Chapter 15

Universal Primers for Detection of Novel Plant Capsid-Less Viruses: Papaya Umbra-like Viruses as Example

Jorge H. Ramirez-Prado and Luisa A. Lopez-Ochoa

Abstract

For diagnosis of positive-sense single-stranded RNA viruses, primers are usually raised against the sequence encoding capsid proteins, since structural proteins are more conserved. This chapter focuses on the design of primers for a group of novel viruses lacking a capsid, known as papaya Umbra-like viruses (unassigned genus) associated with Papaya Sticky Disease, which represent a threat to papaya production. Based on sequence alignments of a region encoding the RNA-dependent RNA Polymerase, universal primers to detect all the known viruses from four countries are proposed. The Forward universal primer can be used in combination with clade- and subclade-specific primers for rapid virus identification. We walk the reader through downloading sequences from nucleotide databases, doing sequence alignments and phylogenetic tree construction to identify conserved and variable regions as valid primer targets; we also show how to design and analyze the primers.

Key words RdRP, RT-PCR, Papaya umbra-like viruses, Papaya Sticky Disease (PDS), Meleira disease, Virus detection

Abbreviations

bp	Base pair
BLAST	Basic Local Alignment Search Tool
CDS	Coding sequence
dsRNA	Double-stranded RNA
Indels	Insertion–deletions
kb	Kilobases
kcal/mol	Kilocalorie per mole
ML	Maximum Likelihood
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
NJ	Neighbor-Joining
PMeV	Papaya meleira virus
PMeV-1	Papaya meleira virus 1
PMeV-2	Papaya meleira virus 2
PMeV-Mx	Papaya meleira virus-Mexican variant

Chhandak Basu (ed.), *PCR Primer Design*, Methods in Molecular Biology, vol. 2392, https://doi.org/10.1007/978-1-0716-1799-1_15,

PpVQ	Papaya virus Q
PRSV-P	Papaya ringspot virus type P
PSD	Papaya Sticky Disease
PCR	Polymerase chain reaction
qPCR	Quantitative PCR
RdRP	RNA-dependent RNA Polymerase
(+) RNA	Positive-sense single-stranded RNA
RT-PCR	Reverse transcription polymerase chain reaction
T_m	Melting temperature

1 Introduction

When new disease symptoms or variation of symptoms of a known disease are first discovered on a crop, it is common to use polymerase chain reaction (PCR) to identify the putative pathogen. PCR is an economical, easy, fast, and reliable method for the detection of viruses and other plant pathogens. With the development of novel technologies such as Next-generation Sequencing (NGS) applied to metagenomics–metaviromics, the discovery of new viruses has been accelerated in recent years [1, 2].

Most known plant viruses have positive-sense single-stranded RNA genomes and are associated with diseases or defined symptoms [1, 3]. Single-stranded positive-sense RNA viruses ((+) RNA viruses) produce double-stranded RNA (dsRNA) during replication. dsRNA can be purified from infected plants and sequenced, allowing the identification of viruses [2, 3]. After sequence assembly and comparison with known viral sequences at databases—i.e., the GenBank, novel viruses are discovered [4]. Next, its etiology is studied by infectivity assays and confirmation of the virus presence by reverse transcription PCR (RT-PCR).

Primers' characteristics and the expected amplicon size depend on the method used for detection. This chapter focuses on end-point PCR because it allows faster optimization, it is cheap, and it is sensitive enough for most plant viruses. However, to detect low levels of viral RNA—e.g., for insect transmission studies—quantitative PCR (qPCR), also called real-time PCR, should be used. The precision and effectivity of virus detection by PCR relies on its inter- and intragenic variability [4]. Because primer binding is key to this method's success, diagnosis of virus variants might be affected if genetic variation occurs at the target site. For virus identification purposes, it is desirable to have a battery of primers to: (1) identify viruses at the species levels, (2) discriminate between virus variants or strains, and, whenever possible, (3) detect viruses from a higher hierarchical level—e.g., genus or family level. Therefore, sequence alignments and phylogenetical analysis must identify

conserved and unique regions from the target sequences for primer design [4].

Most plant viruses encode capsid proteins to enclose and protect nucleic acids; because of their structural function, their sequence remains highly conserved. Proteins involved in viral replication are also among the most conserved [1, 3]. Therefore, for capsid-less (+) RNA viruses, such as papaya umbra-like viruses, the RNA-dependent RNA Polymerase (RdRP) is an ideal target for diagnosis.

Papaya (*Carica papaya* L.) is a tropical crop whose fruit is marketed worldwide, and it is consumed for its taste and nutritional quality. The most important and widespread viral papaya disease is caused by Papaya ringspot virus type P (PRSV-P), a (+) RNA virus [5]. Another viral disease that has recently [6–8] gained worldwide relevance is 'Papaya Sticky Disease' (PSD) or 'papaya meleira'—in Portuguese. It was first reported in Brazil in the 1980s, and it is characterized by the spontaneous exudation of latex in fruits, which turns black upon oxidation. A dsRNA virus called Papaya meleira virus (PMeV) was first proposed as the causal agent for PSD [9]; its 8.7 kilobases (kb) genome is related to Totiviruses [10]. In 2008, PSD symptoms were found in Mexico [11]; however, primers against the 629 nucleotides sequence of PMeV [12] failed to detect the disease in Mexico [13]. In 2015, a partial viral sequence—1154 nucleotides—encoding a protein with 42% identity to the RdRP of umbraviruses was identified from papaya plants in Mexico showing PSD symptoms [13]. Two set of specific primers targeting this sequence also allowed detecting a virus in a plant from Brazil with meleira disease. Since the amplicons' sequences—173 and 491 bp—were highly similar in both countries, the virus found in Mexico was named Papaya meleira virus-Mexican variant (PMeV-Mx) [13]. On 2015, a 4285 nucleotides partial sequence of PMeV-Mx was released at the GenBank—accession number KF214786.1 (our unpublished results). Also in 2015, an umbra-like virus was reported in papaya plants showing variations of symptoms produced by PRSV-P in Ecuador; the new virus was named Papaya virus Q (PpVQ), and it was found associated to PRSV-P [14]. In 2016, an umbra-like virus related to PpVQ and PMeV-Mx was also reported in Brazil in papaya plants displaying PSD symptoms in synergistic association with PMeV. The new umbra-like virus was named Papaya meleira virus 2 (PMeV-2), while PMeV was renamed as PMeV-1 [15]. PMeV-2 RNA was found inside purified PMeV-1 particles, suggesting that trans-encapsidation of the umbra-like virus genome takes place in nature for transmission by insect vectors [15]. In contrast, PMeV-Mx is insect-transmitted and produces PSD symptoms in papaya plants in the absence of a PMeV-1 related virus [6]; PMeV-Mx is also seeds transmitted [16]. In 2018, three partial sequences with similarity to PMeV-2 from papaya plants associated to PSD symptoms in Colombia were

