Characterization of Bundaberg Bee Virus 4 Genome Identified in Hawaii

James Doherty, Vedbar Khadka, Beatrice Kondo, and Jun Panee

ABSTRACT

To identify genetic markers of different cultivars of Stevia rebaudiana (stevia), leaf samples from twenty-three genetically unique stevia plants were subjected to RNA sequencing. Unexpectedly, transcriptome annotation revealed a ‘viral polyprotein’ contig, which was identified as the Bundaberg Bee Virus 4 (BBV4), a virus that has not been found in Hawaii before. BBV4 expression was found in twelve samples, and the expression levels ranged from 0.08-339.18 transcripts per million (TPM). A ‘BBV4-Hawaii’ consensus sequence was generated from 227,427 reads and shares 99.7% identity with the BBV4 reference sample from Australia. A phylogenetic tree constructed using three viral domains placed BBV4 in a monophyletic clade with other members of Iflaviridae, supporting an Iflaviridae classification. A BBV4 intergenic region (IGR) was found to contain a ‘slippery sequence’ within a region that is predicted to contain the unique structure required for Programmed Ribosomal Shifting (PRS). A homologous comparison of the BBV4 IGR to two other Picornavirales that utilize alternative translation revealed an area of overlap in the analogous regions. These results suggest that BBV4 may utilize alternative translation and that its IGR is involved in the process. This study also demonstrates the utility of repurposing large Next Generation Sequence (NGS) datasets with sufficiently deep coverage for viral genome identification, quantification, and characterization. The results herein represent the first full capture of the BBV4 genomic sequence and the second reported observation of the virus worldwide.

Keywords: Alternative translation, data mining, RNA virus, slippery sequence.

I. INTRODUCTION

In most environments, viruses outnumber living cells, but despite this ubiquitous presence, relatively few viruses have been cultivated and identified in a laboratory setting (Roux et al., 2019). In recent years, Next Generation Sequencing (NGS) datasets derived from biological and environmental samples have revealed the presence of many novel RNA viruses. As a result, viruses discovered from NGS data now vastly outnumber those identified from viral cultivation in laboratory settings (Kolundzija et al., 2022). However, many of the novel viruses discovered from NGS data are uncharacterized.

Bundaberg Bee Virus 4 (BBV4) is one such uncharacterized virus discovered in NGS data. The BBV4 sequence was discovered in a 2018 metagenomic analysis and its origin is believed to be from a honeybee hive in Australia (Roberts et al., 2018) and, to date, the virus has not been observed elsewhere. In Hawaii, native bees of the genus *Hylaeus* are the most important pollinators of endemic Hawaiian plants (Cortina et al., 2019). Although *Apis mellifera*, the European honeybee, is non-native to Hawaii, it is nonetheless the second most important pollinator in the island ecosystem (Cortina et al., 2019). Bee viruses such as Deformed Wing Virus (DMV) are known to infect and damage both types of bees (Santamaria, 2020), but the presence and impact of BBV4 on the Hawaiian island ecosystem is unknown.

We have previously generated the transcriptome from twenty-three *Stevia rebaudiana* (stevia) plants grown in Hawaii. During data annotation, we unexpectedly identified segments of the BBV4 genome in the collection of transcripts. We hypothesized that the deep sequencing and broad coverage of the transcriptome data would enable us to capture the full sequence of the BBV4 genome. Through de novo transcriptome assembly and annotation, we generated the consensus sequence of BBV4 from twelve BBV4-infected plants and quantified the expression levels of BBV4 and compared it those of some endogenous plant genes. We further characterized BBV4 viral domains and identified an intergenic region (IGR) of BBV4 that may play a role in translation. This study not only reports the presence of BBV4 in Hawaii for the first time, but it also demonstrates the utility of repurposing large NGS datasets with deep coverage for virus identification, quantification, and characterization.

