nonpolyadenylated) RNA molecules from both the PKW plant and the BSOLV-infected Cavendish (AAA) plant, yielding 120 million and 168 million 125-nt sequencing reads, respectively. The reads were then mapped with zero mismatches to the reference sequences of virus genomes (BSOLV, BSGFV and BSIMV) and the corresponding eBSVs (eBSOLV, eBSGFV and eBSIMV). The results show that, as expected, abundant positive-sense transcripts cover the entire viral genome in BSOLV-infected AAA plant, corresponding to the terminally redundant viral pregenomic RNA (pgRNA) (Fig. 5a). Only 1.4% of viral reads represent the antisense strand of the BSOLV genome, confirming monodirectional transcription of circular double-stranded viral DNA by Pol II, producing pgRNA and also revealing antisense transcripts likely representing antisense strands of dsRNA precursors of viral siRNAs. By contrast, the number of reads from PKW mapped to the BSOLV genome was 1000 times lower than that from the BSOLV-infected AAA plant (Dataset S3). These low-abundance BSOLV-specific reads from the PKW plant have a different hotspot profile compared with those from the BSOLV-infected Cavendish plant (Fig. 5a–d). Analysis of the single-nucleotide resolution map of the BAC clone containing eBSOLV-1 did not reveal reads spanning both junctions of the viral integrant (eBSV positions 53 586–76 442) in the forward orientation, while some reverse reads are present at the downstream junction up to position 77 397 of the BAC clone. The TATAbox (TATATAA) in antisense strand (77 460–77 454) or the 10 repeats of TATATA at positions 77

525–77 584 could serve as a promoter driving transcription in the reverse orientation (Dataset S4) with the resulting transcript(s) containing inverted repeats creating prefect dsRNA precursors of viral siRNAs.

Regarding eBSIMV transcripts, most RNA-Seq reads from PKW represent the antisense strand of the virus genome (87%) and are distributed all along this strand without prominent hotspots (Fig. 5e,f). By contrast, the less abundant viral sense reads were restricted mostly to the 1914-nt region of the virus genome, which is inverted and repeated in the monoallelic eBSIMV integrant. This finding suggests that eBSIMV is transcribed mainly monodirectionally in the antisense orientation (with respect to the BSIMV genome), and the inverted repeat sequences of these antisense transcript(s) create perfect dsRNA precursors of viral siRNAs. Analysis of the single-nucleotide resolution map of the BAC containing eBSIMV (eBSV 91 102–106 919) highlights the presence of reverse reads spanning the downstream junction of the integration up to position 107 989 of the BAC clone with a TATAbox (TATATAA) in antisense strand at 108 020–108 026, which could be responsible for transcription of eBSIMV (Dataset S4).

For eBSGFV transcripts, RNA-Seq reads from PKW are of very low abundance and represent both sense and antisense strands of the corresponding virus genome (Fig. 5g). Hotspot profiles of the viral reads mapped to eBSGFV-7 and eBSGFV-9 loci show that the inverted repeats are enriched in sense reads, while antisense reads are distributed more evenly along both reference sequences (Fig 5h,i). The antisense reads represent 69% of the total reads for eBSGFV-7 and eBSGFV-9. Thus, both eBSGFV alleles appear to be transcribed predominantly in a monodirectional manner in 'antisense' orientation with respect to BSGFV. These results are reinforced by the fact that no transcripts span the upstream junction of eBSV, as evidenced by analysis of the single-nucleotide resolution map of the BAC clone containing eBSGFV-7 (Dataset S4). Conversely, reads spanning the downstream junction of the integration (eBSV located between positions 80 691 and 93 970) up to position 97 450 are present with a TATAbox (TATATAT) in antisense strand at position 97 463–97 469, which could be responsible for transcription of eBSGFV (Dataset S4).

All these data suggest that eBSVs are transcribed very poorly/weakly, and mainly monodirectionally, in PKW.

eBSVs are heavily methylated in PKW

To address whether inverted repeat-derived siRNAs direct cytosine methylation of the corresponding eBSV sequences integrated in the PKW genome, we exploited the methylation-dependent endonuclease McrBC that cleaves double-stranded DNA containing methylcytosine (mC) on one or both strands.

PKW genomic DNA (McrBC-treated and nontreated) was subjected to PCR with five different primer pairs designed to amplify sequences in the eBSIMV locus. Three of the PCR amplicons are located in the inverted repeat regions accumulating siRNAs (Regions 1–3; Fig. 6a) and two are in regions that do not accumulate any siRNAs (Regions 4 and 5; Fig. 6a).
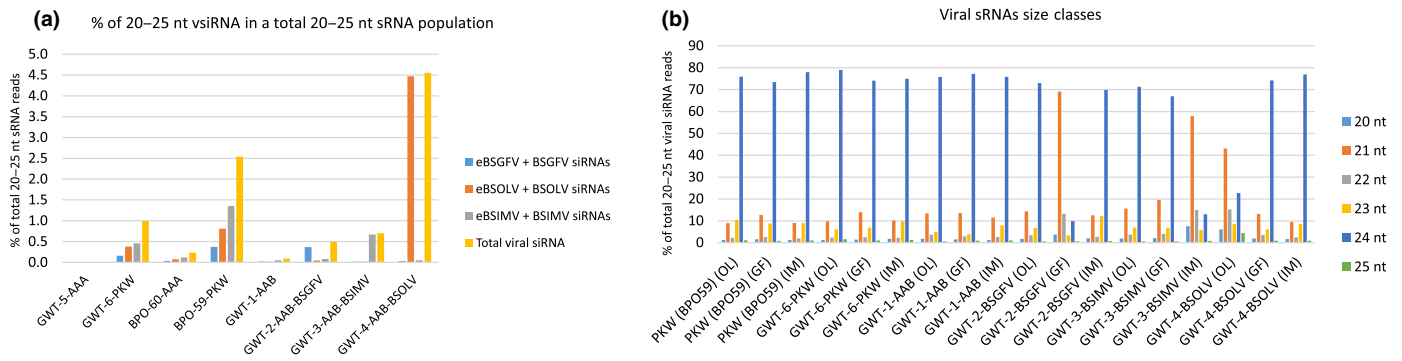
(a) % of 20–25 nt vsiRNA in a total 20–25 nt sRNA population