uploaded at the GenBank—accession numbers MG570380.1 MG570381.1 and MG570382.1 (unpublished). In 2019, a new umbra-like virus was found in Australia in papaya plants displaying PSD symptoms, using NGS [8]; although its sequence has not yet been published, the report says the virus is also seed transmitted. In summary, several papaya umbra-like viruses have been identified and associated to PSD in four countries: Brazil, Mexico, Colombia, and Australia; two of them show seed transmission in Australia and Mexico [8, 13], while the virus found in Ecuador (PpVQ) does not associate with PSD. More studies are required to understand how papaya umbra-like viruses such as PMeV-2 and PMeV-Mx produce PSD and why others like PpVQ do not. The development of universal primers for these capsid-less viruses will contribute to study virus genetic variation as well as to perform fast diagnosis in places were PSD has not yet been reported.

The strategy described here aims to design universal primers able to detect all known papaya umbra-like viruses, as well clade-specific primers targeting conserved and variable RdRP gene regions. The method outline is as follows:

1. Sequence download.
2. FASTA file editing.
3. Multiple sequence alignment.
4. Identification of conserved/variable regions.
5. Phylogenetic inference for grouping.
6. Primers design and analysis.
7. Checking for specificity and cross-reactivity of primers: in silico PCR.

2 Materials

2.1 Software Tools and Web-Based Applications

Computer equipment: All bioinformatics procedures described for this methodology can be carried out on most modern 64-bits desktop/laptop computers. The most CPU/RAM intensive parts of the methodology are executed at online open servers offloading the computational burden from the user's equipment. Online servers: National center for Biotechnology Information (NCBI) GenBank (https://www.ncbi.nlm.nih.gov/nucleotide/) [17]; NCBI BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) [18] MAFFT online version 7 (https://mafft.cbrc.jp/alignment/server/) [19].

Oligo Analyzer Tool is simple and easy-to-use tool that allows determining important primer properties such as melting temperature (T_m), also commonly known as annealing temperature, GC content, primer dimers, primer–primer compatibility, and primer loops. It also allows checking primers for multiplexing, etc. It can

be freely downloaded from the Web page https://oligo-analyzer.software.informer.com/download/

For the reverse primer, use the 'Reverse complement tool' at https://www.bioinformatics.org/sms/rev_comp.html.

Primer-BLAST application at https://www.ncbi.nlm.nih.gov/tools/primer-blast/ [20] is a toolbox to design primers which includes a tool for 'in silico PCR' ligated to a nonspecificity test in which our universal primers can be aligned against the plant host genome (*Carica papaya*) to look for nonspecific amplifications.

3 Methods

PCR primers designed from a single DNA sequence will yield an amplification fragment when that DNA sequence is used as a template, but their specificity—or lack of it—when used with other templates can only be predicted by determining the position of the primers relative to the conserved and variable regions of the template sequence. This can be achieved by means of aligning it to DNA belonging to closely—or not—related strains. On this regard, the first step to designing efficient PCR primers is to obtain sequences related to our DNA of interest and design PCR primers once the conserved/variable regions are pinpointed.

3.1 Sequence Download

For the purposes of our working example, as has been explained above, we will search for nucleotide sequences of the RdRP gene of the papaya meleira umbra-like viruses.

There are several online databases and repositories for genomic data (i.e., DNA, RNA, proteins, and genomes). Three of the most used are GenBank at the NCBI (https://www.ncbi.nlm.nih.gov/nucleotide/), EMBL-EBI at the European Molecular Biology Laboratory (https://www.ebi.ac.uk/), and the DNA Data Bank of Japan DDBJ (https://www.ddbj.nig.ac.jp/index-e.html), and share their data through the International Nucleotide Sequence Database Collaboration (http://www.insdc.org) [21] meaning that a search at any of these will retrieve data deposited on any of them. For sake of simplicity, these methodologies will center on retrieving data from GenBank at NCBI.

There are three main ways of finding sequences at GenBank: by accession numbers, keywords, or homology. Any of the methods, or combinations of these, can be used to gather the relevant sequences.

It should be noted that sequence records at these online databases can be very heterogeneous with regard to their lengths, coverage of the gene of interest, and annotation of other genes, ORFs, or hypothetical proteins present. On our working example, we will be using sequences deposited on GenBank for papaya umbra-like viruses isolated from Mexico [5, 13], Ecuador [14],

Brazil [15], and Colombia (three accessions; unpublished). The completeness of the 5′ and 3′ ends for each record varies as well as the number of genes annotated. Records from Mexico and Brazil include the annotation for a hypothetical protein called ORF1 upstream of RdRP. Records from Colombia are truncated at the 3′ end of the RdRP (missing about 364 nucleotides). The seven records include noncoding regions in addition to the annotated genes. All these features of the records should be taken in consideration when downloading and processing the sequences for the subsequent analysis.

3.1.1 Searching and Downloading by Accession Numbers

The most straightforward way of finding genomic data is through the use of one or several accessions numbers. These are unique identifiers for each DNA, RNA, or protein sequence available in the database. Accession numbers can be gathered from publications presenting/using sequence data.

The GenBank accession numbers for the seven isolates of Papaya umbra-like viruses for our practical case are as follows: Mexico (KF214786, MN203218), Ecuador (KP165407), Brazil (KT921785), and Colombia (MG570380, MG570381, MG570382). To retrieve just the coding sequences (CDS) for the RdRP genes, these steps are followed.

1. Access GenBank database at https://www.ncbi.nlm.nih.gov/nucleotide/.
2. On the search box, type or copy the following list of accession numbers without quotes: 'KF214786, MN203218, KP165407, KT921785, MG570380, MG570381, MG570382' (*see* **Note 1**). A list of the seven sequence records should be displayed. For each record, the following information is shown: title; molecule properties, length (in base pairs), topology (circular/linear), type (DNA/RNA); Accession and Gene Identifier (GI) numbers; links to associated coded proteins and taxonomy data. Below each record are links to display the data in three different formats: GenBank, FASTA (*see* **Note 2**), or as a graphical representation. Note the different length sizes for all the records.
3. There are several ways to download the sequences linked to each record. The easiest and most reliable way (compatible with downward analysis) (*see* **Note 3**) is to use the 'Send to:' menu at the top right. Activate its drop-down menu clicking the arrow (downward-pointing triangle). Select 'Coding sequences' (*see* **Note 4**). Select 'FASTA Nucleotide' as the Format required. Click the 'Create File' button to download the seven records sequences (*see* **Note 5**).