Published Online: July 13, 2023
ISSN: 2684-5199
DOI: 10.24018/ejbio.2023.4.3.461

J. Doherty*
Department of Molecular Biology and Bioengineering, University of Hawaii at Manoa, Honolulu, HI, USA.
(e-mail: jdoher9@hawaii.edu)

V. Khadka
Department of Quantitative Health Services, John A. Burns School of Medicine, University of Hawaii, Honolulu, HI, USA.

B. Kondo
Fischell Department of Bioengineering, University of Maryland, College Park, MD, USA.

J. Panee
Department of Cell and Molecular Biology, John A. Burns School of Medicine, University of Hawaii at Manoa, HI, USA.

*Corresponding Author
II. MATERIALS AND METHODS

A. Sample Collection and RNA Extraction

Leaf samples were collected in three batches from twenty-three genetically distinct S. rebaudiana plants grown in Kunia, Hawaii. Batch one plants were derived from tissue-culture and grown indoors in a sterile room but were fully formed with secondary branching and mature roots at the time of leaf collection. The second and third batches of plants also originated from tissue-culture propagation but were planted in containers of potting soil and grown outside in two different open-air greenhouses, which were on the same parcel of land, but separated by approximately 130 meters. Importantly, leaf samples from batch two plants were collected in May 2019, and leaf samples from batch three plants were collected five months later in October 2019. In contrast to healthy-looking batch one and batch two plants, plants from the third batch (namely: A104, A106, A109, A115, A256, G1-3, S1-2, T54, T55, T58, T60, and A142) had visible signs of biotic stress, presenting with leaf yellowing that ranged from mild to severe.

Collected leaf samples were immediately stored in liquid nitrogen. Frozen samples were then pulverized using a mortar and pestle, and RNA was extracted using the Qiagen RNAeasy kit according to the manufacturer’s specifications. Total RNA was extracted and measured for concentration and quality using the Nanodrop. RNA extraction samples below an A260:280 ratio of 1.8 were rejected.

B. cDNA Library Preparation and Sequencing

High-quality RNA samples were shipped on dry ice to Novogene (Sacramento, CA), where the samples were further tested for degradation. Only RNA samples with an RNA Integrity Number (RIN) above 6.8 were permitted for further analysis. The twenty-three RNA samples that passed the extensive quality control were then subjected to mRNA enrichment and Poly (A) capture. RNA was fragmented via sonication, and a cDNA library was constructed via reverse transcription. Following sequencing on the NovaSeq, reads from the twenty-three samples were subjected to quality scoring and trimming via BBDuk (v.38.84) (Bushnell, 2014) software, and reads with a quality score below 30 were removed.

C. De Novo Transcriptome Assembly

The software BBNorm (v.38.84) (Bushnell, 2014) was used to normalize reads to an average coverage depth of 50X. Normalized reads were then subject to de novo transcriptome assembly using Geneious (Geneious Prime v. 2022.1.1) assembler and contigs were annotated using BLAST2GO (v.6.0) (Gotz et al., 2008). The Geneious assembly was run with the following parameters: Minimum overlap = 25, Minimum overlap identity = 80%, Word length = 24, Index word length = 14, Ignore words repeated more than 100 times, Reanalyze threshold = 24, Maximum mismatches per reads = 15%, and Maximum ambiguity = 4.

D. Phylogenetics

To infer the genetic relatedness of BBV4 to its Picornavirales relatives, fifty-six representative species comprising all eight families of the order Picornavirales were compared. Although the same viral protein domains are present in species of each family, the order of the viral domains is not preserved among families. To circumvent this complication, three protein viral domains-Helicase, 3C Protease, and RNA-dependent RNA Polymerase (RdRP)-were extracted from each taxon and then aligned using MAFFT (v. 7.490) with default parameters. For each respective domain, the sequences were then trimmed from the first conserved residue to the last, and then the domains were concatenated in the same order for all taxa. The concatenated sequences were subjected to multiple alignment a second time. The resultant alignment was imported to jModelTest2.1.10 (Darriba et al., 2012) and run with default parameters to select the best-fit model. FastTree2 (Price et al., 2010) software was used with the best-fit parameters selected from jModelTest2.1.10 results to generate a Maximum Likelihood (ML) tree estimation using the Potato Y Virus as an outgroup. Potato Y Virus was chosen as an outgroup to root the tree because it belongs to the same phylum as Picornavirales, but it is located outside of the Picornavirales order. Branch support values were calculated by FastTree using the Shimodaira-Hasegawa (SH) test in units of SH between 0 and 1. The results were visualized using Geneious (Geneious Prime v.2022.1.1) software.