(b) Viral sRNAs size classes

Fig. 7 Characterization of viral small RNAs (sRNAs) in the seedy diploid PKW plant, triploid interspecific AAB noninfected and BSOLV-, BSGFV- or BSIMV-infected plants and triploid *Musa acuminata* healthy control plants. The graphs show the percentages of 20–25-nt endogenous banana streak virus (eBSV) and BSV-derived sRNAs in the pool of total (host and viral) 20–25-nt reads mapped to the respective virus genomes (eBSOLV, eBSGFV, eBSIMV, BSOLV, BSGFV or BSIMV) with zero mismatches (a) as well as the percentages of each size class of 20–25-nt viral sRNA reads mapped with zero mismatches to each of the BSV species genomes (BSOLV-OL, BSGFV-GF and BSIMV-IM), which have eBSV counterparts present in the banana genome (b). PKW (BPO59 or GWT-6) is homozygous for eBSIMV, heterozygous for eBSGFV with eBSGFV-7 (GF7) infective and eBSGFV-9 (GF9) noninfective alleles and for eBSOLV with eBSOLV-1 (OL1) infective and eBSOLV-2 (OL2) noninfective alleles (24, 25). The four allotriploid AAB hybrids were obtained from interspecific hybridization between a virus-free autotetraploid *M. acuminata* (IDN110 Tetra, AAAA) and the diploid PKW female parent. eBSV genotypes for the AAB interspecific hybrids are as follows: GWT-1 (OL2; GF9); GWT-2 (OL2; GF7); GWT-3 (OL2; GF9); GWT-4 (OL1; GF9); all have one copy of eBSIMV. IC-PCR experiments have shown that GWT-2 was infected by BSGFV, GWT-3 by BSIMV and GWT-4 by BSOLV while GWT-1 was BSV free.

The methylated plasmid spiked into PKW genomic DNA was fully digested with McrBC, yielding two fragments as expected. In the absence of McrBC treatment, all five eBSIMV regions were amplified, yielding PCR products of the expected sizes. Instead, no PCR product was detected when PKW genomic DNA was treated with McrBC before PCR, indicating that all five regions of eBSIMV are methylated, regardless of whether they are a source of siRNAs (Fig. 6a).

We further characterized eBSV methylation status in PKW by treating plant genomic DNA with sodium bisulfite followed by PCR amplification and Sanger sequencing. Each PCR fragment was sequenced in bulk to better represent the overall methylation status of the region. For eBSIMV, we selected the same five regions analysed previously by McrBC treatment. For eBSGFV, we selected two unique regions accumulating siRNAs and three unique regions that do not accumulate detectable sRNAs. For eBSOLV, we selected one unique region covered by siRNAs and one unique uncovered region (Fig. 6b). As a negative control, we analysed a region of the gene encoding the Zonadhesin-related protein located only 14 kb from the eBSGFV integration site. The number of cytosines and the different contexts (CG, CHH and CHG) per PCR amplicon are illustrated in Figs 6(b), S5.

Direct sequencing of PCR products (Fig. 6b) revealed that 100% of the cytosine residues in symmetric contexts CG and CHG (where H represents A, T or C) were methylated in each of the selected regions for each eBSV, with the exception of region 4 of eBSGFV (94% fully methylated and 6% partially methylated CHG sites). By contrast, 30–70% of asymmetric CHH sites were found to be nonmethylated in all selected regions of eBSV loci, while the remaining CHH sites were fully or partially methylated.

The selected region of the Zonadhesin-related protein-coding gene located outside eBSGFV sequence was 100% methylated in

the four CG sites but not methylated at all in the CHG and CHH contexts (Figs 6b, S5). This gene spans few siRNAs along its sequence (data not shown), suggesting that this gene is a so-called gene bodies whose expression is linked to the presence of CG methylation in their coding region (Zilberman *et al.*, 2007; Bewick & Schmitz, 2017).

AAB offspring produce low-abundance siRNAs from eBSV inverted repeats

To investigate the effect of *M. balbisiana* genome (B) ploidy on the regulation of eBSVs, and to understand why, in most interspecific hybrids, including AAB, AB, AAAB, infective eBSVs are able to induce viral infection, we characterized triploid AAB hybrids from a population obtained by an interspecific genetic cross between PKW and IDN 110 tetraploid *M. acuminata* (AAAA).

Using eBSV molecular markers developed previously (Gayral *et al.*, 2008; Chabannes *et al.*, 2013), we genotyped the triploid progeny plants and selected four plants harbouring the eBSIMV allele. Plants GWT-1 and GWT-3 also have the noninfective eBSOLV-2 and noninfective eBSGFV-9 alleles, plant GWT-2 the noninfective eBSOLV-2 and infective eBSGFV-7 alleles, and plant GWT-4 the infective eBSOLV-1 and noninfective eBSGFV-9 alleles. Plant phenotyping by IC-PCR demonstrated that GWT-1 was virus-free, while GWT-2, GWT-3 and GWT-4 were infected by BSGFV, BSIMV and BSOLV, respectively (data not shown), following activation of infective eBSV alleles. sRNA sequencing analysis revealed that GWT-1 accumulated small amounts of viral sRNAs from all three eBSV loci, constituting only 0.09% of the total 20–25-nt sRNA reads, with 0.02% representing eBSGFV, 0.05% eBSIMV and 0.02% eBSOLV loci (Fig. 7a). This is *c.* 10 times and 30 times lower than for the corresponding eBSV loci in the two PKW plants, GWT-6 and
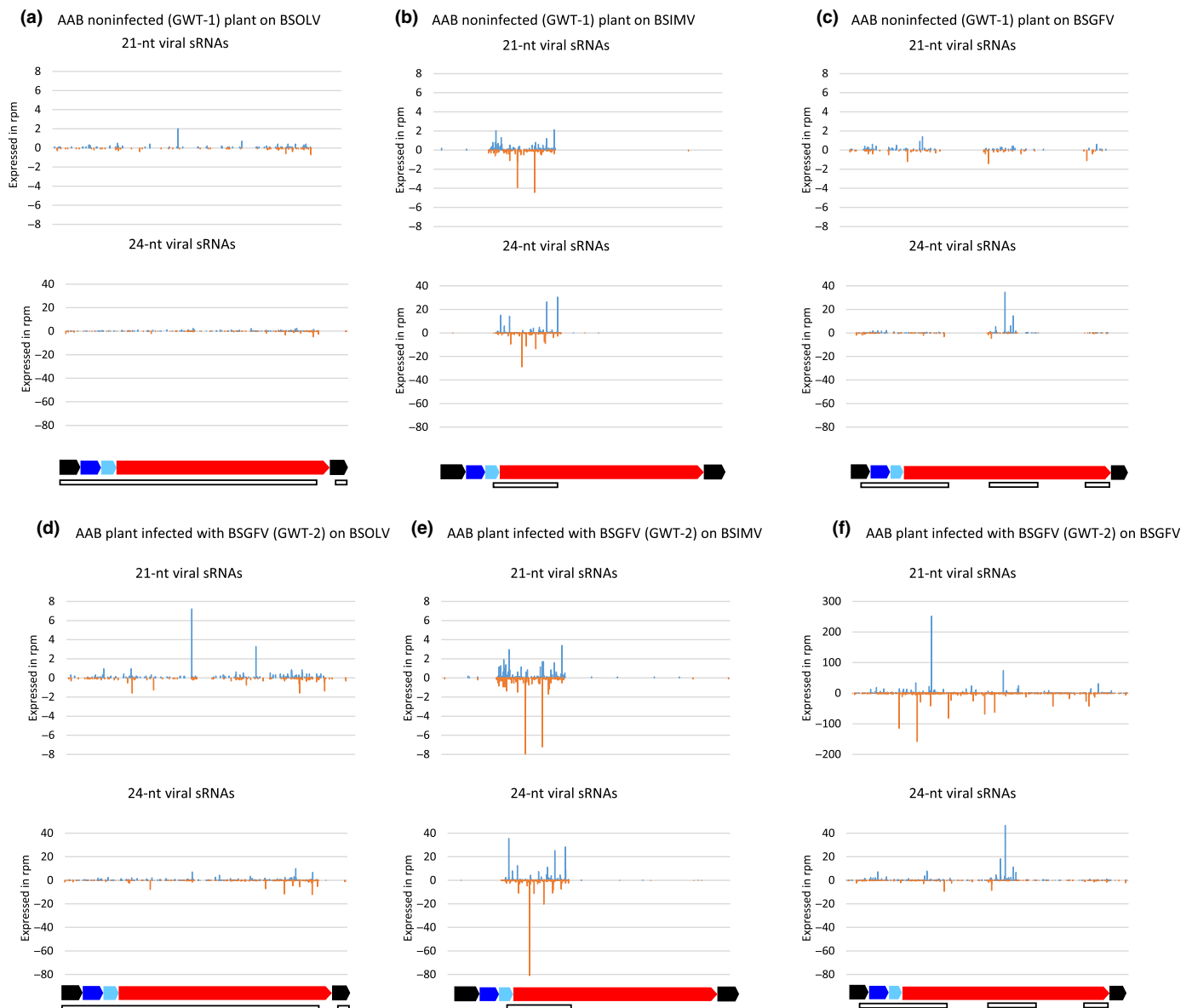
Fig. 8 Single-nucleotide resolution maps of 21-nt and 24-nt endogenous banana streak virus (eBSV) and BSV-derived small RNAs (sRNAs) from triploid interspecific AAB noninfected and BSOLV-, BSGFV- or BSIMV-infected plants. The graphs plot the number of 21-nt and 24-nt viral (eBSV and BSV) sRNA reads with zero mismatches at each nucleotide position of the GenBank genome sequences of BSOLV (a, d, g, j), BSIMV (b, e, h, k) and BSGFV (c, f, i, l) extracted from the noninfected AAB interspecific hybrid (a–c) or BSGFV (d–f), BSIMV (g–i) and BSOLV (j–l) infected AAB interspecific hybrids, respectively. Bars above and below the axis represent sense reads starting and ending at the respective positions (for map details, see Supporting Information Dataset S2). The genome organization of each BSV species is shown schematically below the graph, with the intergenic region, ORF1, ORF2 and ORF3 in black, dark blue, light blue and red, respectively. The empty boxes underneath represent regions of episomal virus that are inverted/repeated in the corresponding integration (eBSV) in PKW plant. Data are expressed in reads per million (rpm).