3.1.2 Searching and Downloading by Keywords

Without accession numbers, relevant sequence records can be retrieved using keywords on the search box. Keywords can be any word contained on any of the multiple fields that conform the record. Usually, the name of the virus should be enough to retrieve significant results, but sometimes special combinations are better suited. For example, the keyword combination 'Papaya Meleira virus' will retrieve around 20 records but will not include the Ecuador record (KP165407). This is due to the fact that the title of the record does not contain the word 'Meleira' instead being labeled as 'umbra virus.' The keyword combination 'Papaya Meleira Umbra virus' will recover zero records because not a single record contains the four words (the search expects to match all words). To overcome this, a keywords combination using 'Boolean operators' (AND, OR, NOT) can be used (*see* **Note 6**).

1. Access GenBank database at https://www.ncbi.nlm.nih.gov/nucleotide/.
2. Use the following keywords combination (without quotes but with the parenthesis): '(papaya AND virus) AND (meleira OR umbra).' A list of records should be displayed.
3. As mentioned previously, many records are only small fragments. To order the records from longest to shortest, activate the 'Sort by' drop-down menu at the top on the middle and select 'Sequence Length.'
4. For consistency with the rest of the example, we will select the same seven records as from Subheading 3.1.1: KF214786.1, MN203218.1, KP165407.1, KT921785.1, MG570380.1, MG570381.1, MG570382.1.

   Check the tick boxes for the corresponding records.
5. As above, to download the sequences, use the 'Send to:' menu at the top right. Activate its drop-down menu. Select 'Coding sequences.' Select 'FASTA Nucleotide' as the Format required. Click 'Create File' button to download the selected sequences.

3.1.3 Searching and Downloading by Homology

If a viral sequence of interest is already known, it can be used as a reference—query—to identify closely related sequences by a homology search of the database using the Basic Local Alignment Search Tool (BLAST) at NCBI. For the purpose of this example, we will use the KF214786.1 record as our query.

1. Access the BLAST tool at https://blast.ncbi.nlm.nih.gov/Blast.cgi.
2. Select the 'Nucleotide BLAST' version (*see* **Note 7**).
3. On the first section (Enter Query Sequence), copy/paste or type the accession number KF214786.1 at the box (*see* **Note 8**).

4. On the 'Query subrange,' use '1231' in 'From' and '2436' in 'To' (both numbers without quotes) (*see* **Note 9**).
5. On the second section (Choose Search Set), to speed the process, we can limit the search to be done only against records from viruses. To this effect, type 'viruses' (without quotes) on the box 'Organism' (*see* **Note 10**).
6. On the third section (Program Selection), at the 'Optimize for' list of options, select 'Somewhat similar sequences (blastn)' (*see* **Note 11**).
7. For the present case, we will use the rest of the default parameters (*see* **Note 12**).
8. Click the blue BLAST button.
9. The results are presented in four tabs: A table of 'Descriptions,' a 'Graphic Summary,' the 'Alignments,' and a 'Taxonomy' report (*see* **Note 13**).
10. At the table of the 'Descriptions' tab, deselect all results and then select those corresponding to the example list of accessions provided: KF214786.1, MN203218.1, KP165407.1, KT921785.1, MG570380.1, MG570381.1, and MG570382.1.
11. Retrieve the selected sequences using the drop-down menu 'Download' located at the top of the table. Select the option 'FASTA (aligned sequences)' (*see* **Note 14**).

3.2 FASTA File Editing

As explained before, for the design of the diagnostic primers, only the coding sequence (CDS) of the RNA-dependent RNA polymerase (RdRP) will be used as the target of PCR. The FASTA formatted files obtained through Subheadings 3.1.1 and 3.1.2 contain, for some strains, the ORF1 CDS in addition to the RdRP CDS. Before generating the multiple sequence alignment, the FASTA file must be manually edited to remove the ORF1 CDS and retain only RdRP CDS.

1. Open the downloaded file from Subheading 3.1.1 or Subheading 3.1.2 using a suitable text editor (*see* **Note 15**), e.g., notepad (MS Windows), TextEdit (Mac OS), or Text Editor (Linux).
2. Identify header lines that contain the words 'RNA-dependent RNA polymerase' and/or 'RdRP.' There should be seven such lines. These are the sequences that we want to retain (*see* **Note 2** as a guide).
3. Identify headers that do not match the above description. Delete these headers and the associated sequence lines.
4. Save the file and exit.

3.3 Multiple Sequence Alignment

Both the multiple sequence alignment construction and phylogeny inference are computationally taxing steps, with the phylogeny being the most demanding one. For multiple sequence alignment, there are many algorithms available implemented on different software packages or accessible through online servers applications. Clustal [22] and Muscle [23] algorithms are very popular options due to both being fast alignment algorithms that use low CPU and RAM resources, but both perform poorly when dealing with highly variable regions and/or long stretches of insertion–deletions (indels), which is often the case with viral sequences. On the other hand, algorithms such as MAFFT (E-INS-i option) [19], Probalign [24], and PRANK [25] are better suited for these kinds of datasets but become computationally expensive very quickly when adding extra sequences. In our experience, MAFFT is a good compromise between alignment accuracy of highly variable regions and resources required. To minimize the computational burdens, and as an operating system independent solution, the following steps are carried out at a publicly available online server running the MAFFT algorithm.

1. Access the MAFFT alignment server at htpts://mafft.cbrc.jp/alignment/server/.
2. As input, load the FASTA file edited on the previous section (Subheading 3.2): On the 'Input' section, use the 'Choose File' button to select and upload the corresponding file.
3. On the next section, defaults can be used with the following considerations.

   *'Direction of nucleotide sequences':

   Option 1: 'Same as input' (default).

   Use this option if using a FASTA file created as detailed in Subheading 3.1.1 or Subheading 3.1.2 (with the 'Coding sequences' option selected at download). All sequences should be on the correct direction of transcription (*see* **Note 4**).

   Option 2: 'Adjust direction according to the first sequence (accurate enough for most cases).'

   Use this option if the FASTA file was created from instructions in Subheading 3.1.3 (NCBI BLAST results), since some of the sequences could be reversed (*see* **Note 14**). This option should also be used if the FASTA file was created by using the 'Complete Record' option when downloading the sequences from NCBI nucleotide database.

   'Job name': Although optional, it is a good practice to fill in to keep track of multiple experiments.

   'Notify when finished': For small sets of sequences, results will be quickly displayed, but it still is a good practice to add an email to get the results link in case the browser quits.

