Kodoja: A workflow for virus detection in plants using *k-mer* analysis of RNA-sequencing data

Amanda Baizan-Edge[1], Peter Cock[2], Stuart MacFarlane[3], Wendy McGavin[3], Lesley Torrance[1,3] and Susan Jones*[2]

1. The School of Biology, University of St Andrews, Biomedical Sciences Research Complex, St Andrews, KY16 9ST, UK
2. Information and Computational Sciences Group, The James Hutton Institute, Dundee, DD2 5DA, UK
3. Cell and Molecular Sciences Group, The James Hutton Institute, Dundee, DD2 5DA, UK

* To whom correspondence should be addressed. Tel: +44(0)3449 285428 Email: Sue.Jones@hutton.ac.uk

Key Words: Plant virus diagnostics, RNA-sequencing, *k-mer* analysis, Raspberry yellow net virus, Beet ringspot virus, Bioinformatics

Repositories: The RNA sequences of 2 raspberry plants exhibiting virus-like symptoms have been deposited in the European Nucleotide Archive and assigned accessions ERR2784286 (D5) and ERR2784287(D6)

Abstract

Background: RNA-sequencing of plant material allows for hypothesis-free detection of multiple viruses simultaneously. This methodology relies on bioinformatics workflows for virus identification. Most workflows are designed for human clinical data, and few go beyond sequence mapping for virus identification.

Methods: We present a new workflow (Kodoja) for the detection of plant virus sequences in RNA-sequence data. Kodoja uses *k-mer* profiling at the nucleotide level and sequence mapping at the protein level by integrating two existing tools Kraken and Kaiju.

Results and Discussion: Kodoja was tested on 3 existing RNA-seq datasets from grapevine, and 2 new RNA-seq datasets from raspberry. For grapevine, Kodoja was shown to be more sensitive than a method based on contig building and Blast alignments (27 viruses detected compared to 19). The application of Kodoja to raspberry, showed that field-grown raspberries were infected by multiple viruses, and that RNA-seq can identify lower amounts of virus material than RT-PCR. This work enabled the design of new PCR-primers for detection of Raspberry yellow net virus and Beet ringspot virus. Kodoja is a sensitive method for plant virus discovery in field samples and enables the design of more accurate primers for detection. Kodoja is available to install through Bioconda and as a tool within Galaxy.

1.0 Introduction

Virus infection is of specific importance in crops cultivated for food and fuel. Viruses cause significant yield and quality losses, and consequently they have important negative economic impact (1). In the UK, *Potato virus Y* causes annual potato crop losses of £30-40 million (2), and in Asia viruses infecting rice (such as *Rice grassy stunt virus*) can cause annual crop losses of $120 million (3). These examples highlight the need for fast and accurate virus detection methods. Viral infection symptoms can include yellowing and stunting, but in many cases symptoms can be absent or masked by other factors. In some cases plant viruses interact synergistically, to cause new or more severe disease symptoms (4). One example is crumbly fruit complex disease of raspberry, which can be caused by the presence of two viruses; *Raspberry bushy dwarf virus* and *Raspberry latent virus* (5). As crops are cultivated in new geographical locations and agricultural practices are intensified, there is an increasing risk of new viruses becoming established, and existing ones widening their host range. Hence, plant virus diagnostics is a field of increasing significance in terms of future food security.

Standard molecular techniques for detection of viruses include methods based on reverse transcriptase polymerase chain reaction (RT-PCR). But such techniques only allow the detection of known viruses, i.e. each test is specific to one virus or a very small number of related viruses (6). Furthermore, viral genomes evolve which can make tests ineffective over time, making disease diagnosis slow and restrictive. Such limitations have recently been overcome through the use of next generation sequencing (NGS) methods for hypothesis-free simultaneous detection of multiple viruses (7). The majority of plant viruses have RNA as their genetic material and those that have DNA genomes produce RNA transcripts. In addition, eukaryote small interfering RNAs (siRNAs) direct antiviral immunity through RNA interference and during this process virus-derived siRNAs are enriched in the host (8). Hence, both RNA and small RNA (sRNA) sequencing are effective methods for virus detection in plants. However, this relies upon two important elements: (a) robust RNA extraction and enrichment protocols, and (b) fast and robust bioinformatics tools for virus identification.

A range of RNA-extraction and enrichment protocols, and bioinformatics workflows, has previously been developed for human clinical samples (for review see (9). Recently such work has resulted in a viral disease diagnosis and actionable clinical management within 48 hours (10). The workflow used in this clinical work comprised the two main elements required for a virus diagnostic tool: (a) identification and removal of host nucleotide sequences, and (b) identification of virus sequences. However, virus detection in clinical samples presents an easier problem than in plants, as the human genome is well annotated (allowing easy removal of host sequences), and human virus data are more prevalent in sequence databases (allowing for easy identification of the virus sequences that are present). In comparison, many crop plant

genomes are incomplete or poorly annotated, and plant virus sequences are under-represented in databases.