E. RNA Structure

The full-length RNA genomic sequence of BBV4 was uploaded to the RNAfold web server using default parameters. The results were visualized using the software Forna (Kerpedjiev et al., 2015).

III. RESULTS AND DISCUSSION

A. Identification and Quantification of BBV4

During the transcriptome annotation process of S. rebaudiana RNAs, CONTIG 22 was annotated as a viral polyprotein. To confirm its identity, its sequence was used as a BLASTn query against the NCBI non-redundant database. The BLASTn query returned a matching sequence with a 99.7% identity score to the reference sequence: MG995705, Bundaberg Bee Virus 4 (BBV4).

The reads for each respective plant sample were individually mapped to the constructed de novo transcriptome, which included CONTIG 22 aka BBV4-Hawaii. In total, 227,427 reads from twelve unique plant samples were mapped to the BBV4-Hawaii contig with an average coverage depth of 4029X. Because BBV4 RNA is exogenous, zero expression is expected in non-infected plants. Therefore, samples that contained a read count above zero were considered positive. BBV4 was detected in all twelve samples collected in Batch 3 (Table 1) but not in samples collected in Batches 1 and 2. In contrast, the endogenous housekeeping genes ACT1, GADPH and PAL are constitutively expressed in all cells. When plant samples were pooled into two groups: BBV4-infected plants vs. non-infected plants, the expression levels of the housekeeping genes were not statistically significant different.

B. The Genomic Layout and Structure of BBV4

Like all Picornavirales, the genomic sequence of BBV4 commences with a long 5’ untranslated region (5’ UTR) that includes an internal ribosome entry site (IRES) that initiates translation in a cap-independent manner (Belsham &
Sonenberg, 1996). Like all other Ifflaviridae, BBV4 contains a long open reading frame (ORF) and begins translation with its structural proteins and concludes translation with its RdRP (Le Gall et al., 2008) (Valles et al., 2017). The BBV4 ORF required manual annotation owing to its 5’ UTR containing sixteen false-start AUG codons that appear before the actual initiation codon. In BBV4, the ORF begins at position 855 and ends at position 8418 (Fig. 1). In members of Ifflaviridae, the encoded structural proteins (VP1-4) are proximal to the 5’ UTR and flanked by an IGR. Translation continues with the non-structural proteins-encoded in the order of Helicase, Protease, and RdRP-and the transcript ends with a poly-A tail that concludes the 3’ UTR. The BBV4 gene adopts a similar structure. The first BBV4 structural viral protein, VP2, marks the boundary between the 5’ UTR and flanked by an IGR. Translation continues with the encoded structural proteins (VP1-4) and ends at position 8418 -start AUG codons that appear before the actual initiation codon. In BBV4, the ORF begins at position 855 and ends at position 8418 (Fig. 1). In members of Ifflaviridae, the encoded structural proteins (VP1-4) are proximal to the 5’ UTR and flanked by an IGR. Translation continues with the non-structural proteins-encoded in the order of Helicase, Protease, and RdRP-and the transcript ends with a poly-A tail that concludes the 3’ UTR. The BBV4 gene adopts a similar structure. The first BBV4 structural viral protein, VP2, marks the boundary between the 5’ UTR and flanked by an IGR. Translation continues with the encoded structural proteins (VP1-4) and ends at position 8418.