BPO-59, respectively (Fig. 7a) and 11 times (GWT-12) to 57 times (GWT-14) lower than the nine self-pollinated PKW off-spring (Figs 2a, 7a). All these data suggest that the reduction in eBSV copies results in reduced transcription and subsequently in the reduction of virus-derived sRNAs.

For GWT-2, GWT-3 and GWT-4 plants, viral sRNAs represented 0.5, 0.7 and 4.5% of the total 20–25-nt sRNAs, respectively, but, in each case, most of these reads are derived from the respective virus (BSGFV, BSIMV or BSOLV). The eBSV-derived sRNAs in GWT-2 (eBSIMV and eBSOLV), GWT-3 (eBSGFV and eBSOLV) and GWT-4 (eBSGFV and eBSIMV) plants accumulate at levels similar to those of eBSV-derived sRNAs in the GWT-1 plant (Fig. 7a).

The size profile of viral sRNAs in the noninfected AAB hybrid plant GWT-1 is similar to that of PKW and its self-pollinated progeny plants, with 24-nt sRNAs being the most abundant (c. 75%) for each eBSV integrant (Fig. 7b). In virus-infected AAB hybrid plants, the size profiles of the respective
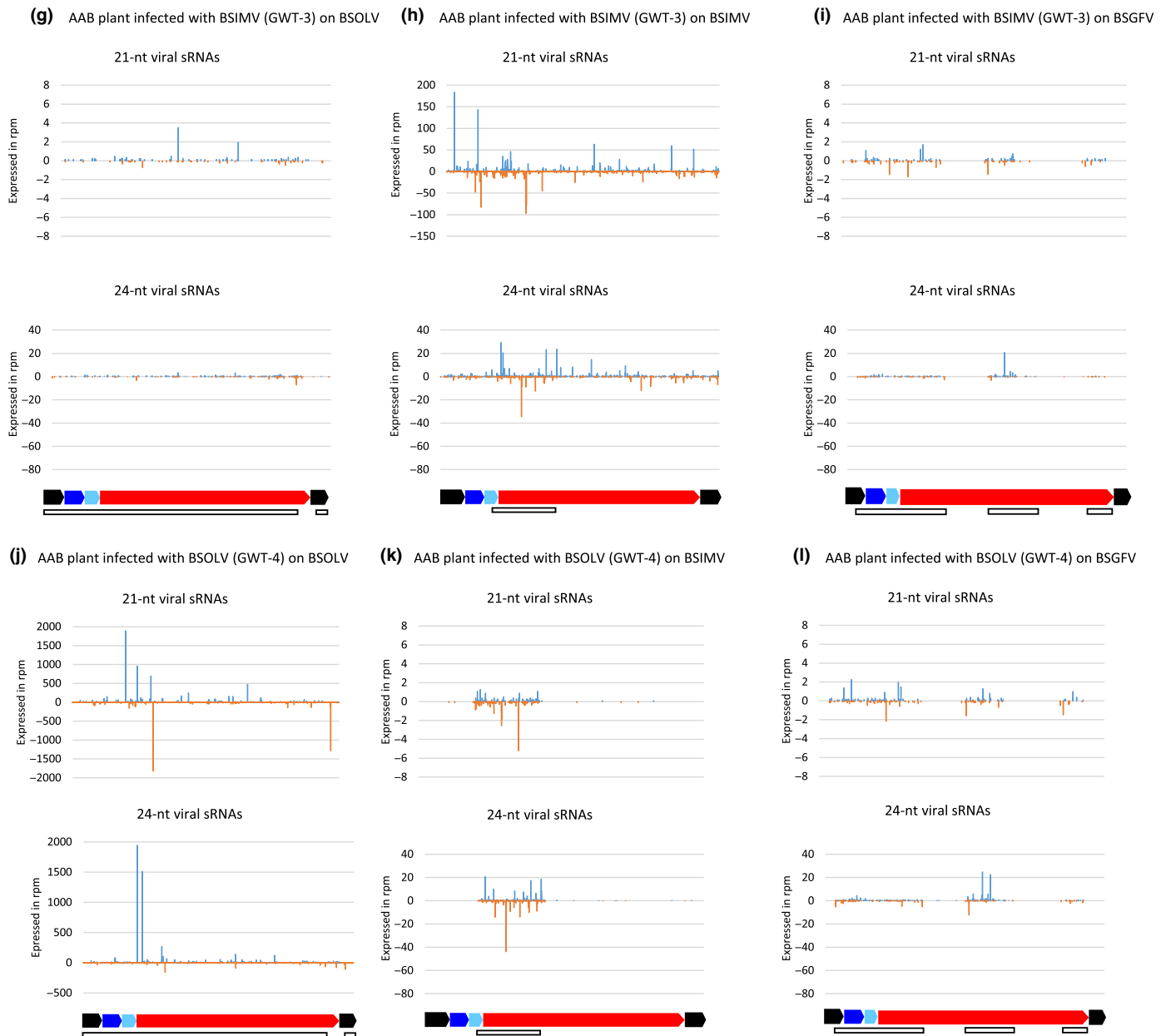
(g) AAB plant infected with BSIMV (GWT-3) on BSOLV
(h) AAB plant infected with BSIMV (GWT-3) on BSIMV
(i) AAB plant infected with BSIMV (GWT-3) on BSGFV

