Phytopathology Research

RESEARCH

Open Access

Identification of new banana endogenous virus sequences highlights the hallmark gene encoded by retroviruses integrated in banana genomes

Huazhou Chen[1], Huaping Li[1*] and Xueqin Rao[1*]

Abstract

Endogenous pararetrovirus sequences (EPRVs) originated from DNA viruses of the family *Caulimoviridae* are widely present in plant genomes. Banana streak viruses (BSVs) are a group of circular double-stranded DNA viruses in the genus *Badnavirus* of the family *Caulimoviridae*. Banana endogenous virus sequences (BEVs) derived from the ancestral genes of badnaviruses and fixed in the genomes of various bananas. However, the genomic characteristics of BEVs remain unknown. In this study, we identified 2 new variants of BEVs GZ5 and GZ13 by sequences analyses, Southern blot, and fluorescent *in situ* hybridization (FISH). BEV GZ5 had one copy of integration in the BB genome of bananas, while BEV GZ13 was only present in the genome of the variety Dajiao. Importantly, BEV GZ5 contained a complete gene of reverse transcriptase (RT) and ribonuclease H (RNase H) (RT/RNase H). In addition, a 340-bp inverted repeat sequence partially overlapping with RNase H was found upstream and downstream of BEV GZ5. However, the amino acid sequences of BEV GZ5 had deletions and mutations compared with BSVs. The bioinformatics analyses showed that BEV GZ5 protein composed of 412 amino acids with a molecular weight of 47.37 kDa and an isoelectric point of 9.40. Leucine, isoleucine, and lysine (Lys) were the main amino acids of BEV GZ5 protein. The analyses revealed that BEV GZ5 protein contained 35 potential phosphorylation sites. Additionally, it was a hydrophilic protein without a signal peptide and transmembrane region. The secondary structure of BEV GZ5 protein consisted of 37.26% α-helix, followed by 36.25% random coil. To our knowledge, this is the first report that novel BEVs with the complete gene of RT/RNase H has been characterized, which provide a basis for further exploration the function and integration mechanism of BEVs in bananas.

Keywords Endogenous pararetrovirus sequences (EPRVs), Banana streak virus (BSV), *Musa*, Banana endogenous virus sequences (BEVs)

Background

Endogenous pararetrovirus sequences (EPRVs), which derived from the family *Caulimoviridae*, are present in a variety of plant genomes, including rice (Chen and Kishima 2016), tobacco (Mette et al. 2002), banana (Chabannes et al. 2021), and other plants. Most EPRVs are fragmented and rearranged compared with the corresponding virus genomes. They likely became integrated in the host genome via illegitimate recombination during

*Correspondence:
Huaping Li
huaping@scau.edu.cn
Xueqin Rao
raoxq@hotmail.com
[1] Guangdong Province Key Laboratory of Microbial Signals and Disease Control, College of Plant Protection, South China Agricultural University, Guangzhou 510642, China

the repair of breaks in host DNA (Feschotte and Gilbert 2012), and then retained in the plant genome (Harper et al. 2002; Gayral and Iskra-Caruana 2009; Iskra-Caruana et al. 2014) as a component of the host plant genome (Yu et al. 2019). However, integrated EPRVs persisted in host genomes as they enhanced the resistance of plants against virus infection by inducing transcriptional or post-transcriptional gene silencing of homologous sequences (Hull et al. 2000). For instance, Staginnus et al. (2007) observed a significant increase in the transcript levels of EPRVs after viral infection. Mette et al. (2002) demonstrated that the enhancer of tobacco EPRVs was silenced in the transgenic tobacco, but expressed in transgenic *Arabidopsis*, suggesting that EPRVs played a crucial role in the biological process of host plant resistance to viruses.

EPRVs were relics of ancient viruses that infected host plants (Chen and Kishima 2016). These EPRVs coevolved with their host plants, resulting in a clear concordance between EPRVs and their hosts. Therefore, the integration time of EPRVs was estimated by analyzing co-evolutionary genes between endogenous viruses and their hosts (Feschotte and Gilbert 2012; Chen et al. 2017; Diop et al. 2018), especially the divergence time of the ancestor host. For example, the origin of lentivirus could potentially be traced back to more than 12 million years, when the host range of rabbit endogenous lentivirus type K expanded to rabbits and hares (Katzourakis et al. 2007; van der Loo et al. 2009). Schmidt et al. (2021) estimated that the integration of beetEPRV3 occurred approximately 13.4–7.2 million years ago based on the divergence times of *Corollinae* and *Nanae*. Therefore, EPRVs served as markers to clarify the phylogenetic relationship between the virus and its host.

Banana streak virus (BSV) belongs to the genus *Badnavirus* in the family *Caulimoviridae* (Jaufeerally-Fakim et al. 2006; Staginnus et al. 2009). The International Committee on Taxonomy of Viruses (ICTV) recommended 80% nucleotide sequence identity in the RT/RNase H-coding region as a criterion for distinguishing species of badnaviruses (Geering et al. 2014). At present, thirteen distinct banana streak viruses (BSVs) have been identified (Lheureux et al. 2007; Geering et al. 2011), such as banana streak OL virus (BSOLV), banana streak GF virus (BSGFV), banana streak VN virus (BSVNV), etc. Additionally, endogenous banana streak virus (eBSV) and banana endogenous badnavirus sequences (BEVs) were discovered in the banana genomes (Geering et al. 2005; Harper et al. 2005; Gayral and Iskra-Caruana 2009; Chabannes et al. 2021). Based on partial sequences of the RT/RNase H genes, BSVs and BEVs were classified into three distinct clades. Clade I and III contained BSVs (Harper et al. 2005; Gayral and Iskra-Caruana 2009; Li et al. 2020). However, BSVs in Clade III were endemic in East Africa (Chabannes et al. 2021). Clade II only comprised different BEVs (Geering et al. 2005; Chabannes et al. 2021).

BEVs were present in various banana genomes (Gayral and Iskra-Caruana 2009; D'Hont et al. 2012; Iskra-Caruana et al. 2014). Following the classification criteria for badnavirus recommended by ICTV (Geering et al. 2014; Chabannes et al. 2021), many BEVs have been identified (Geering et al. 2005). Similarly, BSVs and BEVs from bananas were discovered in Uganda (Harper et al. 2005). At present, BEV NGA (D'Hont et al. 2012), UC, UD, UF, UG, UH, P, and Q (Chabannes et al. 2021) were related to the BEVs reported by Geering et al. (2005). The presence of BEVs in banana genomes correlated with genotypes and varieties of bananas. Chabannes et al. (2021) investigated BEVs among diverse genotypes of bananas in Uganda and discovered that BEV UC and UG were commonly present in bananas with A and B genotypes, respectively. In contrast, BEV UD, UH, and NGA were found only in the A genomes of bananas. Additionally, BEV P and Q were solely in the B genomes of bananas.

BSVs are episomal in the bananas; while eBSVs exist in the banana genomes, some can release the infectious BSVs, but others cannot. Nowadays, an increasing number of BEVs are found in banana genomes. However, none of the BEVs can be activated as viruses due to their currently known limited genes, namely the partial RT/RNase H gene of BSVs (D'Hont et al. 2012; Geering et al. 2014), their gene sequences and structural characteristics in banana genomes are still unknown. In a previous study, we found that BEV GZ5 and GZ13 were new BEVs. BEV GZ5 was located on chromosome 5 of *Musa balbisiana* subsp. PKW (BB) (Rao et al. 2023). However, BEV GZ13 was not found on any chromosomes (Rao et al. 2023) of *M. balbisiana* subsp. PKW (BB) (Davey et al. 2013), *M. acuminata* subsp. DH-Pahang (AA) (Belser et al. 2021), or *M. schizocarpa* subsp. HN8 (SS) (Belser et al. 2018). In this study, we identified that BEV GZ5 and GZ13 were new BEVs by sequence analyses, Southern blot, and fluorescent *in situ* hybridization (FISH). Additionally, we characterized the upstream and downstream sequences of BEV GZ5 and performed bioinformatics analyses, which will lay a foundation for further research on the function of BEVs in bananas.