C. Secondary Structure of BBV4
The secondary structure of BBV4 RNA (Fig. 2) was predicted using the software RNAfold and is based upon free energy calculations (Lorenz et al., 2011). In Fig. 2, the first IGR is followed by a slippery sequence, which has the effect of ballooning the single-stranded RNA in that section. The slippery sequence is flanked in both directions by the type of secondary structure required for Programmed Ribosomal Shifting (PRS) (Brierley et al., 1989) (Namy et al., 2006).

For a slippery sequence to function as a PRS site, it needs to be flanked on the 3’ side by a significant secondary structure and a small spacer of approximately 15nt, which separates the slippery sequence from the structure (Brierley et al., 1989) (Namy et al., 2006). The predicted secondary structure of BBV4 RNA meets these requirements and likely fosters an environment suitable for PRS.

### Table: The Expression Levels of Lundaberg Bee Virus 4 (BBV4) and Three Endogenous Housekeeping Genes in Twenty-Three Stevia Rebaudiana Plants

<table>
<thead>
<tr>
<th>Sample; Batch</th>
<th>BBV4</th>
<th>ACT1</th>
<th>GADPH</th>
<th>PAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>A10; 1</td>
<td>0</td>
<td>4.22</td>
<td>488.07</td>
<td>1558.46</td>
</tr>
<tr>
<td>A15; 1</td>
<td>0</td>
<td>4.94</td>
<td>543.09</td>
<td>611.98</td>
</tr>
<tr>
<td>K1; 1</td>
<td>0</td>
<td>4.55</td>
<td>463.49</td>
<td>1491.79</td>
</tr>
<tr>
<td>A1; 2</td>
<td>0</td>
<td>0.93</td>
<td>813.08</td>
<td>568.12</td>
</tr>
<tr>
<td>A3; 2</td>
<td>0</td>
<td>3.70</td>
<td>446.47</td>
<td>446.85</td>
</tr>
<tr>
<td>A5; 2</td>
<td>0</td>
<td>3.06</td>
<td>347.75</td>
<td>517.44</td>
</tr>
<tr>
<td>A6; 2</td>
<td>0</td>
<td>57.04</td>
<td>666.01</td>
<td>383.93</td>
</tr>
<tr>
<td>A7; 2</td>
<td>0</td>
<td>1.25</td>
<td>777.93</td>
<td>113.44</td>
</tr>
<tr>
<td>A8; 2</td>
<td>0</td>
<td>1.08</td>
<td>868.99</td>
<td>585.37</td>
</tr>
<tr>
<td>A11; 2</td>
<td>0</td>
<td>11.78</td>
<td>932.76</td>
<td>704.56</td>
</tr>
<tr>
<td>H5; 2</td>
<td>0</td>
<td>3.38</td>
<td>502.58</td>
<td>1167.43</td>
</tr>
<tr>
<td>A104; 3</td>
<td>258.63</td>
<td>2.73</td>
<td>563.33</td>
<td>274.26</td>
</tr>
<tr>
<td>A106; 3</td>
<td>47.77</td>
<td>7.46</td>
<td>427.44</td>
<td>811.57</td>
</tr>
<tr>
<td>A109; 3</td>
<td>0.75</td>
<td>8.47</td>
<td>601.01</td>
<td>589.26</td>
</tr>
<tr>
<td>A115; 3</td>
<td>102.49</td>
<td>6.29</td>
<td>446.95</td>
<td>1189.93</td>
</tr>
<tr>
<td>A142; 3</td>
<td>339.18</td>
<td>7.28</td>
<td>640.97</td>
<td>1299.86</td>
</tr>
<tr>
<td>A256; 3</td>
<td>0.08</td>
<td>6.32</td>
<td>963.16</td>
<td>917.34</td>
</tr>
<tr>
<td>G1; 3</td>
<td>5.47</td>
<td>2.84</td>
<td>664.79</td>
<td>1524.40</td>
</tr>
<tr>
<td>S12; 3</td>
<td>108.21</td>
<td>2.08</td>
<td>589.42</td>
<td>320.01</td>
</tr>
<tr>
<td>T54; 3</td>
<td>0.79</td>
<td>3.19</td>
<td>848.54</td>
<td>510.51</td>
</tr>
<tr>
<td>T55; 3</td>
<td>8.72</td>
<td>3.17</td>
<td>620.61</td>
<td>647.34</td>
</tr>
<tr>
<td>T58; 3</td>
<td>1.37</td>
<td>5.91</td>
<td>849.51</td>
<td>346.01</td>
</tr>
<tr>
<td>T60; 3</td>
<td>0.14</td>
<td>2.02</td>
<td>871.01</td>
<td>229.44</td>
</tr>
<tr>
<td>p-value</td>
<td>1.43E-10</td>
<td>0.393599</td>
<td>0.08785</td>
<td>0.851738</td>
</tr>
</tbody>
</table>