(j) AAB plant infected with BSOLV (GWT-4) on BSOLV
(k) AAB plant infected with BSOLV (GWT-4) on BSIMV
(l) AAB plant infected with BSOLV (GWT-4) on BSGFV

Fig. 8 Continued.

virus-derived siRNAs are similar to those obtained for virus-infected Cavendish (AAA) plants, where the 21-nt class is predominant (Rajeswaran *et al.*, 2014) (Figs 2d, 7b). The eBSV-derived sRNAs in GWT-2 (eBSIMV and eBSOLV), GWT-3 (eBSGFV and eBSOLV) and GWT-4 (eBSGFV and eBSIMV) plants have size profiles similar to those in the virus-free GWT-1 and PKW plants. These results suggest that PTGS directed by 21-nt siRNAs is induced in AAB interspecific hybrids upon virus infection, similar to that induced by corresponding viruses in AAA plants (Fig. 2d).

Analysis of single-nucleotide-resolution maps of eBSV-derived sRNAs (Fig. 8) revealed that, like PKW and its self-pollinated progeny, viral siRNA accumulation is restricted to the inverted repeats of the three eBSV loci in the virus-free GWT-1 plant (Fig. 8a–c). This is also the case for two of the three eBSV loci in GWT-2 (eBSIMV and eBSOLV), GWT-3 (eBSGFV and eBSOLV) and GWT-4 (eBSGFV and eBSIMV) plants (Fig. 8d, e,g,i,k,l). By contrast, the entire virus genome is covered with viral BSGFV-derived siRNAs in GWT-2, BSIMV-derived siRNA in GWT-3 and BSOLV-derived siRNA in GWT-4 plants (Fig. 8f,h,j). The hotspot profiles for each virus appear as a mixture of the virus sRNA hotspots (evident from the respective virus-infected AAA plants) (Fig. 3d–f) and the eBSV inverted repeat sRNA hotspots (evident from GWT-1 plant (Fig. 8a–c) and two of the three GWT-2 to GWT-4 plants lacking the respective virus).

Discussion

eBSVs are regulated by RdDM and TGS

The presence of infective eBSVs in *M. balbisiana* genomes is a strong hindrance to banana breeding programmes world-wide as many interspecific hybrids are found to be infected with BSV released from such eBSVs (Thomas *et al.*, 2015). It is also a constraint for germplasm exchange, as this represents a potential avenue for the concomitant transfer of contaminating viruses (Thomas *et al.*, 2015). Interestingly, to date, all seedy diploid BB banana have been shown to be healthy carriers of infective eBSV; here, we sought to understand the basis of this natural resistance.

Analysis of total RNA molecules from both PKW and BSOLV-infected triploid AAA plants showed that eBSOLV is *c.* 1000 times less transcribed than the virus in an infected plant (Fig. 5). Although the level of transcripts is very low, they are distributed along almost the entire eBSV sequence for the three infective species studied and seem to be produced monodirectionally in an antisense orientation (Fig. 5). We speculate that this transcription may be driven by plant Pol II promoters located adjacent to integrated viral sequences. Our initial RNA-Seq experiment targeted polyadenylated and nonpolyadenylated transcripts > 125 nts, but subsequent deep sRNA sequencing revealed that all three eBSV loci produce predominantly 24-nt siRNAs, probably driven by Pol IV (Blevins *et al.*, 2015; Zhai *et al.*, 2015), and much less abundant siRNAs of other size classes. This suggests that RdDM and TGS operate at these loci, whereas Pol II transcription of eBSVs is negligible. Our analysis of single-nucleotide-resolution maps of eBSV sRNAs revealed that they are derived almost exclusively from inverted and repeated regions of eBSVs (Figs 4, S3, S4), echoing results from (da Fonseca *et al.*, 2016), who showed that only the inverted repeats of an integrated cucumber mosaic virus (CMV) RNA1 sequence in soybean generate viral siRNA.

Analysis of viral sRNA profiles of eBSOLV-1 and eBSOLV-2 homozygous self-pollinated PKW, confirmed that eBSV-derived viral sRNA accumulation in these newly created PKW plants follows the same mechanism as in the wild parental PKW plant (Fig. S4). Indeed, this mechanism always targets inverted and repeated regions.

Bisulfite sequencing revealed that eBSV sequences are highly methylated in all cytosine contexts, with 100% methylated symmetric CG and CHG sites and up to 50% of methylation in nonsymmetric CHH sites – a hallmark of RdDM and TGS. This silencing mechanism of *Musa balbisiana* would be analogous to that in *A. thaliana*, where DCL3-dependent RdDM generates 24-nt siRNAs and plays a primary role in transcriptional silencing at heterochromatic loci and in genome defence against foreign DNA such as TEs (Castel & Martienssen, 2013; Matzke & Mosher, 2014). Integrated sequences could be transcribed initially by Pol II, producing low levels of relatively long transcripts; these transcripts fold to create perfect dsRNA precursors from the inverted repeat area, forming dsRNA precursors of siRNA processed by DCL2, DCL3 and DCL4 enzymes. The 23–24-nt eBSV-derived Pol II siRNA and, to a lesser extent, 21–22-nt siRNAs (McCue