4. For typical cases, most of the defaults for the 'Advanced settings' can be used with the following exception:
   'Strategy': MAFFT has several methods for alignment strategies. For the case of highly variable viral sequences, the 'E-INS-i' method is the most appropriate.
5. Click on the 'Submit' button. Wait time will depend on the number/size of sequences and load of the server.
6. Results are displayed in the CLUSTAL alignment format.
7. 'Clustal format' and 'FASTA format' links on top can be used to download the alignment on the indicated formats. The 'FASTA' format is widely accepted by other bioinformatics applications. The 'Clustal' format is a more human-readable format.

3.4 Identification of Conserved/Variable Regions

Alignments on the Clustal format are useful for visually inspecting and identifying conserved/variable regions. Sequences are divided (usually) on 60 nucleotides long lines. A bottom line is added indicating the level of conservation: Asterisks (*) below columns indicate 100% identity. Periods (.) indicate a conserved change (purine to purine or pyrimidine to pyrimidine substitutions).

1. Open on a suitable text editor application the multiple sequence alignment on Clustal format (*see* **Notes 3**, **15**, and **16**).
2. To find appropriate target regions for the design of diagnostic primers, we will need to find long enough regions (at least 18 nucleotides) of 100% (or almost) identical positions. These can be identified by long stretches of uninterrupted asterisks (*). Keep in mind that some identical regions could be split across two lines (end of one and start of the following).
3. If long enough 100% identical regions are not present, stretches of identical nucleotides interrupted by a very small number of either conserved substitutions (best) or mismatches can be used as explained before: one or two conserved substitutions (.) near the 5′ end or middle of the primer may be acceptable, or even useful to design more general degenerate primers. At this point, three possible scenarios exist:
   A. Long enough identical/conserved target regions are identified for at least a forward and a reverse primer (more than one target region for each one is ideal).
   B. Only a long enough identical/conserved target region is identified for either a forward or a reverse primer.
   C. There are not long enough identical/conserved target regions for neither of the PCR diagnostic primers.

4. For scenarios A and B, copy the identified candidate regions to a text file, noting their positions on the sequence. For scenarios B and C, to be able to find appropriate regions, we will need to divide the sequences into less variable (more closely related) groups.

3.5 Phylogenetic Inference for Grouping

As stated above, the phylogenetic inference is a very computationally, intensive bioinformatic step. For taxonomy analysis, lineage divergence, and/or very detailed evolutionary histories determinations, a precise phylogenetic inference should be carried out. Similar to the multiple sequence alignment situation, phylogenetic distance methods like Neighbor-Joining (NJ) [26] or UPGMA [27] are quite fast but again are not optimal for alignments containing highly variable regions and/or multiple large indels. Maximum Likelihood (ML) methods are more appropriate for these cases, and tools like PAML [28] or PhyML [29] are good implementations. Previously to the phylogeny construction, the best model and parameters for the inference must be tested. This step can take as much (or longer) than the actual phylogeny reconstruction. Bootstrapping (calculation of confidence levels for internal nodes) is even more computationally demanding.

However, if our only aim is to divide the divergent sequences into groups of more closely related sequences to aid on the pinpointing of conserved target regions, such a fine analysis is not required, and a distance method such as NJ is suitable. For this, we will use the phylogenetic tools associated with the MAFFT server fused before.

1. Go to the MAFFT alignment server at https://mafft.cbrc.jp/alignment/server/ and repeat **steps 1–5** of Subheading 3.3 (multiple sequence alignment).
2. At the results page, click the 'Phylogenetic tree' button.
3. For the 'Settings,' select 'NJ Conserved sites' (*see* **Note 17**). Method and the 'Jukes-Cantor' [30] 'Substitution model' (for our purposes, there is no need for 'Bootstrap' calculation).
4. Click 'Go!' button.
5. On the results page, there are many options to display the tree. Use the 'View tree on Phylo.io' button. Tree should be presented on a new window.
6. Visually inspect the distribution of sequences in the different clades and determine accordingly the most likely groupings.
7. Create as many copies of the FASTA file edited on Subheading 3.2 as the number of groupings determined on the previous step. Rename the file copies accordingly.

8. Manually edit each FASTA file deleting the appropriate sequences to keep only the relevant sequences for the intended group.
9. Using these new FASTA files, repeat for each one the steps of Subheadings 3.3 and 3.4 to find suitable target regions for the design of primers.

3.6 Primer Design and Analysis

If several conserved/unique regions are found after multiple sequence alignment, for end-point PCR, select amplicon length from 200 bp to 1 kb, but ideally close to 500 bp (*see* **Note 18**). Apply regular rules for primer parameters (*see* **Note 19**). For this example, to design the Forward universal primer, a 23-nucleotide stretch of 100% identity among all sequences located at position 360 from the start codon (ATG) was selected (Fig. 1, left). A second conserved region 300 nucleotides downstream from the first and spanning 20 nucleotides was selected for reverse primer (Fig. 1, right). The first segment exhibits 100% identity among all accessions, while the second has seven of the eight accessions identical, with exception of PMeV-2 from Brazil (accession KT921785) which only has two mismatches (*see* **Note 20**). Thus, for the universal forward primer, we can use the sequence as it is; while for the universal reverse primer, it is necessary to include two degenerations, as shown in Fig. 1 (*see* **Notes 20** and **21**). Because apart from these two conserved regions, the rest of the alignment revealed heterogeneity (not shown), the Phylogeny of the RdRP gene was inferred (Fig. 2) in order to select sequences for 'group-specific' primers.

Perform primer analysis as follows: Install 'Oligo Analyzer' in your computer and open it, select the primer tab (Fig. 3). To design the universal Forward primer, copy the 23 nucleotides sequence (5′ GAGAAACCTGTCTACTCTTGTTT 3′) of the first conserved region in the alignment (in Fig. 1, right) and paste it on the first Primer box. Name the primer (optional) in the primer name box. We recommend replacing the 'Default parameters' in 'Oligo

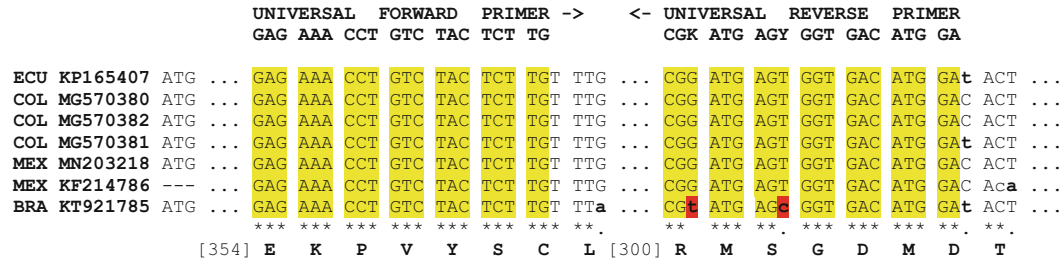

Fig. 1 Multiple sequence alignment of the Umbra-like virus RdRP gene. On bottom, identical positions are represented by asterisks (*) and conserved substitutions by dots (.). On top, sequence of universal primers, where K = G + T and Y = T + C