Note: Unit of transcripts per million (TPM) is used to quantify the gene expression levels. ACT1, Actin; GADPH, Glyceraldehyde 3-phosphate dehydrogenase; and PAL, Phenylalanine ammonia lyase.
Notably, the BBV4 slippery sequence UUUAAAG is located 12 bases downstream of the overlap region in Fig. 3. These findings, especially when considered together, suggest that this region may be involved in BBV4 translation.

D. A homologous Comparison Suggests A General Function of the BBV4 IGR

Fig. 3 shows a multiple alignment of the IGRs of BBV4, Canine Picocidistrovirus (CPDV) and Cricket Paralysis Virus (CrPV). The IGR of BBV4 had 53-57% sequence homology to the IRES regions of CPDV and CrPV. In CrPV and CPDV, this overlap region is associated with a second, internal IRES, which initiates Cap-independent translation (Woo et al., 2012). The alignment of the three IGRs revealed an area of overlap that occurs proximally to the BBV4 slippery sequence.

CrPV belongs to the genus Dicistroviridae, whose members are all known to dicistronically translate their genome. However, CPDV is exceptional in the sense that it belongs to Picornaviridae, whose members, like those of flaviviridae, are supposed to be monocistronic. Despite this classification, the CPDV IGR was experimentally proven to contain a second IRES site capable of independent translation (Woo et al., 2012).

The multiple alignment of the overlap IGR region of BBV4, CrPV and CPDV (three viruses from three different Picornavirales families) revealed significant homology to areas with a known function in the two dicistronic viruses, CrPV and CPDV, which may offer a clue as to the function of this region in BBV4. In many Picornavirales, the IGR between structural and non-structural proteins appears to be important in translation, and differences and permutations in this region appear to imbue different Picornavirales viruses with varying aspects of alternative translation. To bolster this idea, consider that the Foot and Mouth Disease Virus (FMDV), another Picornavirales, also does something unusual at the terminus of the genomic region that encodes its last structural protein (VP) that precedes the IGR. In a process distinct from either the IRES strategy or PRS, in FMDV, the last glycine codon of the VP domain is unconventionally recoded as a stop codon, effectively truncating the protein product in a process deemed as the StopGo translation (Atkins et al., 2007).

E. Small RNA Viruses Maximize Proteome Diversity to Compensate for Small Genomes and Minimal Genes

Flaviviridae and most other Picornaviruses are described in scientific literature as having a single, long ORF (Valles et al., 2017). The word ‘long’ is an interesting qualifier because one long ORF, which is immediately apparent, may also contain other smaller and less obvious ORFs. Viruses with compact genomes utilize diverse translational strategies to compensate for their small genomes and to compete and meet rapidly changing conditions presented in host environments (Jaafar & Kieft, 2019). Most documented observations of ribosomal framework occur in viral genomic sequences that contain two obvious, long ORFs (Firth et al., 2008) (Loughran et al., 2011). But this association may result from simple bias because long ORFs are easily detected by sequence analysis and predictive software (Loughran et al., 2011), whereas hidden, overlapping, and small ORFs are not obvious and are easily missed.