Results

Nucleotide and amino acid sequences analyses of new BEVs

According to the classification criteria of badnaviruses, BEV GZ5 and GZ13 were identified from different banana samples. BEV GZ5 was detected in the samples collected from Maoming, Qinzhou, and Yulin, while GZ13 was found in the sample from Wuzhou. To analyze whether BEV GZ5 and GZ13 were BSVs, their nucleotide sequences and amino acid sequences compared with those of BSVs from GenBank. The results showed that the nucleotide sequence identity of BEV GZ5 ranged from 62.4% to 67.9%, and the amino acid sequence identity of BEV GZ5 ranged from 61.8% to 71.4% compared with BSVs. Similarly, BEV GZ13 shared 64.9%–68.5% sequence identity at the nucleotide level and 62.6%–70.9% sequence identity at the amino acid level when compared with BSVs. These results suggested that BEV GZ5 and GZ13 were not BSVs.

To analyze whether BEV GZ5 and GZ13 were new BEVs, their nucleotide and amino acid sequences were compared with the BEVs from GenBank and those reported by Rao et al. (2023). The nucleotide sequence identities of BEV GZ5 and GZ13 shared 59.8%–77.2% and 61.5%–77.6% with those of BEVs from GenBank (Fig. 1a), while the amino acid sequence identities of BEV GZ5 and GZ13 had 63.9%–80.1% and 65.4%–83.8% with those of BEVs from GenBank (Fig. 1b), respectively. In addition, when compared with other BEVs reported by Rao et al. (2023), the nucleotide identities of BEV GZ5 and GZ13 showed 64.9%–75.9% and 65.0%–77.7%, while the amino acid identities of BEV GZ5 and GZ13 were 65.2%–78.6% and 63.2%–84.4%, respectively. These results suggested that BEV GZ5 and GZ13 were new BEVs.

To demonstrate the evolutionary relationships of BEV GZ5 and GZ13 with other BEVs, a phylogenetic tree based on the partial genomes of RT/RNase H gene was constructed. The results indicated that BEVs in this study and from GenBank as well as BSVs from GenBank were grouped into three distinct clades. Both Clade I and II contained BEVs, while Clade I and III contained BSVs. However, BEV GZ5 was clustered into a different branch from other BEVs in Clade II, GZ13 was in a different branch closer to BEV 23 (Bat36, AY189430) than other BEVs in Clade II (Fig. 1c). BEV GZ5 and GZ13 were new BEVs confirmed by the phylogenetic relationship constructed based on the partial sequences of the RT/RNase H-encoding gene.

Southern blot analyses of BEV GZ5

To analyze the endogenous characteristic of BEV GZ5, Southern hybridization was performed on the main banana cultivars in Guangdong. The results demonstrated that the BEV GZ5 probe only generated hybridization signals in bananas with BB genomes, such as Fenza 1 and Guangfen 1. The size of the hybridization band corresponded to the banana genomes. However, the bananas with AAA genomes, such as Williams and Brazilian bananas, did not show any hybridization signals (Fig. 2a). Compared with the undigested genomes of Fenza 1 (ABBB) and Guangfen 1 (ABB) (lanes 1 and 2) (Fig. 2b), the results indicated that the digested banana genomes of Fenza 1 (ABBB) and Guangfen 1 (ABB) had only one hybridization band of approximately 9400 bp (lanes 3 and 4) (Fig. 2b), respectively. These findings suggested that BEV GZ5 had only one integration site in bananas with BB genome.

Identification of BEV GZ5 by FISH

To confirm the chromosomal location of BEV GZ5, we performed FISH tests using the digoxigenin-labeled BEV GZ5 probes on the mitotic chromosomes of Guangfen 1 (ABB), one of the main banana cultivars in Guangdong. We observed only one hybridization signal in the chromosome of Guangfen 1 (Fig. 2c), which confirmed the results of the chromosomal location (Rao et al. 2023) and Southern blot analyses in this study. It suggested that BEV GZ5 was only present in the BB genomes of bananas with one integration site.

Analyses of endogeny and location of BEV GZ13

To confirm that BEV GZ13 originated from banana genomes, Southern blot was performed. The results showed that the BEV GZ13 probes only reacted with Dajiao (ABB) genome and not with those of the other five tested banana varieties, namely Brazilian (AAA), Williams (AAA), Fenza 1 (ABBB), Guangfan 1 (ABB), and Jinfen (ABB) (Fig. 2d). Additionally, the size of the hybridization band was similar to that of the banana genome. It suggested that BEV GZ13 integrated only in the genome of Dajiao (ABB).

Analyses of the upstream and downstream sequences of BEV GZ5 in Guangfen 1

To investigate the upstream and downstream sequences of BEV GZ5, a 10,000-bp gene fragment was amplified through PCR based on the BEV GZ5 location in the banana BB genome, then cloned and sequenced. After assembly, a 9734-bp nucleotide sequence was obtained, which contained a total of 1743-bp BEV GZ5, and a 340-bp inverted repeat sequences (IRS) upstream, and downstream of BEV GZ5 with partial sequences overlapping with the RNase H gene of BEV GZ5 (Fig. 3a). In addition, a gene encoding a zinc finger protein (ZinF) was discovered upstream of BEV GZ5, and a 621-bp gene sequence (GBV1-1) was found downstream of BEV GZ5, which shared 70.6% nucleotide sequence identity with that of grapevine badnavirus 1 isolate VLJ-178 (GBV1-VLJ-178); however, no functional proteins were predicted.

Analyses of upstream and downstream sequence of BEV GZ5 revealed that the upstream sequence (1–3853bp) of BEV GZ5 showed 99.9% nucleotide sequence identity with that of *M. balbisiana* subsp. PKW (BB). Similarly, the downstream sequence