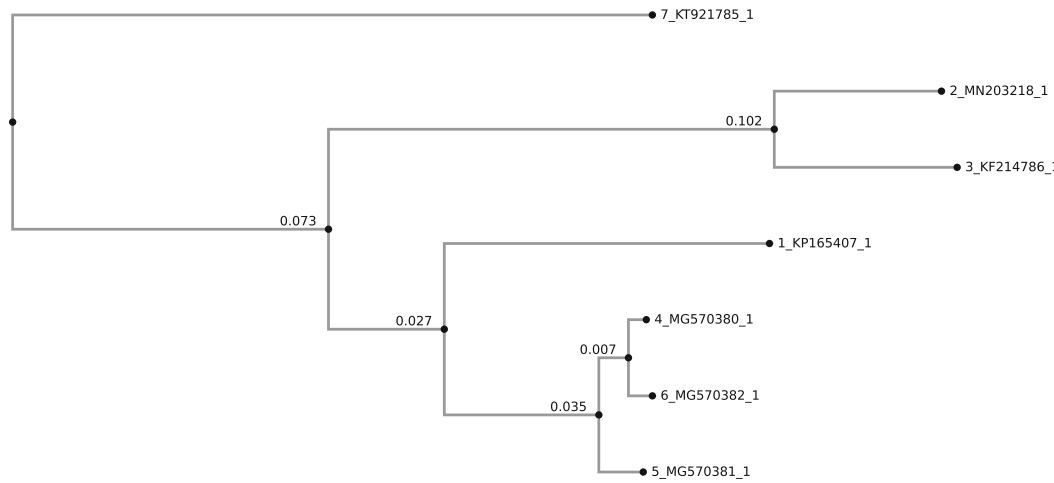

Fig. 2 Phylogeny for the RdRP gene from Papaya umbra-like viruses. Two clades are formed: The first includes only PMeV-2 from Brazil (KT921785); the second clade with two subclades includes in one branch PMeV-Mx (KF214786) and PMeV-2 (MN203218) from Mexico and in the second branch PpVQ from Ecuador (KP165407) along the Colombian accessions (MG570380, MG570381, MG570382)

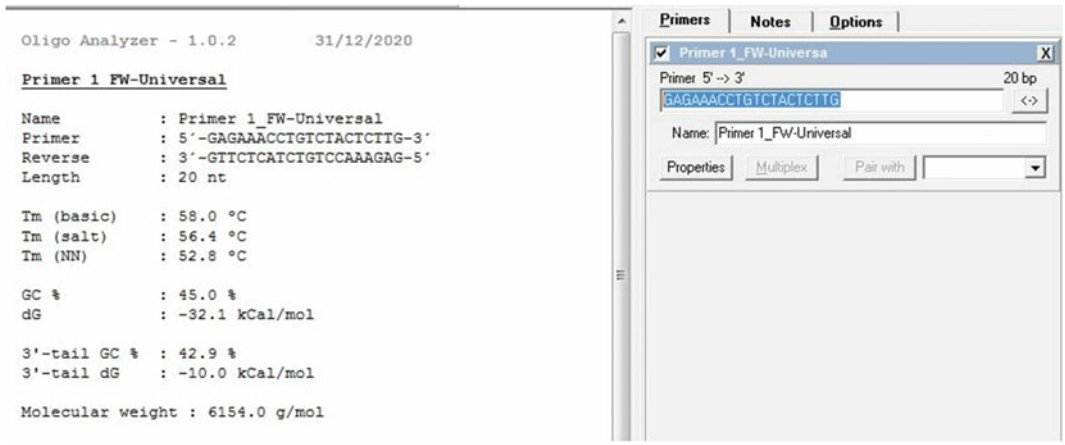

Fig. 3 Oligo Analyzer screenshot. Primers, Notes, and Options Tab are on the top right side. The Primer tab is open and shows the Forward universal primer sequence, and nt number (20 bp) is indicated. The 'reverse primer' button, the primer name box, the primer pair box with a drop-down menu. The 'properties,' 'multiplex,' and 'pair with' are also on the top section. At the bottom right side are the 'multiplex all' and 'Add primer' buttons

Analyzer' for the ones suggested here (*see* **Note 22**). In our experience, primers perform well in the laboratory. Click on the 'properties' button; the result will be displayed at the left. T_m at 'Salt concentration' is the most important, but it depends on your working conditions. Look for secondary structures at the bottom of the left window, if any, check the *d*G value. Values of −9 kcal/mol or more negative could affect PCR. In this example, when the

23 nucleotides sequence stretch is used, four structures appear below the 'primer self-annealing' legend and four under 'primers loops' (not shown) although the G values are acceptable; the T_m is 59.3 and the GC content is 39.1%. In order to demonstrate the effect of changing the number of nucleotides on the primer properties, remove the three 'T's at the primer 3′ end in 'oligo Analyzer' (5′ GAGAAACCTGTCTACTCTTG 3′) and press the 'properties button.' Now the secondary structures' G values increase (are less negative), the GC content raises from 39% to 40.9% and the primer ends in a 'G,' all these are improvements, although the T_m is reduced from 59 °C to 56.4 °C (Fig. 3). Select the final Universal Forward Primer sequence based on the T_m of the Reverse primer, adjust the primer length as needed (*see* **Note 21**).

For the 'universal degenerate Reverse primer,' a mismatch (G/T) and a conserved substitution (T/C) will be included in the second consensus sequence (Fig. 1, right), this primer will detect all known papaya umbra-like viruses (*see* **Note 21**). Copy the consensus sequence (CGGATGAGT GGTGACATGGA) and make the two substitutions (CGTATG AGC GGTGACATGGA). Obtain the reverse-complement in the 'Reverse complement tool' at https://www.bioinformatics.org/sms/rev_comp.html (*see* **Note 20**). Use oligo Analyzer to calculate primer properties. The universal reverse primer sequence should be a mix of 5′ TCCATGTCACC ACTCATCCG 3′ and TCCATGTCACC GCTCATACG, with 20 nucleotides and a T_m near 60.5 °C (*see* **Note 20**). To order the final Universal Degenerated Reverse primer, use this format 5′ TCCATGTCACC(G/A)CTCAT(A/C)CG 3′. Select the final Universal Forward primer sequence with T_m = 59.3 °C for primer synthesis (5′ GAGAAACCTGTCTACTCTTGTTT 3′) (*see* **Note 21**). Keep in mind that a degenerated primer is a mix of two or more sequences and that increasing degeneracy will allow identifying more potentially unknown virus variants; however, the annealing will be less specific, thus also increasing the possibility of false positives (*see* **Note 21**). If you want to detect only the known viruses, synthesize the two clade-specific reverse primers: (1) specific for PMeV-2 from Brazil and (2) the reverse primer for the group of viruses at clade 2 (from Mexico, Colombia and Ecuador) (5′ TCC ATGTCACCACTCATCCG 3′), with the universal forward primer.