To generate the requisite protein diversity necessary to compete with complex host cells, it is well established that small RNA viruses employ alternative translation strategies, such as recoding and ribosomal shifting, and some even utilize subgenomic RNA transcription and translation (Valles et al., 2014). Using short, overlapping, and hidden reading frames is part of a multi-faceted effort to maximize the proteome and compensate for a small genome (Chung et al., 2008). In BBV4, the proposed ribosomal frame shift would have the effect of introducing a stop codon shortly after the frame shift and truncating the protein. In this way, PRS could be utilized to alter the ratio of structural proteins (VPs) to non-structural proteins, which are known to be required in unequal proportion in other Picornavirales (Loughran et al., 2011).

---

Fig. 3. Homological comparison of the intergenic region (IGR) of BBV4 to the internal IRES of Cricket Paralysis Virus (CrPV) and Canine Picocidistrovirus Virus (CPDV). CPDV and CrPV internal IRES regions were aligned to the corresponding IGR of BBV4. Agreement to the consensus sequence are highlighted in colors, the disagreements are greyed out, and the gaps are filled in with dash signs.

DOI: http://dx.doi.org/10.24018/ejbio.2023.4.3.461
F. Phylogenetic Placement of BBV4 as Support for Iflaviridae Classification

To evaluate the phylogenetic relationship between BBV4 and other viruses, the genomic sequences of three domains (RdRP, Helicase, and 3C Protease) of BBV4 were used in the Maximum Likelihood (ML) phylogenetic tree approximation generated with the software FastTree2 (Price et al., 2010) (Fig. 4). Fifty-six Picornavirales species from all eight Picornavirales families and a distantly related Potato Y Virus were allowed to sort using the JTT theory of evolution. In the resultant phylogenetic tree, taxa from the Picornavirales family Iflaviridae form a monophyletic clade, which contains BBV4-Hawaii.

Evidence from the phylogenetic tree estimation (Fig. 4) in conjunction with the annotated order of encoded BBV4 viral domains (Fig. 1) provide supporting evidence to classify BBV4 as an Iflaviridae virus.

IV. CONCLUDING REMARKS AND FUTURE DIRECTIONS

The discovery of BBV4 in *S. rebaudiana* leaf samples in Oahu, Hawaii is the second known report of the virus since its 2018 discovery in Australia, and this discovery designates BBV4 as a global virus. Aside from its sequence and gene structure, very little is known about the virus and its host range or even its potential for causing a disease state. Although Iflaviridae viruses are assumed to infect arthropods, there are notable exceptions (Saqib et al., 2015) and the host range of BBV4 has yet to be determined. BBV4 was found in samples at relatively high levels, but it is unclear if *S. rebaudiana* is a natural host, an intermediate host, or merely a vector for transmission.

The primary limitation of this work is the original study design, which was intended for the purpose of generating a catalog of *S. rebaudiana* mRNAs derived from. Although a differential gene expression analysis of BBV4-positive samples vs BBV4-negative samples yielded interesting results, uncontrolled independent variables hindered meaningful investigation of BBV4-associated changes in gene expression. Conversely, the same study design intended for transcriptome generation provided the deep coverage necessary to fully capture the BBV4 genomic sequence, which might have otherwise been missed in a background of endogenous plant genes.

The novel appearance of BBV4 in Hawaii may interest ecological epidemiologists to determine if the virus threatens the island ecosystem and the extent of its spread. Many exciting molecular biology discoveries, including the IRES element, were first observed in Picornaviruses (Atkins et al., 2007). Because of its small size and compact genome, BBV4 may become a good model for investigating alternative translations employed by similarly small viruses. Further investigation of BBV4 may provide insight into the interplay between ribosomal kinetics and viral sequence recoding and how this interplay can expand the proteome without enlarging the genome.

**CONFlict OF INTEREST**

Authors declare that they do not have any conflict of interest.
REFERENCES