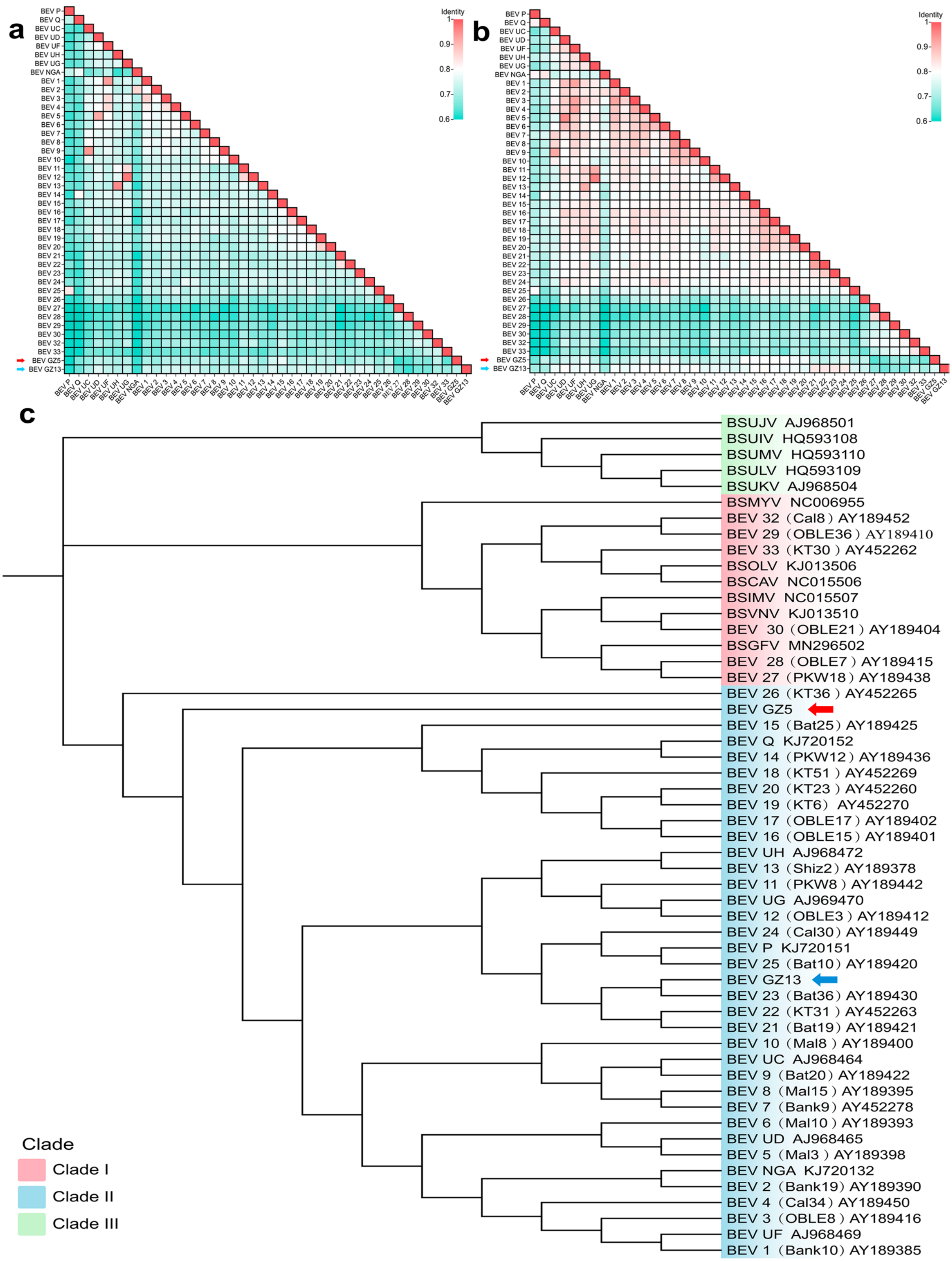

Fig. 1 Analyses of BEV GZ5 and GZ13 based on the partial sequences of RT/RNase H with BEVs from GenBank. **a** Alignment of nucleotide sequence. **b** Alignment of amino acid sequence. **c** Phylogenetic tree based on the partial RT/RNase H region

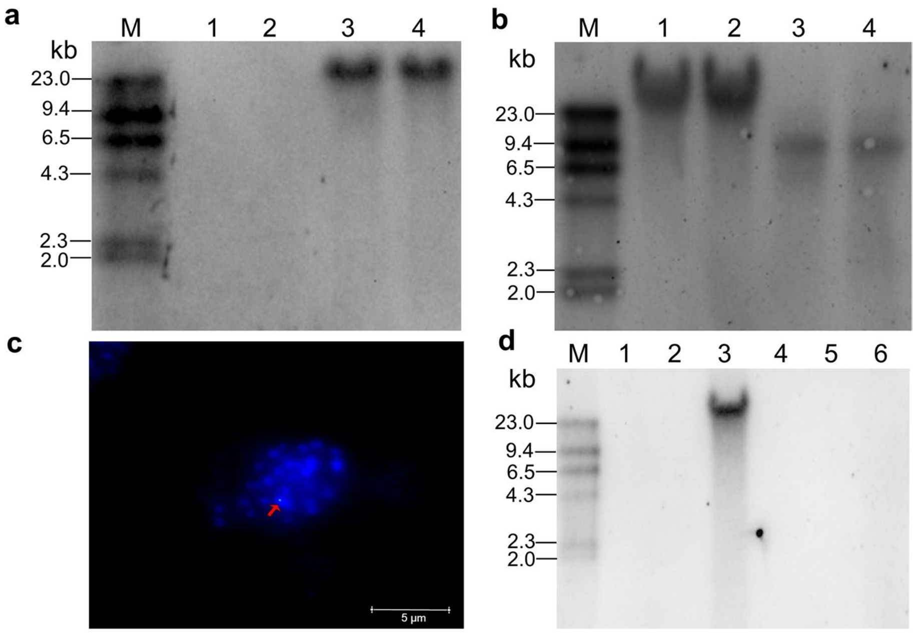

Fig. 2 Identification of BEV GZ5 and GZ13. **a** Southern blot analysis of BEV GZ5 in undigested total banana genomic DNA. 1, Brazilian (AAA); 2, Williams (AAA); 3, Fenza 1 (ABBB); 4, Guangfen 1 (ABB). **b** Southern blot analyses of BEV GZ5 in banana total genomic DNA. 1–2, Undigested total genomic DNA of Fenza 1 and Guangfen 1; 3–4, Total genomic DNA of Fenza 1 and Guangfen 1 digested by *Eco*R V. **c** Localization of BEV GZ5 integrations in the chromosomes of Guangfen 1 by FISH. The bright spot indicated by the red arrows was BEV GZ5, while blue represented the banana chromosomes counterstained with DAPI. **d** Southern blot analyses of BEV GZ13 in undigested banana total genomic DNA. 1, Brazilian; 2, Williams; 3, Dajiao (ABB); 4, Fenza 1; 5, Guangfan 1; 6, Jinfen (ABB). M, Digoxigenin labeled marker

(5597–9734bp) of BEV GZ5 shared 99.0% nucleotide sequence identity with that of *M. balbisiana* subsp. PKW (BB). Therefore, the upstream and downstream sequences of BEV GZ5 had the highest nucleotide sequence identities with the BB genomes of bananas.

When comparing the nucleotide sequence of BEV GZ5 with the corresponding region of the badnaviruses, the highest sequence identity was 71.67% at the nucleotide level with grapevine roditis leaf discoloration-associated virus isolate w4. Furthermore, the amino acid sequence of BEV GZ5 was aligned with the corresponding region of BSVs; it was found that BEV GZ5 contained the intact encoding gene of RT/RNase H, which was common to all the badnaviruses (Fig. 3b). Additionally, the amino acid sequence of BEV GZ5 included typical motifs of the RT/RNase H domain. However, it differed from BSVs in that there were 13 amino acid mutation sites and 2 deletion sites (Fig. 3b). Among them, 9 of the 13 amino acid mutation sites and 1 deletion site were in the RT region, 2 of them in RNase H region, and 1 deletion site in the intergenic region, respectively. Motif 3 in the RT region was the most variable motif in BEV GZ5 compared with the corresponding region of the BSVs.

Amino acid sequence composition and physicochemical properties of the BEV GZ5 protein

To identify the fundamental characteristics and potential structures of the protein encoded by BEV GZ5, we conducted bioinformatics analyses, which will aid in determining the function of the BEVs. The amino acid sequence composition and physicochemical properties of BEV GZ5 protein were analyzed (Table 1). The results revealed that the proteins of the BEV GZ5 and the chosen ten BSVs consisted of 410–414 amino acid residues with a relative molecular weight of 46.57–47.56 kDa. Their isoelectric point ranged from 9.14 to 9.42, with 44 to 53 negatively charged residues, such as aspartic acid (Asp) and glutamate acid (Glu), and 61 to 67 positively charged residues, including arginine (Arg) and lysine (Lys). The aliphatic amino acid coefficient ranged from 85.75 to 95.17. The instability coefficient of the BEV GZ5 and the ten BSVs proteins ranged from 30.27 to 46.16, indicating that they were unstable (with coefficients more than 30). The predicted hydrophilicity values ranged from -0.454 to -0.311, which suggested that they were hydrophilic (with negative values). All the proteins of BEV GZ5 and the ten BSVs consisted of 20 amino acids with similar proportions of different amino acids (Additional file 1:

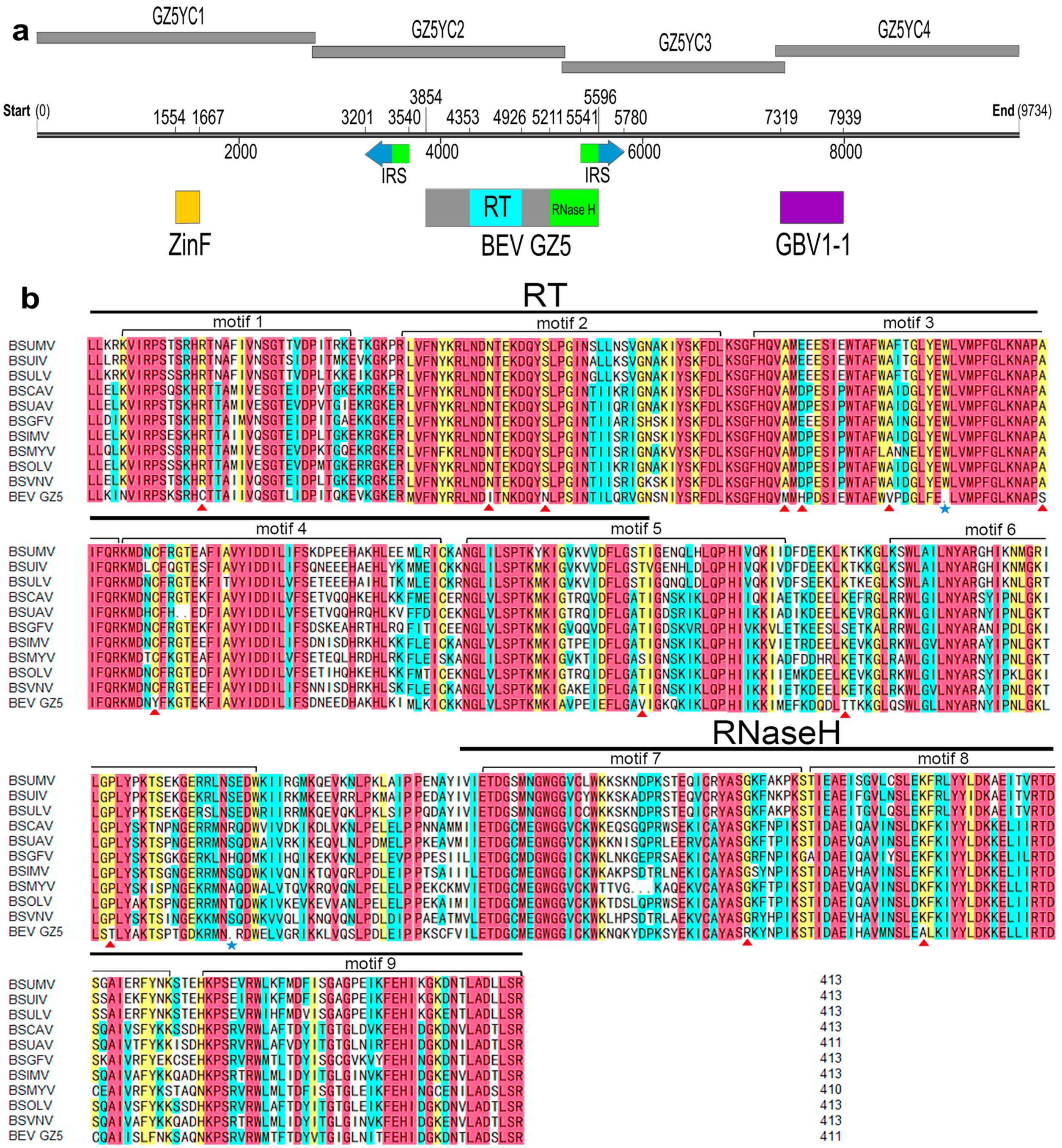

Fig. 3 Schematic diagram of BEV GZ5 and the corresponding protein alignment of BEV GZ5 and BSVs. **a** Schematic diagram of BEV GZ5. GZ5YC1–GZ5YC4 indicated BEV GZ5 was amplified in four segments. The yellow box represented zinc finger protein (ZinF), the gray box containing RT/RNase H indicated BEV GZ5, the light blue represented RT, the green box represented RNase H, and the purple box denoted GBV1-1. The arrows showed the reverse repeat sequence (IRS), the green box in the arrows represented that the region of IRS overlapped with RNase H, and the dark blue one in the arrows represented the non-encoding region. **b** Amino acid sequence of BEV GZ5 compared with the corresponding protein of BSVs, orange indicated 100% sequence identity, yellow represented more than 75% sequence identity, blue denoted greater than 50% sequence identity, and white expressed less than 25% sequence identity. The red triangles were the amino acid mutation sites, and the blue stars were the amino acid deletion sites

Table 1 Primary structure analysis of corresponding proteins BEV GZ5 and BSVs

BSV/BEVs	Amino acid	Molecular weight/kDa	Isoelectric points	Asp (-)	Glu (-)	Arg (+)	Lys (+)	Instability coefficient	Aliphatic-aa coefficient	Hydrophilicity
BEV GZ5	412	47.37	9.40	25	19	17	45	42.16	94.42	-0.311
BSVNV	414	47.14	9.26	24	25	17	46	30.27	95.17	-0.334
BSOLV	414	47.48	9.24	24	29	19	47	37.94	90.68	-0.366
BSGFV	414	47.24	9.14	21	31	21	43	37.64	92.10	-0.359
BSUAV	411	47.26	9.29	25	25	23	40	36.51	94.16	-0.316
BSMYV	410	46.57	9.28	21	26	20	41	34.60	91.10	-0.312
BSCAV	414	47.61	9.27	25	27	22	43	38.90	88.31	-0.440
BSIMV	414	47.19	9.26	24	26	20	43	32.83	95.17	-0.355
BSUIV	414	47.56	9.26	20	33	19	47	44.31	85.97	-0.419
BSUMV	414	47.33	9.42	19	31	21	46	37.67	85.75	-0.450
BSULV	414	47.44	9.41	18	32	22	44	46.16	86.43	-0.454

Table S1). Among them, leucine, isoleucine, and Lys were more abundant, while cysteine was the least one. The difference of amino acids might be related to the properties of the viral proteins.

Prediction of subcellular localization, transmembrane regions, secondary structure, and signal peptide regions of BEV GZ5 proteins

The subcellular localizations of proteins encoded by BEV GZ5 and ten BSVs were predicted by Wolfpsort, the results indicated that they were primarily present in peroxisomes and the cytoplasm (Table 2). The transmembrane region prediction for BEV GZ5 protein showed that it lacked a transmembrane region. Furthermore, the prediction of signal peptide region of BEV GZ5 protein indicated that it did not possess a signal peptide region (Additional file 2: Figure S1). These results suggested that the BEV GZ5 protein was hydrophilic with a negative charge in most amino acid sites (Fig. 4a). The hydrophilicity results obtained from the ProtScale online tool were consistent with the results from the online software ExPASy Proteomic. The secondary structure of the BEV GZ5 protein was composed of 37.26% α-helix, 36.25% random coil, 20.28% extended strand, and 6.41% β-fold.

Phosphorylation sites of BEV GZ5 protein

The prediction of phosphorylation sites showed that the BEV GZ5 protein had 35 potential phosphorylation sites (Fig. 4b). Of these, 17 were identified as potential serine phosphorylation sites, 11 as threonine phosphorylation sites, and 7 as tyrosine phosphorylation sites. These results suggested that serine was the primary site for potential phosphorylation on the BEV GZ5 protein.