et al., 2015; Zhao *et al.*, 2016) will trigger low levels of DNA methylation at eBSV loci, with subsequent transcription by Pol IV initiating the canonical RdDM pathway as described for transcriptionally active LTR transposons (Borges & Martienssen, 2015). The Pol IV single-stranded RNA copied into dsRNA by RDR2 is processed by DCL3 to produce 24-nt-long siRNAs. As established in *A. thaliana*, once incorporated into the Argonaute protein AGO4 with a Pol V scaffold, these transcripts will result in DRM2 recruitment to reinforce methylation at cognate eBSV loci in all cytosine contexts (Chan *et al.*, 2004; Zhong *et al.*, 2014), leading to the establishment and maintenance of epigenetic silencing of those loci. The *in silico* genome-wide identification and characterization of DCL, AGO and RDR genes in bananas by (Ahmed *et al.*, 2021) and the presence of Musa genes analogous to the Arabidopsis Pol IV and Pol V in the Phytozome database (https://phytozome.jgi.doe.gov/pz/portal.html) by using the completed genome assembly (v1) of *M. acuminata* (data not shown) reinforce the idea that eBSV would be regulated by the TGS/PolIV pathway. Interestingly, the noninverted/repeated regions were also strongly methylated in all three cytosine contexts, whereas these regions are devoid of siRNA. These results are similar to those described by Ahmed *et al.* (2011), who found evidence at a genome-wide scale in *A. thaliana* that a significant fraction of TEs were densely methylated at CG, CHG and CHH sites in the absence of matching siRNAs. These latter authors suggest that those TE are unlikely to be targeted by the RdDM machinery but that DNA methylation spreads from adjacent RdDM targeted sites. In our model, the inverted/repeated areas would thus serve as an initial site for methylation of eBSV via RdDM, which then extends to the noninverted/repeated area of eBSV to reinforce the silent state of those integrations by an as yet unknown mechanism. In PKW, the amount of vsiRNA produced and the methylation of eBSVs likely allow this plant to be a healthy carrier of infective integrations and be naturally resistant to cognate episomal viruses (to be described later).

Where BSV persistently infects *M. acuminata* banana plants, the involvement of both PTGS and TGS pathways has been demonstrated. However, despite the presence of 24-nt viral siRNAs, the episomal viral DNA was largely free of cytosine methylation (Rajeswaran *et al.*, 2014) and thus able to evade siRNA-directed DNA methylation to avoid TGS (Pooggin, 2013). In this study, we found that, upon release of BSV from the B genome of AAB hybrids, the PTGS pathway generating abundant 21–22-nt viral siRNAs is induced from active transcripts produced from the BSV. Taken together, our data show that both PTGS and TGS are involved in repression of eBSV loci. The components/proteins that mediate both PTGS and RdDM/TGS in *M. acuminata* and *M. balbisiana* genomes remain to be functionally characterized and compared.

Are eBSV-derived siRNAs responsible for the natural resistance of seedy BB diploid banana to episomal BSV infection?

The three BSV species integrated sequentially into the PKW genome, probably as concatenated viral genomes, several hundred thousand years ago (Gayral *et al.*, 2010; Chabannes *et al.*, 2013).

However, before genome shuffling generated the current molecular structures, these ancient neo-integrations had to be very quickly tamed by plant defence systems to prevent the large-scale production of infectious viral particles. When a pararetrovirus integrates as tandem repeats, the terminally redundant genome-length pgRNA produced by direct Pol II transcription can give rise to episomal virus infection, as shown for endogenous petunia vein clearing virus (ePVCV) (Richert-Poggeler *et al.*, 2003). Initial integrations of BSV sequences may also have occurred in tandem repeats, subsequently evolving through unequal recombination, point mutations and insertions, leading to the present-day structures of eBSV loci in PKW (Chabannes *et al.*, 2013).

Cumulative amounts of total eBSV-derived vsRNA in the healthy interspecific AAB hybrid (IDN110-T × PKW, GWT-1) plant were between 10 and 57 times lower than in PKW and its self-pollinated progeny plants (Figs 2a, 7a), whereas the eBSV dosage in this hybrid is only three times lower. When the values obtained for the four AAB hybrids (ignoring eBSV values where the virus has awakened) are averaged over the 11 BB diploid plants studied in this work, there is a 20-fold decrease in total eBSV-derived vsRNA (Figs 2a, 7a, S6). Thus, a synergistic effect seems to occur when two *M. balbisiana* genomes or two alleles of eBSV combine to produce abundant viral sRNAs, which may contribute to the repression of eBSVs, even when these plants are subjected to stresses known to activate eBSV. By contrast, in hybrids harbouring a single copy of each eBSV, the repressed state of each eBSV locus would not be maintained strongly enough under stress conditions, triggering the release of episomal BSV. Once released, active transcription from the episomal virus will activate PTGS.

The genomic environment in the *c.* 50 kb surrounding the three eBSV loci does not seem to influence viral siRNA accumulation; eBSGFV and eBSIMV, both in protein-coding gene-rich regions (Chabannes *et al.*, 2013), produce the lowest and highest quantity of viral siRNAs in PKW (Fig. 2a). However, the presence of TEs close to all three eBSV loci may have served as an initial signal to the silencing machinery to control eBSV when initial integration in tandem repeats took place. The higher accumulation of vsiRNA from eBSIMV could be related to the fact that this integration is the least rearranged, probably requiring increased surveillance to avoid virus emergence by direct Pol II transcription as discussed above, unlike eBSOLV and eBSGFV, which require at least one additional recombination step to reform a fully linear viral genome (Gayral, 2008; Iskra-Caruana *et al.*, 2010; Chabannes & Iskra-Caruana, 2013). How recent viral integrants with no homology to host genomes are detected and silenced *de novo* is still a black box, and it remains to be investigated whether the mechanism resembles that described during the reconstruction of *de novo* silencing of an active plant retrotransposon (Mari-Ordonez *et al.*, 2013).

The maintenance of 21-nt (*c.* 10% of total vsRNA) and 22-nt (*c.* 2–4%) long viral siRNA accumulation is probably responsible for the natural resistance of PKW to mealybug-mediated infections by guiding the *RNA-induced silencing complex (*RISC) to the complementary sequence of the viral RNA, where AGO1 will catalyse cleavage and/or translational repression of the target RNA resulting in PTGS. Similarly, Valli *et al.* (2023) suggest that pararetroviral-derived endogenous siRNAs, enriched in 22-nt forms in tomato constitute an adaptative heritable record of past experiences potentially targeting all forms of plant pararetroviruses. The abundant 24-nt eBSV-derived siRNAs could also potentially guide AGO4-RISC to episomal viral DNA, interfering with its transcription in the nucleus at the earliest stages of virus infection allowing PKW to react and set up its defence system immediately after intrusion of a cognate BSV to prevent its replication. Conversely, banana plants lacking eBSV sequences and pre-existing eBSV-derived siRNAs, such as *M. acuminata* Cavendish, are susceptible to episomal infection. Indeed, despite induction of both PTGS and TGS upon virus infection, the virus is able to escape these defence mechanisms and continue to multiply its circular double-stranded DNA, evading cytosine methylation and TGS in the nucleus (Rajeswaran *et al.*, 2014). Plants and pathogens are in a constant race, and differences in defence timing and intensity influence disease outcomes (Kondratev *et al.*, 2020).