We recently reviewed the bioinformatics tools and workflows currently available for virus detection from NGS data (9). From this we concluded that the majority were optimised for human NGS data, few went beyond sequence identity for virus identification, and many required significant computational knowledge for installation and/or use. Two tools, Taxonmer (11) and VirusDetect (12) are available as web servers and provide the potential for the analysis of RNA-sequence data from plants (2). However, the review highlighted the fact that whilst three published tools had been tested on plant data, projects focused on detecting viruses in plants have not used them. Instead, projects have used standalone mapping and assembly algorithms outside of a workflow, as this approach has generally offered greater flexibility during the analysis.

Any virus identification workflow needs to be capable of: (a) conducting quality control measures on raw data files, including trimming of poor quality reads and adaptor sequences, (b) identifying host sequences and (c) identifying viral sequences. The identification of known viruses can be done by mapping to a database of existing virus sequences, but the identification of new strains or novel viruses requires expert knowledge and additional analyses beyond a workflow.

Many of the published virus detection workflows use contig assembly and mapping algorithms to identify viral sequences (9). But, both assembly and mapping can be very computationally intensive, meaning that workflows can have long run times for large datasets. Assembly and mapping methods also result in unassembled reads being left unidentified. One alternative way to identify virus reads in RNA-seq datasets is to use *k-mer* profiling, which has been successfully implemented in Taxonomer (11). RNA and DNA sequences can be treated as character strings and divided into multiple substrings of length *k*. In this way a sequence can be represented by *k-mer profiles*, and these profiles can be compared for taxonomic assignment. K-mer profiles have been used in a range of similarity searches in bioinformatics. In metagenomics it allows alignment-free similarity analyses between sequences (13), and in taxonomic profiling, binning methods use k-mer profiles to cluster sequences and allow draft genome recovery (14). Such methods have also successfully been applied to the identification of viral haplotypes within a population without using a reference genome (15).

The Kodoja workflow, presented here, combines a set of unique features that make it applicable to a wide range of researchers working with NGS datasets. Our aim was to develop a workflow that went beyond assembly and mapping methods, that was specifically optimised for plant datasets, and was accessible to the non-bioinformatician. Kodoja is a workflow that allows virus sequences to be identified from mixed (comprising both plant and potentially viral, bacterial and fungal nucleotides)

RNA-seq data. Kodoja is unique in that it is (a) specific for plant NGS data, (b) uses k-mer profiling at the nucleotide level and sequence alignment at the protein level for virus classification by integrating the existing tools Kraken (16) and Kaiju (17), (c) is available for local installation through Bioconda (18), and (d) is available as a tool within the Galaxy web-based analytical environment (19).

2.0 Methods

2.1. The Kodoja workflow

The Kodoja workflow combines two existing tools, Kraken (16) for taxonomic classification using k-mers at the nucleotide level and Kaiju (17) for sequence matching at the protein level. Kodoja has three main components, summarized in Figure 1: (a) kodoja_build for database generation for Kraken and Kaiju (b) kodoja_search for the taxonomic classification of RNA-seq reads, and (c) kodoja_retrieve for extraction of viral sequences by species for downstream analysis.

2.1.1. Kodoja_build: Database generation

For virus classification, the main Kodoja components (Kraken (16) and Kaiju (17)) each require a database generated from the genome or proteome of known plant viruses, and (if available) the genome or proteome of the plant host. Data download and database generation are achieved using the kodoja_build module. This module downloads genomes and protein sequence files from RefSeq (20), and then implements code from Kraken and Kaiju to generate tool-specific databases. The user can specify if all viruses or only plant viruses are included in the databases. If a host genome is available (either provided by the user or in RefSeq (20)), this can also be added to the database for host sequence classification.

To make Kodoja easy to use, ready-made plant-specific viral databases for Kraken and Kaiju are provided for download at https://doi.org/10.5281/zenodo.1406071. These were generated by downloading all complete virus and viroid genomes and protein sequence files in NCBI RefSeq (Release 89)(20) and selecting plant viruses using information from the Virus-Host DB (21). For Kraken, *k-mer* size is specified when building the database, and a *k-mer* size of 31 was used for the RNA-seq datasets.

2.1.2. Kodoja_search: Taxonomic classification of virus reads

Kodoja_search is the main Kodoja component. RNA-seq reads are first quality checked using Trimmomatic (22) which trims and discards low-quality reads. FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is used for summarizing the read quality after trimming, and the FASTQC report forms part of Kodoja's results. Kraken (16) is then used for the nucleotide-level classification. Kraken is a sequence classification algorithm for assigning taxonomic labels to short sequences (16). It does this through dividing each sequence into *k-mers* and querying each against a *k-mer*

database. *K-mers* which are shared between organisms are mapped to the lowest common ancestor, and this information is then used to build a subtree of the general taxonomy tree for the classification of the sequences. In the tree, each node has a weight equal to the number of *k-mers* in the sequence associated with the node's taxon. Each root-to-leaf path in the tree is scored by adding all the weights in the path. The leaf of the path with the largest score is the classification used for the sequence (16). The use of the *k-mer* database makes the classification algorithm very fast compared to alignment based methods (11).