Discussion

Endogenous viral sequences are common in host plant genomes (Ndowora et al. 1999; Kunii et al. 2004; Geering et al. 2005; Yu et al. 2019), and play a significant role in shaping the genome of the host plant (Richert-Poggeler et al. 2021). BEVs with partial RT/RNase H sequences

Table 2 Subcellular localization prediction of BEV GZ5 and corresponding proteins of BSVs

BSVs/BEV GZ5	Peroxisome	Cytoplasm	Cytoplasm-Nuclear	Nuclear	Mitochondria	endoplasmic reticulum
BEV GZ5	13	8.5	6	2.5	2	1
BSVNV	13	9.5	6.5	2.5	2	
BSOLV	15	6.5	5.5	3.5	2	
BSGFV	12	9.5	7	3.5	2	
BSUAV	16	6.5	5	2.5	2	
BSMYV	13	6.5	6	4.5	3	
BSCAV	14	5.5	6	5.5	2	
BSIMV	15	8.5	6	2.5	1	
BSUIV	10	9.5	7	3.5	2	2
BSUMV	12	9.5	6.5	2.5	3	
BSULV	11	9.5	7	3.5	3	

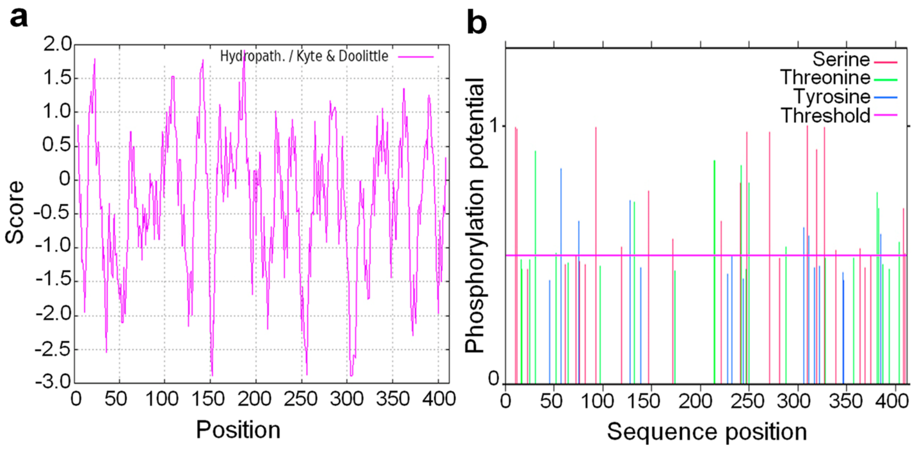

Fig. 4 Prediction of hydrophilicity and potential phosphorylation sites of BEV GZ5 protein. **a** Hydrophobicity or hydrophilicity prediction. The abscissa represented the position of the amino acid, while the ordinate represented the hydrophobicity score. The graph displayed hydrophobicity as positive values and hydrophilicity as negative values. **b** Potential phosphorylation sites prediction

had been found in various genotypes of banana (Geering et al. 2005; Chabannes et al. 2021). However, we found that BEV GZ5 had the complete genes of RT/RNase H. To our knowledge, this is the first report that BEVs with the complete genes of RT/RNase H has been characterized. Our findings provide a theoretical foundation for exploring the integration mechanisms of BEVs.

As molecular fossils, BEVs provide strong evidence for the viral ancestors infecting bananas. However, it is important to note that viruses mainly classified based on their relevant biological information such as infectivity, morphology, and transmission. This suggests that it is difficult to recognize ancestral viruses without biological information (Vassilieff et al. 2023). To distinguish between endogenous pararetroviral sequences and homologous episomal viruses, Staginnus et al. (2009) suggested adding the prefix 'E-' or 'e-' or the suffix '-EPRS' to the integrated viral sequences. Assuming that EPRVs were molecular fossils of ancestral viruses, Geering et al. (2010) classified BEVs into tentative genera based on the classification criterion of badnaviruses. Therefore, many tentative species of BEVs were identified based on the threshold of 80% nucleotide sequence identity of the RT/RNase H gene (Geering et al. 2010; D'Hont et al. 2012; Chabannes et al. 2021). According to this classification criterion, we identified two new tentative species of BEVs, BEV GZ5 and GZ13. Southern hybridization confirmed the endogeny of the two BEVs. However, BEV GZ5 was distinct from BEV GZ13 and the BEVs from GenBank (Geering et al. 2010; D'Hont et al. 2012; Chabannes et al. 2021) in that it contained intact encoding genes of RT/RNase H and had an inverted repeat sequence at its both ends. The nucleotide sequence upstream and downstream of BEV GZ5 showed the highest identities with the BB genomes of bananas, respectively. Furthermore, the genetic structure of BEV GZ5 was different from both eBSVs (Cote et al. 2010; Iskra-Caruana et al. 2010; Chabannes et al. 2013) and episomal BSVs (Lheureux et al. 2007; James et al. 2011). Therefore, BEV GZ5 was classified as novel BEVs.

Endogenous pararetroviruses could be identified as a distinct category of transposable elements (TEs), as they lacked the repeats of retrotransposons with long terminal repeats (LTRs) (Jakowitsch et al. 1999). EPRVs could potentially be domesticated as TEs and integrated in host genomes during evolution (Yu et al. 2019). DNA transposons carrying terminal inverted repeat sequences were generally transposed by transposases encoded by their autonomous elements, and the target site sequences at both ends of the element were duplicated when a transposon was actively inserted into the genome (Liu et al. 2012). In tobacco, Jakowitsch et al. (1999) discovered a 63-bp inverse repeat sequence that might be related to the EPRVs recombination process in tobacco genes. In this study, we found a pair of IRS that partial sequences overlapped with the RNase H gene at both ends of BEV GZ5, which was similar to the transposable unit, suggesting that the IRS might play a role in the integration of BSVs in the banana genome or in the acquisition of episomal BSV genes by bananas. Although there was no direct evidence of BSV genes integration and acquisition, the results of Southern hybridization and FISH demonstrated the integration of BEV GZ5 in the bananas with BB genomes. In addition, the IRS at both ends of BEV GZ5 might help the replication of endogenous viral genes via gene homologous recombination (White et al. 1994) or promote chromosomal rearrangements through the identical mechanism (Hughes and Coffin 2001) to improve the environmental adaptability of bananas.

BSVs had integrated in the genomes of bananas in a rearranged and fragmented manner (D'Hont et al. 2012). The integration of EPRVs in host plants had enhanced their resistance by silencing genes at the transcriptional and post-transcriptional levels (Hull et al. 2000). For example, the significant quantity of small RNAs derived from EPRVs may improve host resistance to related viral infections (Huang and Li 2018). Small RNAs derived from EPRVs might suppress potentially pathogenic EPRVs (Schmidt et al. 2021). Therefore, the EPRVs integration in host plants was considered a beneficial component of host resistance against viruses (Zhang et al. 2015; Yang et al. 2016). It is interesting to know that the RT/RNase H gene was present in both retroviruses and retrotransposons (Xiong and Eickbush 1990; Harper et al. 2002). Unlike most BEVs from GenBank, BEV GZ5 contained the intact encoding genes of both RT and RNase H. Importantly, both BEV GZ5 and BEVs from GenBank contained the conserved structural domain of RT/RNase H, which might protect bananas against the pathogens with RT/RNase H genes. It suggests that the hallmark gene of RT and RNase H encoded by retroviruses has been integrated in banana genomes.

RT-like sequences were found in various elements, including plant and animal DNA viruses, transposable elements in fruit flies, yeast, trypanosomes, slime mold, and mitochondrial introns (Xiong and Eickbush 1988), suggesting that they might share an ancestor. Rao et al. (2023) demonstrated that BEVs and BSVs coevolved with bananas and originated from a common badnavirus ancestor, some BEVs predated the differentiation of *Musa* ancestor. This study revealed that different types of BEVs shared a partial RT/RNase H region, indicating that BEVs had a common origin with retroviruses, transposable elements, and mitochondrial introns, which advanced the coevolutionary timeline among BEVs, BSVs, and bananas.

The distribution of BEVs varied across the chromosomes of bananas, despite their presence in multiple banana genomes (Geering et al. 2005; Perrier et al. 2011; Chabannes et al. 2021). Chabannes et al. (2021) discovered different distribution patterns of BEVs in bananas. This study found that BEV GZ5 was present in BB genomes of bananas at a low copy, as confirmed by Southern hybridization and FISH. However, it did not found in AA and SS genomes of bananas. BEV GZ13 was only present in Dajiao (ABB) and not in other ABB banana varieties. It hypothesized that the integration of BEV GZ5 in BB genomes of bananas was earlier than that of BEV GZ13. Additionally, due to external factors, the integration of BEV GZ13 was a low-frequency event and was challenging to spread among different banana varieties with the same genotype.