In summary, our data suggest that eBSV-derived siRNAs can maintain the RdDM (24-nt siRNA) at cognate loci to repress infective eBSV copies, preventing the release of episomal transmissible virus and simultaneously targeting any cognate (incoming) mealybug-transmitted virus in a sequence-specific manner at both transcriptional (24-nt siRNA) and post-transcriptional (21–22-nt siRNAs) levels. Furthermore, our data imply a minimum threshold level of vsiRNA accumulation necessary for the plant to be immune to eBSV release and BSV infection.

Finally, our work highlights the crucial role of host regulation of expression of these integrated viral sequences, which can represent either a selective advantage that allows BB banana plants to resist cognate viruses or a disadvantage in some interspecific hybrids leading to the development of disease. Our study signposts a future requirement to study the expression, regulation and further molecular characterization of these seemingly ubiquitous integrated sequences, to improve breeding strategies and enable the production of horticultural and agronomic plant material free of known or suspected infective sequences.

Acknowledgements

Competing interests

None declared.

Author contributions

POD, RR, NL, FS, JMD and MC conducted the experiments. JS and SR performed the bioinformatic analyses. MP, MLI-C, POD and MC designed the experiments. POD, MP and MC wrote the paper. MLI-C and RR edited the manuscript.

ORCID

Matthieu Chabannes ORCID https://orcid.org/0000-0001-5754-5982

Pierre-Olivier Duroy ORCID https://orcid.org/0000-0002-4045-5240

Marie-Line Iskra-Caruana ORCID https://orcid.org/0000-0003-4486-2449

Nathalie Laboureau ORCID https://orcid.org/0000-0002-0105-3919

Mikhail Pooggin ORCID https://orcid.org/0000-0003-2308-393X

Rajeswaran Rajendran ORCID https://orcid.org/0000-0002-4842-1256

Sébastien Ravel ORCID https://orcid.org/0000-0001-6663-782X

Data availability

All the datasets and scripts have been uploaded on the open-source CIRAD dataverse. These data can be accessed using the links provided: BPO-59: https://doi.org/10.18167/DVN1/9K4J0D. BPO-60: https://doi.org/10.18167/DVN1/5LFLWU. BPO-64: https://doi.org/10.18167/DVN1/WJPM4N. BPO-66: https://doi.org/10.18167/DVN1/TPDEX8. GWT-1: https://doi.org/10.18167/DVN1/PNUURS. GWT-2: https://doi.org/10.18167/DVN1/XIOTAO. GWT-3: https://doi.org/10.18167/DVN1/EVHGJO. GWT-4: https://doi.org/10.18167/DVN1/TYWQ0R. GWT-5: https://doi.org/10.18167/DVN1/0JUVIK. GWT-6: https://doi.org/10.18167/DVN1/IHRVQS. GWT-7: https://doi.org/10.18167/DVN1/XF2A3E. GWT-9: https://doi.org/10.18167/DVN1/9IHYTJ. GWT-10: https://doi.org/10.18167/DVN1/JHKGIO. GWT-11: https://doi.org/10.18167/DVN1/LB5HTS. GWT-12: https://doi.org/10.18167/DVN1/YKFVR1. GWT-13: https://doi.org/10.18167/DVN1/59AVAI. GWT-14: https://doi.org/10.18167/DVN1/ZNXDNL. GWT-15: https://doi.org/10.18167/DVN1/NPRFXF. GWT-16: https://doi.org/10.18167/DVN1/IX2M3J. GWT-17: https://doi.org/10.18167/DVN1/HPQU43. GWT-18: https://doi.org/10.18167/DVN1/L4LBY7. GWT-19: https://doi.org/10.18167/DVN1/YTRCBN. Scripts for RNA-Seq and sRNA analyses: https://doi.org/10.18167/DVN1/K1OKGY.

References

Aboughanem-Sabanadzovic N, Allen TW, Frelichowski J, Scheffler J, Sabanadzovic S. 2023. Discovery and analyses of caulimovirid-like sequences in upland cotton (*Gossypium hirsutum*). *Viruses* **15**: 1648.

Ahmed FF, Hossen MI, Sarkar MAR, Konak JN, Zohra FT, Shoyeb M, Mondal S. 2021. Genome-wide identification of DCL, AGO and RDR gene families and their associated functional regulatory elements analyses in banana (*Musa acuminata*). *PLoS ONE* **16**: e0256873.

Ahmed I, Sarazin A, Bowler C, Colot V, Quesneville H. 2011. Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. *Nucleic Acids Research* **39**: 6919–6931.

Akbergenov R, Si-Ammour A, Blevins T, Amin I, Kutter C, Vanderschuren H, Zhang P, Gruissem W, Meins F Jr, Hohn T et al. 2006. Molecular characterization of geminivirus-derived small RNAs in different plant species. *Nucleic Acids Research* **34**: 462–471.

Baulcombe DC. 2022. The role of viruses in identifying and analyzing RNA silencing. *Annual Review of Virology* **9**: 353–373.

Bewick AJ, Schmitz RJ. 2017. Gene body DNA methylation in plants. *Current Opinion in Plant Biology* **36**: 103–110.

Blevins T, Podicheti R, Mishra V, Marasco M, Wang J, Rusch D, Tang H, Pikaard CS. 2015. Identification of Pol IV and RDR2-dependent precursors of 24 nt siRNAs guiding *de novo* DNA methylation in Arabidopsis. *eLife* **4**: e09591.

Blevins T, Rajeswaran R, Aregger M, Borah BK, Schepetilnikov M, Baerlocher L, Farinelli L, Meins F Jr, Hohn T, Pooggin MM. 2011. Massive production of small RNAs from a non-coding region of Cauliflower mosaic virus in plant defense and viral counter-defense. *Nucleic Acids Research* **39**: 5003–5014.

Blevins T, Rajeswaran R, Shivaprasad PV, Beknazariants D, Si-Ammour A, Park HS, Vazquez F, Robertson D, Meins F Jr, Hohn T et al. 2006. Four plant Dicers mediate viral small RNA biogenesis and DNA virus induced silencing. *Nucleic Acids Research* **34**: 6233–6246.

Borges F, Martienssen RA. 2015. The expanding world of small RNAs in plants. *Nature Reviews. Molecular Cell Biology* **16**: 727–741.

Castel SE, Martienssen RA. 2013. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nature Reviews. Genetics* **14**: 100–112.

Chabannes M, Baurens FC, Duroy PO, Bocs S, Vernerey MS, Rodier-Goud M, Barbe V, Gayral P, Iskra-Caruana ML. 2013. Three infectious viral species lying in wait in the banana genome. *Journal of Virology* **87**: 8624–8637.