3.7 Checking for Specificity and Cross-Reactivity of Primers: In Silico PCR

Since primers were designed on the most conserved regions of the RdRP, there is a possibility that their target regions (or a fraction of it) could be shared with hosts' proteins with polymerase activity or RNA binding domains producing false positives.

A way to check for this is to carry out an in silico PCR using as templates the genomes of the target virus and its host (in this case, *Carica papaya*). This in silico PCR methodology is implemented at NCBI Primer-BLAST online application. Its interface is very similar to the already described NCBI BLAST.

1. Access Primer-BLAST application at https://www.ncbi.nlm.nih.gov/tools/primer-blast/.
2. On the first section (PCR Template), copy/paste or type the accession number KF214786.1 at the box (*see* **Note 8**).
3. On the second section (Primer Parameters), copy/paste or type the sequences for the Forward and Reverse primers (reverse primer must be the reverse-complement). Leave the rest of the default values (*see* **Note 23**).
4. Leave the default values for the third section (Exon/intron selection).
5. In the fourth section (Primer Pair Specificity Checking Parameters), change 'Organism' to 'Carica papaya (taxid:3649)' (without quotes) (*see* **Note 10**). This is the host genome.
6. Leave the rest of the default values and click the blue 'Get Primers' button.
7. A primer specificity report, as shown in Fig. 4 is presented. The report is divided into three parts. The first section includes data of the target sequence with its name and size (Range). The line labeled 'Specificity of primers' reports if the tested pair of primers would be able (or not) to amplify a PCR product on the host's genome. The second part is a graphical representation that depicts the annotations (if any) on the target sequence including proteins' active sites and/binding sites. The primers and the expected PCR product are shown as blue arrows connected by a thin line. The third section contains data for each primer, including the strand to which they will anneal,

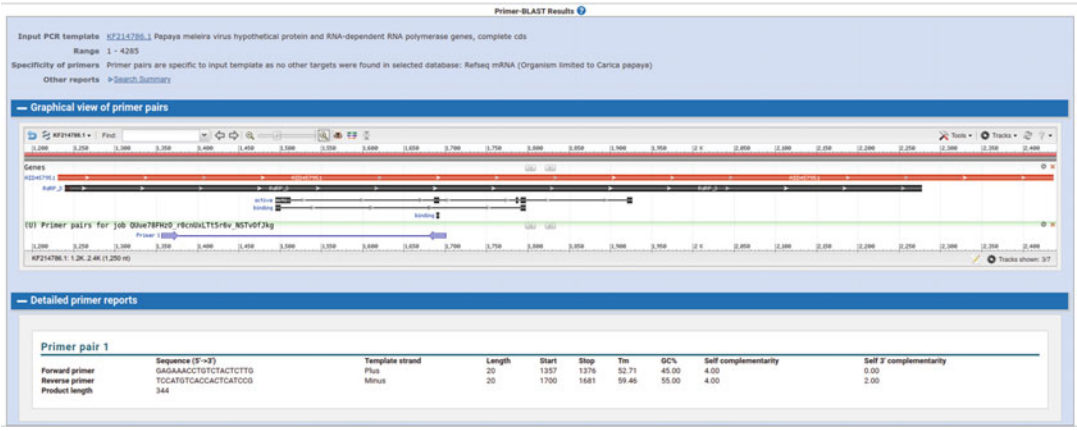

Fig. 4 Primer-BLAST specificity check. Example of report obtained for the specificity of umbra-like virus universal primer pair. At top: PCR template and primer specificity reports. In this example, universal primers are specific to the viral gene (RdRP) and do not target the host *Carica papaya*. At middle: graphical view of the *in silico* amplicon aligned to the target gene and protein (RdRP). At bottom: 'Detailed primer reports' including amplicon length and primer characteristics is shown

length, start, and end positions of annealing on target, T_m, GC content, as well as Self-complementarity and Self 3′ complementarity values. The last line reports the expected length of the PCR product.

4 Notes

1. List of accession numbers can be created by the using the following valid separators: ' ' (single blank space), ', '(comma followed by a single blank space), ',' (comma without a following blank space). Also, lists of accession numbers can be created by typing or pasting them (one by line) on text documents. Copying the resulting column of accession numbers and pasting them on the search box will produce the same result.
2. The FASTA format is a plain text format to store DNA/RNA/Protein sequences. Its simplest form is composed of two lines: a header line and a sequence line. The header line starts with the greater than sign (>) followed immediately by the sequence identifier. The sequence can be split into multiple lines. Here is a typical header from Genbank followed by the first three lines of its associated sequence:

```
>KF214786.1 Papaya meleira virus hypothetical protein and
RNA-dependent RNA polymerase genes, complete cds
ATGAACATTTCGAACATTCCCGTGGGACGTCTCACCTCCCATGCCGGTTTTAATTTGTT-
GAAGCTGGCAA
GCAAGCTTGGAAAGCAAAATCCTCCTTCCAAGTTTGCGGGTGGGCGAAGGC-
CAAATCCTGGTGAGTCTGG
CACCTCACACGGTGGGCCTTCTTCATCATCTCGACCTTCCCGAAGGCGAGG-
TAAGTTTGCCCTTAGGAGG
```

Although the header line, when wrapped, can extend more than a line, it should not be separated by a hard return. The actual sequence identifier ends at the first blank space (in this case KF214786.1 is the identifier). Anything after the first blank space but before a return is considered as human-readable metadata. If a file contains more than one sequence (a multisequence FASTA file), each sequence must have a unique identifier up to the first blank space.

Everything after the 'first return' and up to the 'next greater than' sign (or end of file) is considered the sequence. The sequence can be contained on a single line or divided on lines of arbitrary size.

3. The file automatically generated is in a simple text format. Copying and pasting a FASTA sequence (header and sequence lines) onto a MS Word document or similar word processing

software will produce unusable files. Although the format may look like a FASTA file, these applications include hidden code that makes them unsuitable for use with any bioinformatics application, unless an option to save as 'simple text' is available.