Bioinformatics analyses were used to identify the potential functions of the BEV GZ5 protein. The results revealed that the RT/RNase H protein encoded by BEV GZ5 was a hydrophilic protein and lacked a transmembrane structure. It was primarily present in peroxisomes and the cytoplasm. Phosphorylation was the most fundamental and crucial mechanism for regulating and controlling protein viability and function (Keck et al. 2015). The potential phosphorylation sites suggested that the BEV GZ5 protein might play a significant role in the transcriptional regulation of proteins.

Most EPRVs in the host genome frequently led to early termination of the open reading frame (ORF) or translational frame shifts due to nucleotide substitutions, insertions, or deletions (Feschotte and Gilbert 2012). The integration of BSVs in the banana genome led to genetic mutations, rearrangements, and, in some cases, fragmentation of the entire banana genome triggered by natural selection or host interactions. D'Hont et al. (2012) reported the discovery of 24 BEVs integrations across ten chromosomes in *M. acuminate* subsp. DH-Pahang, all of which were highly recombinant and fragmented, thus failed to form active viruses. Similarly, Chabannes et al. (2021) discovered a significant genetic diversity of BEVs, which often showed translational frame shifts. In this study, compared with RT and RNase H of BSVs, BEV GZ5 had amino acid mutation and deletion. However, BEVs in this study and from GenBank showed partial common genes of RT and RNase H, which suggested that the integration of BEVs in the banana genomes were different. Valli et al. (2023) conducted extensive research on viral functional genes and their associated EPRVs in four representative eggplant genomes; they identified four distinct endogenous viral genomes and other associated EPRVs. Therefore, it was important to explore the distribution of the other functional genes of BSVs in the banana genome, which would provide further insight into the genetic evolutionary mechanism of BSVs and their interaction with bananas.

Conclusions

In this study, we identified two new BEVs, BEV GZ5 and GZ13. BEV GZ13 only had a partial RT/RNase H region, while BEV GZ5 had the complete genes of RT/RNase H, with a 340-bp IRS partially overlapping with RNase H showed upstream and downstream sequences of BEV GZ5. BEV GZ5 was present in different banana varieties with BB genomes, whereas BEV GZ13 was only present in the Dajiao genome (ABB). The new BEVs were different from BSVs; however, both of them shared partial RT/RNase H sequences with retroviruses. This suggested that the hallmark gene of RT and RNase H encoded by retroviruses had integrated in banana genomes, which would advance the co-evolution of BEVs, BSVs, and bananas. This research provides a theoretical foundation for further studying the integration mechanism of different BEVs in bananas.

Methods

DNA extractions, PCR cloning, and sequencing

During 2021–2022, the banana samples collected from Maoming in Guangdong Province, and Qinzhou, Wuzhou, and Yulin in Guangxi Province were stored at -80°C, respectively. To determine whether BSVs infected these samples, immunocapture-PCR (IC-PCR) was conducted on the total DNA (Le Provost et al. 2006). Total DNA from the bananas was extracted using the plant genomic DNA extraction kit following the instructions recommended by Tiangen (Beijing, China). The degenerate primers of Badv-RT were used for PCR amplification with high-fidelity enzymes from Vazyme (Nanjing, China), and the PCR reaction condition was performed as described by Rao et al. (2023). The PCR product was analyzed by electrophoresis in 1 % agarose gel, and the target fragments were purified by a gel recovery kit (Axygen, MA, USA). The purified PCR product of each DNA fragment was cloned into pMD 19T (Sanggon, Shanghai, China) and sequenced. Five clones of each positive sample were sequenced by Sangon (Shanghai, China).

Phylogenetic tree of BEVs

To determine the phylogenetic relationship of new BEVs with other BEVs and BSVs, BEV GZ5 and GZ13 in this study along with BEVs and BSVs from GenBank were analyzed. These sequences were aligned using Clustal W implemented in MEGA11 software (Tamura et al. 2021) and corrected manually when necessary. The phylogenetic tree was constructed by MEGA11 software (Tamura et al. 2021) using the Neighbor-joining (NJ) method with 1000 Bootstrap replicates (Saitou and Nei 1987).

Southern blot

To analyze whether BEV GZ5 and GZ13 were located in banana genomes, we conducted Southern blot to examine their endogeny. Genomic DNA were extracted from different banana varieties, such as Brazilian (AAA), Williams (AAA), Fenza 1 (ABBB), Guangfen 1 (ABB), Jinfen (ABB), and Dajiao (ABB) through the new plant genomic DNA extraction kit from Tiangen (Beijing, China) according to the manufacturer's instructions. The probe primers of GZ5TZ and GZ13TZ were designed based on the gene sequences of BEV GZ5 and GZ13 following the principles of probe design of Southern hybridization (Table 1), and synthesized through PCR amplification using the Roche PCR DIG Probe Synthesis kit (Roche, Switzerland). Southern blot analysis was performed as described previously (Rao et al. 2023). The hybridization signals were detected by the bio-macromolecular analyzer (Bio-Rad, USA).

FISH of BEV GZ5

To confirm the existence of endogenous BEV GZ5 in banana genomes, FISH was performed using the specific probes of BEV GZ5 on the young roots of Guangfen 1, the major cultivated varieties with BB genome in Guangdong Province. The specific DNA probe of BEV GZ5 was generated using the Roche PCR DIG Probe Synthesis kit (Roche, Basel, Switzerland). The banana chromosome preparations and FISH conducted on the young growing root tips of Guangfen 1 following the protocols described by Chabannes et al. (2013). Banana chromosomes were counterstained with 4 ', 6 ' -diamidino-2-phenylindole (DAPI) (Solarbio, Beijing, China) and mounted in anti-fade solution. The observations and images were made using a Leica inverted fluorescence biomicroscope DMi8 (Leica, Wetzlar, Germany).

Sequence amplification and analyses of upstream and downstream sequences of BEV GZ5

BEV GZ5 was located on chromosome 5 of *M. balbisiana* subsp. PKW (BB) with one locus (Rao et al. 2023). To analyze the sequence characteristics near BEV GZ5 in banana genomes, primers were designed based on the genome sequence of *M. balbisiana* subsp. PKW (BB) to amplify the upstream and downstream sequences of BEV GZ5 in the BB genomes of bananas (Table 3). PCR amplification of extracted DNA was carried out using high-fidelity enzymes following the protocol, 10 μL 2 × HiFi PCR StarMix, 1 μL 10 μM forward primer and 1 μL 10 μM of reverse primer, 7 μL ddH$_2$O, and 1 μL DNA. The PCR profile was as follows, 94℃ for 5 min; 35 cycles of 94℃ for 1 min, primer-specific annealing temperature for 1 min, and 72℃ for 1 min; and 72℃ for 10 min, then stored at 4 ℃. The target fragments were gel purified and cloned in pMD 19T vectors (Sanggon, Shanghai, China) according to the manufacturer's instructions. Plasmid DNA was extracted with plasmid DNA purification system (Axygen, USA) according to the manufacturer's instructions. The upstream and downstream sequences of BEV GZ5 were sequenced. Each clone was sequenced at least twice. Finally, the nucleotide sequences obtained in both directions of BEV GZ5 were assembled, respectively. After that, the nucleotide sequences and amino acid sequences of BEVs were analyzed by DNAMAN software. Additionally, the identification of motifs of amino acid and prediction of protein structural domains were carried out using MEME Suite (http://meme-suite.org/) and ProScan (http://www.ebi.ac.uk/interpro/search/sequence/), respectively.