Chabannes M, Gabriel M, Aksa A, Galzi S, Dufayard JF, Iskra-Caruana ML, Muller E. 2021. Badnaviruses and banana genomes: a long association sheds light on Musa phylogeny and origin. *Molecular Plant Pathology* **22**: 216–230.

Chabannes M, Iskra-Caruana ML. 2013. Endogenous pararetroviruses–a reservoir of virus infection in plants. *Current Opinion in Virology* **3**: 615–620.

Chan SW, Zilberman D, Xie Z, Johansen LK, Carrington JC, Jacobsen SE. 2004. RNA silencing genes control *de novo* DNA methylation. *Science* **303**: 1336.

Chiba S, Kondo H, Tani A, Saisho D, Sakamoto W, Kanematsu S, Suzuki N. 2011. Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. *PLoS Pathogens* **7**: e1002146.

Cote FX, Galzi S, Folliot M, Lamagnere Y, Teycheney PY, Iskra-Caruana ML. 2010. Micropropagation by tissue culture triggers differential expression of infectious endogenous Banana streak virus sequences (eBSV) present in the B genome of natural and synthetic interspecific banana plantains. *Molecular Plant Pathology* **11**: 137–144.

Dallot S, Acuna P, Rivera C, Ramirez P, Cote F, Lockhart BE, Caruana ML. 2001. Evidence that the proliferation stage of micropropagation procedure is determinant in the expression of banana streak virus integrated into the genome of the FHIA 21 hybrid (Musa AAAB). *Archives of Virology* **146**: 2179–2190.

Deleris A, Gallego-Bartolome J, Bao J, Kasschau KD, Carrington JC, Voinnet O. 2006. Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense. *Science* **313**: 68–71.

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M et al. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**: 213–217.

Didion JP, Martin M, Collins FS. 2017. Atropos: specific, sensitive, and speedy trimming of sequencing reads. *PeerJ* **5**: e3720.

Diop SI, Geering ADW, Alfama-Depauw F, Loaec M, Teycheney PY, Maumus F. 2018. Tracheophyte genomes keep track of the deep evolution of the Caulimoviridae. *Scientific Reports* **8**: 572.

Duroy PO, Perrier X, Laboureau N, Jacquemoud-Collet JP, Iskra-Caruana ML. 2016. How endogenous plant pararetroviruses shed light on Musa evolution. *Annals of Botany* **117**: 625–641.

da Fonseca GC, de Oliveira LFV, de Morais GL, Abdelnor RV, Nepomuceno AL, Waterhouse PM, Farinelli L, Margis R. 2016. Unusual RNA plant virus

integration in the soybean genome leads to the production of small RNAs. *Plant Science* 246: 62–69.

Gayral P. 2008. *Evolution des pararétrovirus endogènes de plantes: Le cas des séquences intégrées du Banana streak virus chez le bananier (*Musa *sp.).* Montpellier, France: Université de Montpellier II.

Gayral P, Blondin L, Guidolin O, Carreel F, Hippolyte I, Perrier X, Iskra-Caruana ML. 2010. Evolution of endogenous sequences of banana streak virus: what can we learn from banana (*Musa* sp.) evolution? *Journal of Virology* 84: 7346–7359.

Gayral P, Noa-Carrazana JC, Lescot M, Lheureux F, Lockhart BE, Matsumoto T, Piffanelli P, Iskra-Caruana ML. 2008. A single Banana streak virus integration event in the banana genome as the origin of infectious endogenous pararetrovirus. *Journal of Virology* 82: 6697–6710.

Gilbert C, Feschotte C. 2018. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Current Opinion in Genetics & Development* 49: 15–24.

Gong Z, Han GZ. 2018. Euphyllophyte paleoviruses illuminate hidden diversity and macroevolutionary mode of Caulimoviridae. *Journal of Virology* 92: 1218.

Ho T, Broome JC, Buhler JP, O'Donovan W, Tian T, Diaz-Lara A, Martin RR, Tzanetakis IE. 2024. Integration of Rubus yellow net virus in the raspberry genome: a story centuries in the making. *Virology* 591: 109991.

Iskra Caruana M-L, Thomas JE, Chabannes M. 2019. Banana streak. In: Jones David R, ed. *Handbook of diseases of banana, Abacá and Enset.* Wallingford, UK: Royaume-Uni: CABI, 393–408.

Iskra-Caruana ML, Baurens FC, Gayral P, Chabannes M. 2010. A four-partner plant-virus interaction: enemies can also come from within. *Molecular Plant–Microbe Interactions* 23: 1394–1402.

Ito H. 2013. Small RNAs and regulation of transposons in plants. *Genes & Genetic Systems* 88: 3–7.

Jakowitsch J, Mette MF, van Der Winden J, Matzke MA, Matzke AJ. 1999. Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. *Proceedings of the National Academy of Sciences, USA* 96: 13241–13246.

Kondratev N, Denton-Giles M, Bradshaw R, Cox M, Dijkwel P. 2020. Camellia plant resistance and susceptibility to petal blight disease are defined by the timing of defence responses. *Molecular Plant–Microbe Interactions* 33: 982–995.

Kovacova V, Janousek B. 2012. Bisprimer–a program for the design of primers for bisulfite-based genomic sequencing of both plant and Mammalian DNA samples. *The Journal of Heredity* 103: 308–312.

Le Provost G, Iskra-Caruana ML, Acina I, Teycheney PY. 2006. Improved detection of episomal Banana streak viruses by multiplex immunocapture PCR. *Journal of Virological Methods* 137: 7–13.

Lheureux F. 2002. *Etude des mécanismes génétiques impliqués dans l'expression des séquences EPRVs pathogènes des Bananiers au cours de croisements interspécifiques*, PhD. Montpellier, France: Université Sciences et Techniques du Languedoc USTL Montpellier.

Lheureux F, Carreel F, Jenny C, Lockhart BE, Iskra-Caruana ML. 2003. Identification of genetic markers linked to banana streak disease expression in inter-specific Musa hybrids. *Theoretical and Applied Genetics* 106: 594–598.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26: 589–595.

Lockhart BE, Menke J, Dahal G, Olszewski NE. 2000. Characterization and genomic analysis of tobacco vein clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. *Journal of General Virology* 81: 1579–1585.

Mari-Ordonez A, Marchais A, Etcheverry M, Martin A, Colot V, Voinnet O. 2013. Reconstructing *de novo* silencing of an active plant retrotransposon. *Nature Genetics* 45: 1029–1039.

Martin G, Baurens FC, Droc G, Rouard M, Cenci A, Kilian A, Hastie A, Dolezel J, Aury JM, Alberti A *et al*. 2016. Improvement of the banana '*Musa acuminata*' reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics* 17: 243.

Matzke MA, Mosher RA. 2014. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nature Reviews. Genetics* 15: 394–408.

McCue AD, Panda K, Nuthikattu S, Choudury SG, Thomas EN, Slotkin RK. 2015. ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. *EMBO Journal* 34: 20–35.