4. The 'Coding sequences' option will automatically select the fragments of sequence corresponding to the coding portion(s) of the annotated gene(s) present on the record. For double-stranded DNA sequences, if an annotated gene is transcribed on the complementary strand, this option will retrieve the appropriate reverse-complement sequence.
5. If all records are left unchecked, every result will be formatted and downloaded. This is especially useful when retrieving more than 20 sequences. By default, only the first 20 records are shown on the first page. Accession lists longer than this will be split into multiple pages. If all tick boxes are left unchecked, everything will be downloaded even if it is on a different page.
6. Boolean operators are the words (all in capitals) AND, OR, NOT. By default, the search algorithm implicitly includes the AND Boolean operator after each keyword. This is a strict operator that will only be true if all keywords are matched. The OR Boolean operator will be true if either the keywords before it or right after it are matched. Combinations can be thus created by pairing keywords with parenthesis and the necessary Boolean operators. For example, the records recovered by the keyword combination '(papaya AND virus) AND (meleira OR umbra)' must include three keywords: both the papaya and virus words plus another keyword that could be either meleira or umbra. Conversely, the NOT operator will discard any record that contains the immediately following keyword (or combination of keywords). More information can be found here https://www.nlm.nih.gov/pubs/techbull/ja97/ja97_pubmed.html.
7. When doing a homology search on a database through BLAST, there are four options of algorithms to use depending on the reference sequence used (query) and the database to be searched against. For the purpose of primer design, we need to recover nucleotide sequences so the nucleotide database should be searched against to. If our query is an mRNA or coding sequence, then nucleotide BLAST (blastn) should be used. Starting with a protein sequence, it is also possible to search the nucleotide database thorough means of the tblastn algorithm which will compare the amino acids of the query against a translated version of the nucleotide database.
8. The Query box can accept accession numbers, gene identifiers (GI), or copied and pasted FASTA sequences (even without a header line, although this is not advisable). For nucleotide

sequences, these may have numbers (that will be ignored) and the letter 'N' for ambiguities, but not any of the other ambiguities codes (e.g., R and Y).

9. The Query subrange can be used to limit the search to be carried out with only a determined portion of the query provided. For this case, the CDS for the RdRP gene is located on nucleotides 1231 to 2436 for this record (both inclusive).
10. While typing, suggestions will appear. Multiple names for the same organism or group, with small variations (different capitalization or pluralization, for example, 'viruses,' 'Viruses' or 'Vira'), may be suggested. As long as the accompanying taxid number is the same, there is no difference in which one selected (for viruses taxid:10239).
11. Three variations of the blastn algorithm are available: megablast, discontiguous megablast, and blastn. The first two options tend to be too strict and are useful when very little variation is expected among the sequences. 'Optimize for Somewhat similar sequences (blastn)' is more permissive of accumulated mutations, which is the case for viruses. A more detailed explanation can be found on this guide: https://www.ncbi.nlm.nih.gov/blast/BLAST_guide.pdf.
12. All BLAST algorithms have a number of parameters that can be fine-tuned depending on the particular characteristics of the query and/or search required. It is beyond the scope of this chapter, an explanation of all of the parameters but NCBI maintains many online resources that can be consulted to better understand the capabilities of these parameters. Some useful links are provided (this is not an exhaustive list): https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs. https://www.ncbi.nlm.nih.gov/books/NBK1734/.
13. It is also beyond the scope of this chapter an explanation of all the data displayed on the results, but the reader is directed to this handy introductory handout: https://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf.
14. In Subheadings 3.1.1 and 3.1.2 to recover only the coding sequences from the sequences, we selected that option from the 'Send to:' menu (*see* **Note 4**). The tool to download records from a BLAST results page does not have this option, but a similar function can be obtained with the 'FASTA (aligned sequences)' option. Since our query was limited to the subrange that encompasses the RdRP CDS, using the 'FASTA (aligned sequences)' option will download only sequences that matched that fragment of the original record. However, note that some results maybe shorter than expected

if there is too high variability on the ends. Also, while the 'Coding sequences' option of the 'Send to:' menu will obtain the CDS on the right translation direction (reverse-complementing the sequence if necessary), the 'FASTA (aligned sequences)' option will not (and should be taken in consideration when making multiple sequence alignments).

15. As mentioned on **Note 3**, FASTA files should be edited and manipulated using text editors that open and save files in simple text format. Some text editors also have the option to save in 'rich text format,' which also must be avoided.
16. The visual inspection of a multiple sequence alignment on Clustal format relies on the proper alignment of the characters displayed on the file. To accomplish this, a monospaced font (also called fixed-pitch, fixed-width, or nonproportional font) must be used. Some examples of monospaced fonts available on most systems are Courier, Courier New, Lucida Console, Monaco, or Consolas. Typical default document fonts like Arial, Times New Roman, Calibri, and Helvetica are not monospaced and will wrongly display the alignment file. If the alignment looks wrong, select all (Ctrl+A) and change the font accordingly.
17. Sequences from Colombia are truncated at the 3′ end and are missing about 364 nucleotides. The 5′ end also differs for some of the sequences aligned. These differences produce gaps on both ends of the alignment. If these gaps are used during the phylogenetic inference, the algorithm considers each one as an evolutionary event of insertion or deletion wrongly inflating the evolutionary distances between the sequences. Selecting the 'NJ Conserved sites' option will trim all the gaps from the alignment and only consider columns with nucleotides present in all of the records.
18. Visualization of amplicons smaller than 200 bp on agarose gels could be affected by an excess of primers in the reaction tube or when there is no amplification—e.g., in the negative controls—a spot of primers can be seen masking bands absence/presence. Use agarose concentration according to amplicon size. Larger than 1 kb amplicons will take more thermocycler time, it also slows the diagnosis process.
19. Primer length and sequence determine annealing temperature (T_m), which in turn determines PCR success. T_m should not differ more than 5 °C between Forward and Reverse primers, but a difference of 2 °C or less is ideal. A primer length of 18–20 nucleotides is recommended with a T_m from 54 °C to 60 °C, although T_ms closer to 60 °C are ideal to reduce self-annealing (loops/harping formation) and primer dimers (homo- and heterodimers) possibilities.