Table 3 Primers in this study

Primers	Sequence(5′- 3′)	Application
GZ5TZ-F	ATGGACAATTACTTTAAAGGTACCG	Probe of BEV GZ5
GZ5TZ-R	CCATGCAGCCATCTGTTTCCA	
GZ13TZ-F	TGGACGAGTGTTTTAAAGGTACTG	Probe of BEV GZ13
GZ13TZ-R	ACTTCCAGTCTTGGTCATTCATCC	
GZ5YC1-F	AGTGATGACAGGTAGGTGAAGC	Amplification of fragment GZ5YC1
GZ5YC1-R	CCACTAGATGGGCACTTCAGAT	
GZ5YC2-F	AGTTTCTCCATCTGAAGTGCCCA	Amplification of fragment GZ5YC2
GZ5YC2-R	CCATGCAGCCATCTGTTTCCA	
GZ5YC3-F	ATGGACAATTACTTTAAAGGTACCG	Amplification of fragment GZ5YC3
GZ5YC3-R	CCATTGTTCCTTCGTCTTGTCG	
GZ5YC4-F	TGAGAGCCTAAGATGAAGTGGG	Amplification of fragment GZ5YC4
GZ5YC4-R	TCAATGAAGAATCAGTACCAGATG	

Bioinformatics analysis of BEV GZ5 protein

To investigate the characterization and function of BEV GZ5 protein, the physicochemical properties, amino acid sequence composition, hydrophobicity, transmembrane structure, signal peptide, subcellular localization, phosphorylation site, and secondary structures were predicted using bioinformatics softwares available online. The nucleotide sequences of BEV GZ5 were translated into amino acid sequences using Editseq software (V7.1.0.44). The physicochemical properties such as molecular weight, isoelectric point, and amino acid composition of BEV GZ5 and ten BSVs were analyzed using the ExPASy Proteomic (http://web.expasy.org/) and Wolfpsort (https://wolfpsort.hgc.jp). Furthermore, the potential subcellular localization of BEV GZ5 and the ten BSVs were predicted using Wolfpsort (https://wolfpsort.hgc.jp). The transmembrane structural domain of BEV GZ5 protein was analyzed using the TMHMM 2.0 Server (http://web.expasy.org/ProtScal). Additionally, the signal peptide region of the BEV GZ5 protein was predicted using the SignalP 4.1 server (http://www.cbs.dtu.dk/servers/SignalP). The hydrophilicity of the BEV GZ5 protein was analyzed using the online tool ProtScale (http://web.expasy.org/ProtScale). The potential phosphorylation sites of the BEV GZ5 protein were predicted using the online software NetPhos 3.1 Server (http://www.cbs.dtu.dk/services/NetPhos) with a threshold value of 0.5. The secondary structure of BEV GZ5 protein was predicted through SOPMA (http://npsa-prabi.ibcp.fr/cgi-bin/scpred_sopma.pl) after manual correction.

Abbreviations

Arg	Arginase
Asp	Aspartic acid
BEVs	Banana endogenous virus sequences
BSV	Banana streak virus
BSOLV	Banana streak OL virus
BSGFV	Banana streak GF virus
BSVNV	Banana streak VN virus
eBSV	Endogenous banana streak virus
EPRVs	Endogenous pararetrovirus sequences
FISH	Fluorescent *in situ* hybridization
GBV1	Grapevine badnavirus 1
Glu	Glutamate acid
ICTV	International Committee on Taxonomy of Viruses
IRS	Inverted repeat sequences
LTRs	Long terminal repeats
Lys	Lysine
ORF	Open reading frame
RT	Reverse transcriptase
RNase H	Ribonuclease H
TEs	Transposable elements
ZinF	Zinc finger protein

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s42483-024-00256-7.

Additional file 1: Table S1. Amino acid composition and content analyses (%) of BEV GZ5 and corresponding proteins of BSVs.

Additional file 2: Figure S1. Prediction of transmembrane region and signal peptide region of BEV GZ5 protein.

Acknowledgements

Not applicable.

Authors' contributions

HP and XQ designed experiments. HZ performed the experiments and analyzed the data. HZ and XQ wrote the manuscript, HP revised the manuscript. All authors read and approved the final manuscript.

Funding

This study was supported by National Key R&D Program of China (2019YFD1001800) and China Agriculture Research System of MOF and MARA (CARS-31).

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 23 January 2024 Accepted: 21 May 2024
Published online: 02 August 2024

References

Belser C, Istace B, Denis E, Dubarry M, Baurens FC, Falentin C, et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nat Plants. 2018;4:879–87. https://doi.org/10.1038/s41477-018-0289-4.

Belser C, Baurens FC, Noel B, Martin G, Cruaud C, Istace B, et al. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. Commun Biol. 2021;4:1047. https://doi.org/10.1038/s42003-021-02559-3.

Chabannes M, Baurens FC, Duroy PO, Bocs S, Vernerey MS, Rodier-Goud M, et al. Three infectious viral species lying in wait in the banana genome. J Virol. 2013;87:8624–37. https://doi.org/10.1128/JVI.00899-13.

Chabannes M, Gabriel M, Aksa A, Galzi S, Dufayard JF, Iskra-Caruana ML, et al. Badnaviruses and banana genomes: a long association sheds light on *Musa* phylogeny and origin. Mol Plant Pathol. 2021;22:216–30. https://doi.org/10.1111/mpp.13019.

Chen S, Kishima Y. Endogenous pararetroviruses in rice genomes as a fossil record useful for the emerging field of palaeovirology. Mol Plant Pathol. 2016;17:1317–20. https://doi.org/10.1111/mpp.12490.

Chen S, Zheng H, Kishima Y. Genomic fossils reveal adaptation of non-autonomous pararetroviruses driven by concerted evolution of noncoding regulatory sequences. PLoS Pathog. 2017;13:e1006413. https://doi.org/10.1371/journal.ppat.1006413.

Côte FX, Galzi S, Folliot M, Lamagnere Y, Teycheney PY, Iskra-Caruana ML. 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. Mol Plant Pathol. 2010;11:137–44. https://doi.org/10.1111/j.1364-3703.2009.00583.x.

Davey MW, Gudimella R, Harikrishna JA, Sin LW, Khalid N, Keulemans J. A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter- and intra-specific *Musa* hybrids. BMC Genomics. 2013;14:683. https://doi.org/10.1186/1471-2164-14-683.

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature. 2012;488:213–7. https://doi.org/10.1038/nature11241.

Diop SI, Geering A, Alfama-Depauw F, Loaec M, Teycheney PY, Maumus F. Tracheophyte genomes keep track of the deep evolution of the *Caulimoviridae*. Sci Rep. 2018;8:572. https://doi.org/10.1038/s41598-017-16399-x.

Feschotte C, Gilbert C. Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet. 2012;13:283–96. https://doi.org/10.1038/nrg3199.

Gayral P, Iskra-Caruana ML. Phylogeny of banana streak virus reveals recent and repetitive endogenization in the genome of its banana host (*Musa* sp.). J Mol Evol. 2009;69:65–80. https://doi.org/10.1007/s00239-009-9253-2.

Geering A, Olszewski NE, Harper G, Lockhart B, Hull R, Thomas JE. Banana contains a diverse array of endogenous badnaviruses. J Gen Virol. 2005;86:511–20. https://doi.org/10.1099/vir.0.80261-0.

Geering AD, Scharaschkin T, Teycheney PY. The classification and nomenclature of endogenous viruses of the family *Caulimoviridae*. Arch Virol. 2010;155:123–31. https://doi.org/10.1007/s00705-009-0488-4.

Geering AD, Parry JN, Thomas JE. Complete genome sequence of a novel badnavirus, banana streak IM virus. Arch Virol. 2011;156:733–7. https://doi.org/10.1007/s00705-011-0946-7.

Geering AD, Maumus F, Copetti D, Choisne N, Zwickl DJ, Zytnicki M, et al. Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. Nat Commun. 2014;5:5269. https://doi.org/10.1038/ncomms6269.

Harper G, Hull R, Lockhart B, Olszewski N. Viral sequences integrated into plant genomes. Annu Rev Phytopathol. 2002;40:119–36. https://doi.org/10.1146/annurev.phyto.40.120301.105642.

Harper G, Hart D, Moult S, Hull R, Geering A, Thomas J. The diversity of banana streak virus isolates in Uganda. Arch Virol. 2005;150:2407–20. https://doi.org/10.1007/s00705-005-0610-1.

Huang Y, Li Y. Secondary siRNAs rescue virus-infected plants. Nat Plants. 2018;4:136–7. https://doi.org/10.1038/s41477-018-0118-9.