Mölder F, Jablonski K, Letcher B, Hall M, Tomkins-Tinch C, Sochat V, Forster J, Lee S, Twardziok S, Kanitz A *et al*. 2021. Sustainable data analysis with Snakemake [v.2; peer review: 1 approved, 1 approved with reservations]. *F1000Research* 10(33): 29032.

Pooggin MM. 2013. How can plant DNA viruses evade siRNA-directed DNA methylation and silencing? *International Journal of Molecular Sciences* 14: 15233–15259.

Qu F, Ye X, Morris TJ. 2008. Arabidopsis DRB4, AGO1, AGO7, and RDR6 participate in a DCL4-initiated antiviral RNA silencing pathway negatively regulated by DCL1. *Proceedings of the National Academy of Sciences, USA* 105: 14732–14737.

Raja P, Jackel JN, Li S, Heard IM, Bisaro DM. 2014. Arabidopsis double-stranded RNA binding protein DRB3 participates in methylation-mediated defense against geminiviruses. *Journal of Virology* 88: 2611–2622.

Raja P, Sanville BC, Buchmann RC, Bisaro DM. 2008. Viral genome methylation as an epigenetic defense against geminiviruses. *Journal of Virology* 82: 8997–9007.

Rajeswaran R, Seguin J, Chabannes M, Duroy PO, Laboureau N, Farinelli L, Iskra-Caruana ML, Pooggin MM. 2014. Evasion of short interfering RNA-directed antiviral silencing in *Musa acuminata* persistently infected with six distinct banana streak pararetroviruses. *Journal of Virology* 88: 11516–11528.

Ricciuti E, Laboureau N, Noumbissie G, Chabannes M, Sukhikh N, Pooggin MM, Iskra-Caruana ML. 2021. Extrachromosomal viral DNA produced by transcriptionally active endogenous viral elements in non-infected banana hybrids impedes quantitative PCR diagnostics of banana streak virus infections in banana hybrids. *The Journal of General Virology* 102: 1670.

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

Seguin J, Otten P, Baerlocher L, Farinelli L, Pooggin MM. 2014. MISIS: a bioinformatics tool to view and analyze maps of small RNAs derived from viruses and genomic loci generating multiple small RNAs. *Journal of Virological Methods* 195: 120–122.

Seguin J, Otten P, Baerlocher L, Farinelli L, Pooggin MM. 2016. MISIS-2: a bioinformatics tool for in-depth analysis of small RNAs and representation of consensus master genome in viral quasispecies. *Journal of Virological Methods* 233: 37–40.

Thomas JE, Iskra-Caruana M-L, Kumar L, Roux N, Chabannes M, Sebastien M, Mathieu Y, Chase R, Houwe I. 2015. Position paper on a strategy to distribute banana (Musa) germplasm with endogenous Banana streak virus genomes. [WWW document] URL https://drive.google.com/file/d/0B6WMCDtu_LjpU0pIM1AyWGFqNkk/view?resourcekey=0-1fX6r3J04wit7N4NNj4scA [accessed 12 September 2024].

Valli AA, Gonzalo-Magro I, Sanchez DH. 2023. Rearranged endogenized plant pararetroviruses as evidence of heritable RNA-based immunity. *Molecular Biology and Evolution* 40: msac240.

Vassilieff H, Geering ADW, Choisne N, Teycheney PY, Maumus F. 2023. Endogenous caulimovirids: fossils, zombies, and living in plant genomes. *Biomolecules* 13: 1069.

Vaucheret H, Voinnet O. 2024. The plant siRNA landscape. *Plant Cell* 36: 246–275.

Wang Z, Miao H, Liu J, Xu B, Yao X, Xu C, Zhao S, Fang X, Jia C, Wang J *et al*. 2019. *Musa balbisiana* genome reveals subgenome evolution and functional divergence. *Nature Plants* 5: 810–821.

Zhai J, Bischof S, Wang H, Feng S, Lee TF, Teng C, Chen X, Park SY, Liu L, Gallego-Bartolome J *et al*. 2015. A one precursor one siRNA model for pol IV-dependent siRNA biogenesis. *Cell* 163: 445–455.

Zhao JH, Fang YY, Duan CG, Fang RX, Ding SW, Guo HS. 2016. Genome-wide identification of endogenous RNA-directed DNA methylation loci associated with abundant 21-nucleotide siRNAs in Arabidopsis. *Scientific Reports* 6: 36247.

Zhong X, Du J, Hale CJ, Gallego-Bartolome J, Feng S, Vashisht AA, Chory J, Wohlschlegel JA, Patel DJ, Jacobsen SE. 2014. Molecular mechanism of action of plant DRM *de novo* DNA methyltransferases. *Cell* 157: 1050–1060.

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2007. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics* 39: 61–69.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Dataset S1 Summary of total number of reads of each sRNA size-class (20–25 nt) in forward and reverse orientation mapped on each reference genome.

Dataset S2 Single-base resolution maps of 20–25 nt viral sRNAs in forward and reverse orientation mapped on each reference genome.

Dataset S3 Summary of total number of 125-nt RNA reads in forward and reverse orientation mapped on each reference genome.

Dataset S4 Single-base resolution maps of 125-nt viral RNAs in forward and reverse orientation mapped on each reference genome.

Fig. S1 RNA blot hybridization analysis of eBSV-derived siRNA in eBSV-free *Musa acuminata* Cavendish (plant and the seedy diploid *Musa balbisiana* PKW plant).

Fig. S2 Characterization of endogenous banana sRNAs in noninfected *Musa acuminata* plants or individually infected with three distinct BSV species, in the seedy diploid PKW plant and a self-pollinated progeny and triploid interspecific AAB noninfected and BSOLV-, BSGFV- or BSIMV-infected plants.

Fig. S3 Single-nucleotide resolution maps of 21-nt and 24-nt eBSV-derived sRNAs from the seedy diploid *Musa balbisiana* PKW plant and a self-pollinated progeny.

Fig. S4 Single-nucleotide resolution maps of the total 20–25-nt eBSV-derived sRNAs from self-pollinated PKW plants homozygous for eBSOLV-1 or eBSOLV-2 alleles.

Fig. S5 DNA methylation analyses of PKW eBSV determined by bisulfite sequencing.

Fig. S6 Comparing amounts of vsiRNA accumulated from eBSV in B genome haploid or diploid plants.

Methods S1 Illumina deep-sequencing and bioinformatics analysis.

Notes S1 Raw chromatograms of PCR products amplified from eBSIMV, eBSOLV and eBSGFV genomic DNA of PKW initially treated with sodium bisulfite.

Table S1 Listing of short DNA oligonucleotide probes used for detection of eBSV-derived siRNAs in PKW plant.

Table S2 Listing of the degenerated primers designed by the BIS-PRIMER software used to quantify eBSV methylation in PKW plants.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.