20. Please note that the sequence in Fig. 3 is just for the purpose of representation of the primer position at the conserved region, but the actual universal Reverse primer sequence has to be 'reverse-complemented' since it will be part of the complementary strand. To get the reverse-complement, copy the sequence and paste it in the box of the 'Reverse complement tool' at https://www.bioinformatics.org/sms/rev_comp.html, make sure than the 'reverse- complement Tab' is selected, press submit. Also, note that in virus diagnosis it is common to use the Reverse primer for first-strand cDNA synthesis; this is done to reduce background, however this can also generate false negatives (if the primer is not recognizing that particular strain of virus) or loss of sensitivity (when a degenerate primer is used). In our experience, diagnosis works very well using random hexamers for first-strand synthesis followed of PCR with specific primers, this cDNA can also be used in the PCR with host gene primers—e.g., the 'actin' gene in papaya—as positive control [13].
21. Keep in mind than combinations of the degenerated bases will be produced when the primer is synthesized, therefore variations on the calculated T_m should be expected. As stated before, having a universal primer would allow detecting virus variants that might have not yet being reported, however the diagnosis sensitivity could be reduced.
22. To adjust parameters in 'Oligo analyzer,' go to the 'Options' tab and fill in the boxes the following values: 3′-tail length '7,' Salt concentration '50' mM, DNA Conc. '500' pM, click the Save button.
23. By default, the PCR product is expected to have a length between 70 and 1000 bp. The optimal T_m for the primers is, by default, between 57 °C and 63 °C. If the primers designed in Subheading 3.7 are outside these default values of product length and/or T_m, the values should be changed accordingly.

Acknowledgments

This work was funded by the CONACYT research grant A1-S-19850 to L.L.O.

References

1. Dolja VV, Koonin EV (2018) Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. Virus Res 244:36–52. https://doi.org/10.1016/j.virusres.2017.10.020
2. Roossinck MJ, Martin DJ, Roumagnac P (2015) Plant virus metagenomics: advances in virus discovery. Phytopathology 105:6, 716–727
3. Dolja VV, Krupovic M, Koonin EV (2020) Annu Rev Phytopathol 58:23–53. https://

doi.org/10.1146/annurev-phyto-030320-041346

4. Rubio L, Galipienso L, Ferriol I (2020) Detection of plant viruses and disease management: relevance of genetic diversity and evolution. Front Plant Sci 11:1092. https://doi.org/10.3389/fpls.2020.01092
5. Alcala-Briseño RI, Casarrubias-Castillo K, López-Ley D et al (2020) Network analysis of the Papaya Orchard Virome from two agroecological regions of Chiapas, Mexico. mSystems 5:e00423-19. https://doi.org/10.1128/mSystems.00423-19
6. Garcia-Camara I, Tapia-Tussell R, Magaña-Alvarez A et al (2019) Empoasca papayae (Hemiptera: Cicadellidae)-mediated transmission of Papaya meleira virus-Mexican variant in Mexico. Plant Dis 103:2015–2023
7. Sá-Antunes TF, Maurastoni M, Madroñero LJ (2020) Battle of three: the curious case of papaya sticky disease. Plant Dis 104:2754–2763. https://doi.org/10.1094/PDIS-12-19-2622-FE
8. Campbell P (2018) New test to offer early detection of papaya sticky disease. Papaya Press. https://australianpapaya.com.au/website/wp-content/uploads/2018/05/PAPAYAPRESS-MAY.pdf
9. Maciel-Zambolim E, Kunieda-Alonso S, Matsuoka K, De Carvalho M, Zerbini F (2003) Purification and some properties of Papaya meleira virus, a novel virus infecting papayas in Brazil. Plant Pathol 52:389–394
10. Abreu EFM, Daltro CB, Nogueira EOPL et al (2015) Sequence and genome organization of papaya meleira virus infecting papaya in Brazil. Arch Virol 160:3143–3147
11. Perez-Brito D, Tapia-Tussell R, Cortes-Velazquez A et al (2012) First report of papaya meleira virus (PMeV) in Mexico. Afr J Biotechnol 11:13564–13570
12. Abreu PMV, Piccin JG, Rodrigues SP et al (2012) Molecular diagnosis of Papaya meleira virus (PMeV) from leaf samples of *Carica papaya* L. using conventional and real-time RTPCR. J Virol Methods 180:11–17
13. Zamudio-Moreno E, Ramirez-Prado J, Moreno-Valenzuela O et al (2015) Early diagnosis of a Mexican variant of Papaya meleira virus (PMeV-Mx) by RT-PCR. Genet Mol Res 14:1145–1154
14. Quito-Avila DF, Alvarez RA, Ibarra MA et al (2015) Detection and partial genome sequence of a new umbra-like virus of papaya discovered in Ecuador. Eur J Plant Pathol 143:199–204
15. Sa Antunes TFS, Amaral RJV, Ventura JA et al (2016) The dsRNA virus papaya meleira virus and an ssRNA virus are associated with papaya sticky disease. PLoS One 11:e01552
16. Tapia-Tussell R, Magaña-Alvarez A, Cortes-Velazquez A et al (2015) Seed transmission of Papaya meleira virus in papaya (*Carica papaya*) cv. Maradol. Plant Pathol 64:272–275
17. Resource Coordinators NCBI (2016) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 44:D7–D19. https://doi.org/10.1093/nar/gkv1290
18. Zhang Z, Schwartz S, Wagner L et al (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214. https://doi.org/10.1089/10665270050081478
19. Katoh K, Rozewicki J, Yamada KD (2019) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform 20:1160–1166. https://doi.org/10.1093/bib/bbx108
20. Ye J, Coulouris G, Zaretskaya I et al (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinform 13(1):134. https://doi.org/10.1186/1471-2105-13-134
21. Karsch-Mizrachi I, Takagi T, Cochrane G & International Nucleotide Sequence Database Collaboration (2018) The international nucleotide sequence database collaboration. Nucleic Acids Res 46:D48–D51. https://doi.org/10.1093/nar/gkx1097
22. Sievers F, Higgins DG (2021) The clustal omega multiple alignment package. In: Katoh K (ed) Multiple sequence alignment. Methods in molecular biology, vol 2231. Humana Press, New York, NY, pp 3–16. https://doi.org/10.1007/978-1-0716-1036-7_1
23. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340
24. Roshan U, Livesay DR (2006) Probalign: multiple sequence alignment using partition function posterior probabilities. Bioinformat 22:2715–2721. https://doi.org/10.1093/bioinformatics/btl472
25. Löytynoja A (2014) Phylogeny-aware alignment with PRANK. In: Russel D (ed) Multiple sequence alignment methods. Methods in molecular biology (Methods and protocols), vol 1079. Humana Press, Totowa, NJ, pp 155–170. https://doi.org/10.1007/978-1-62703-646-7_10
26. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425.

https://doi.org/10.1093/oxfordjournals.molbev.a040454

27. Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kans Sci Bull 38:1409–1438
28. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 8:1586–1591. https://doi.org/10.1093/molbev/msm088
29. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321. https://doi.org/10.1093/sysbio/syq010
30. Jukes TH, Cantor CR, Munro HN, Allison JB (1969) Evolution of protein molecules. In: Mammalian protein metabolism. Academic, New York