Hughes JF, Coffin JM. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nature Genet. 2001;29:487–9. https://doi.org/10.1038/ng775.

Hull R, Harper G, Lockhart B. Viral sequences integrated into plant genomes. Trends Plant Sci. 2000;5:362–5. https://doi.org/10.1016/s1360-1385(00)01723-4.

Iskra-Caruana ML, Baurens FC, Gayral P, Chabannes M. A four-partner plant-virus interaction: enemies can also come from within. Mol Plant-Microbe Interact. 2010;23:1394–402. https://doi.org/10.1094/MPMI-05-10-0107.

Iskra-Caruana ML, Duroy PO, Chabannes M, Muller E. The common evolutionary history of badnaviruses and banana. Infect Genet Evol. 2014;21:83–9. https://doi.org/10.1016/j.meegid.2013.10.013.

Jakowitsch J, Mette MF, van Der Winden J, Matzke MA, Matzke AJ. Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. Proc Natl Acad Sci USA. 1999;96:13241–6. https://doi.org/10.1073/pnas.96.23.13241.

James AP, Geijskes RJ, Dale JL, Harding RM. Development of a novel rolling-circle amplification technique to detect banana streak virus that also discriminates between integrated and episomal virus sequences. Plant Dis. 2011;95:57–62. https://doi.org/10.1094/PDIS-07-10-0519.

Jaufeerally-Fakim Y, Khorugdharry A, Harper G. Genetic variants of banana streak virus in Mauritius. Virus Res. 2006;115:91–8. https://doi.org/10.1016/j.virusres.2005.06.015.

Katzourakis A, Tristem M, Pybus OG, Gifford RJ. Discovery and analysis of the first endogenous lentivirus. Proc Natl Acad Sci USA. 2007;104:6261–5. https://doi.org/10.1073/pnas.0700471104.

Keck F, Ataey P, Amaya M, Bailey C, Narayanan A. Phosphorylation of single stranded RNA virus proteins and potential for novel therapeutic strategies. Viruses. 2015;7:5257–73. https://doi.org/10.3390/v7102872.

Kunii M, Kanda M, Nagano H, Uyeda I, Kishima Y, Sano Y. Reconstruction of putative DNA virus from endogenous rice tungro bacilliform virus-like sequences in the rice genome: implications for integration and evolution. BMC Genomics. 2004;5:80. https://doi.org/10.1186/1471-2164-5-80.

Le Provost G, Iskra-Caruana ML, Acina I, Teycheney PY. Improved detection of episomal banana streak viruses by multiplex immunocapture PCR. J Virol Methods. 2006;137:7–13. https://doi.org/10.1016/j.jviromet.2006.05.021.

Lheureux F, Laboureau N, Muller E, Lockhart B, Iskra-Caruana ML. Brief report molecular characterization of banana streak acuminata Vietnam virus isolated from *Musa acuminata* siamea (banana cultivar). Arch Virol. 2007;152:1409–16. https://doi.org/10.1007/s00705-007-0946-9.

Li WL, Yu NT, Wang JH, Li JC, Liu ZX. The complete genome of banana streak GF virus Yunnan isolate infecting Cavendish *Musa* AAA group in China. Peer J. 2020;8:e8459. https://doi.org/10.7717/peerj.8459.

Liu R, Koyanagi KO, Chen S, Kishima Y. Evolutionary force of AT-rich repeats to trap genomic and episomal DNAs into the rice genome: lessons from endogenous pararetrovirus. Plant J. 2012;2(5):817–28. https://doi.org/10.1111/tpj.12002.

Mette MF, Kanno T, Aufsatz W, Jakowitsch J, van der Winden J, Matzke MA, et al. Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. Embo J. 2002;21:461–9. https://doi.org/10.1093/emboj/21.3.461.

Ndowora T, Dahal G, LaFleur D, Harper G, Hull R, Olszewski NE, et al. Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. Virology. 1999;255:214–20. https://doi.org/10.1006/viro.1998.9582.

Perrier X, De Langhe E, Donohue M, Lentfer C, Vrydaghs L, Bakry F, et al. Multidisciplinary perspectives on banana (*Musa* spp.) domestication. Proc Natl Acad Sci USA. 2011;108:11311–8. https://doi.org/10.1073/pnas.1102001108.

Rao X, Chen H, Lu Y, Liu R, Li H. Distribution and location of BEVs in different genotypes of bananas reveal the coevolution of BSVs and bananas. Int J Mol Sci. 2023;24(23):17064. https://doi.org/10.3390/ijms242317064.

Richert-Poggeler KR, Vijverberg K, Alisawi O, Chofong GN, Heslop-Harrison J, Schwarzacher T. Participation of multifunctional RNA in replication, recombination and regulation of endogenous plant pararetroviruses (EPRVs). Front Plant Sci. 2021;12:689307. https://doi.org/10.3389/fpls.2021.689307.

Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4:406–25. https://doi.org/10.1093/oxfordjournals.molbev.a040454.

Schmidt N, Seibt KM, Weber B, Schwarzacher T, Schmidt T, Heitkam T. Broken, silent, and in hiding: tamed endogenous pararetroviruses escape elimination from the genome of sugar beet (*Beta vulgaris*). Ann Bot. 2021;128:281–99. https://doi.org/10.1093/aob/mcab042.

Staginnus C, Gregor W, Mette MF, Teo CH, Borroto-Fernandez EG, Machado ML, et al. Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. BMC Plant Biol. 2007;7:24. https://doi.org/10.1186/1471-2229-7-24.

Staginnus C, Iskra-Caruana ML, Lockhart B, Hohn T, Richert-Poggeler KR. Suggestions for a nomenclature of endogenous pararetroviral sequences in plants. Arch Virol. 2009;154:1189–93. https://doi.org/10.1007/s00705-009-0412-y.

Tamura K, Stecher G, Kumar S. MEGA11: molecular evolutionary genetics analysis version 11. Mol Biol Evol. 2021;38:3022–7. https://doi.org/10.1093/molbev/msab120.

Valli AA, Gonzalo-Magro I, Sanchez DH. Rearranged endogenized plant pararetroviruses as evidence of heritable RNA-based immunity. Mol Biol Evol. 2023;40:msac240. https://doi.org/10.1093/molbev/msac240.

van der Loo W, Abrantes J, Esteves PJ. Sharing of endogenous lentiviral gene fragments among leporid lineages separated for more than 12 million years. J Virol. 2009;83:2386–8. https://doi.org/10.1128/JVI.01116-08.

Vassilieff H, Geering A, Choisne N, Teycheney PY, Maumus F. Endogenous caulimovirids: fossils, zombies, and living in plant genomes. Biomolecules. 2023;13:1069. https://doi.org/10.3390/biom13071069.

White SE, Habera LF, Wessler SR. Retrotransposons in the flanking regions of normal plant genes: a role for copia-like elements in the evolution of gene structure and expression. Proc Natl Acad Sci USA. 1994;91:11792–6. https://doi.org/10.1073/pnas.91.25.11792.

Xiong Y, Eickbush TH. Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol Biol Evol. 1988;5(6):675–90. https://doi.org/10.1093/oxfordjournals.molbev.a040521.

Xiong Y, Eickbush TH. Origin and evolution of retroelements based upon their reverse transcriptase sequences. Embo J. 1990;9:3353–62. https://doi.org/10.1002/j.1460-2075.1990.tb07536.x.

Yang J, Liu D, Wang X, Ji C, Cheng F, Liu B, et al. The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. Nature Genet. 2016;48:1225–32. https://doi.org/10.1038/ng.3657.

Yu H, Wang X, Lu Z, Xu Y, Deng X, Xu Q. Endogenous pararetrovirus sequences are widely present in *Citrinae* genomes. Virus Res. 2019;262:48–53. https://doi.org/10.1016/j.virusres.2018.05.018.

Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nat Biotechnol. 2015;33:531–7. https://doi.org/10.1038/nbt.3207.