In the next step full length sequence reads are translated and classified at the protein-level using Kaiju (17). Kaiju translates the sequences into six frames and splits the resulting translations into fragments using translation termination codons (UAG, UAA, UGA). Kaiju balances precision and sensitivity by using a minimum fragment length parameter. We used a minimum fragment length of 15 and the number of mismatches permitted was one. Fragments are queried against a protein database using a modified version of the backwards search algorithm in the Burrows–Wheeler transform (23). A key component of sequence classification for both Kraken and Kaiju is the tool-specific database. We have provided pre-computed plant virus databases that can be used directly with the Kodoja workflow, but custom databases can also be made using kodoja_build (see section 2.1.1).

Implementation of the kodoja_search module results in reads being assigned to taxonomic classes by both Kraken and Kaiju. Reads assigned to the same virus class by both tools (set intersection) are designated as stringent assignments; and reads assigned to a virus class by either Kraken or Kaiju (set union) are assigned as non-stringent assignments. The assignments are given in a results summary, which includes the reads counts for each type of assignment. Full results from Kraken and Kaiju are also provided so that users can analyse these data further, outside of the Kodoja workflow.

2.1.3 Kodoja_retrieve: Extraction of viral reads

This module can be used to extract species-specific sequences for downstream analysis outside of the Kodoja workflow. The user can specify retrieval of reads classified to a species, and/or genus, using either stringent or non-stringent assignments. The ability to retrieve and download all reads assigned to a specific virus gives the user the potential to assemble complete viral genomes for further analysis.

2.1.4 Kodoja workflow availability

Kodoja is available for direct installation and use at the command line in Linux through Bioconda (18) (https://anaconda.org/bioconda/kodoja). Alternatively, the code can be downloaded from github (https://github.com/abaizan/kodoja). Kodoja is also provided as a package in Galaxy, an open source web-based analytical environment for data analysis (19). This is available on GitHub (https://github.com/abaizan/kodoja_galaxy) and the Galaxy Tool Shed (https://toolshed.g2.bx.psu.edu/view/abaizan/kodoja).

Developing Kodoja as a package within Galaxy makes it available to researchers with a local installation of Galaxy, and allows analysis to be completed with no command line input. By using an open source workflow platform in this way, the tool can also potentially be used on a cloud-based Galaxy server.

2.2 Benchmarking Kodoja using existing datasets

Kodoja was tested on three publicly available RNA-seq grapevine datasets (24) analysed for the presence of viral sequences (25). In the original work sequencing data for 11 grapevine samples was obtained, including multiple samples from skin, grain, and seed (24). In the analysis work viral sequences were identified using contig building and subsequence Blast alignment of contigs to a reference viral database (25). For the Kodoja benchmarking, we selected one library from grain (G1R1) (Sequence Read Archive (SRA) identifier SRR866540), skin (S3R1) (SRA: SRR866571) and seed (S3R3) (SRA:SRR866576); representative of those datasets with the largest and most diverse viromes. These datasets are denoted GV1, GV2 and GV3 respectively in the current analysis.

2.2.1. Assembly and alignment for confirmation

To confirm the viruses predicted by Kodoja, kodoja_retrieve was used to extract reads assigned to each virus. Reads for each virus were then assembled using Trinity (26) with minimum contig length of 200 nucleotides. The longest contig for each virus was then aligned against the NCBI non-redundant nucleotide database using Blastn, and the match with lowest e-value selected for taxonomic comparison. Where too few reads were available for contig assembly, all reads assigned to a virus species by Kodoja were aligned.

2.3. Applying Kodoja to virus detection in Raspberry (*Rubus idaeus*)

Kodoja was then applied to RNA-seq libraries generated from two raspberry plants of variety Glen Dee (denoted D5 and D6) collected from a commercial raspberry plantation in Angus, Scotland, UK. Both plants showed viral infection symptoms: D5 showed vein yellowing and D6 showed leaf blade yellowing (Figure 2).

2.3.1 RNA-sequencing

Symptomatic leaves were collected from each plant (D5 and D6) and frozen at -80ºC for long-term storage (15 months). Two samples of leaf were placed in a clean, autoclaved 2 ml Eppendorf tube together with a sterile 3 mm glass bead, frozen with liquid nitrogen and then powdered using a bead beater (Qiagen TissueLyser). Then 100 mg of powdered leaf was resuspended in a mixture of 450 µl Qiagen RNeasy Plant Mini Kit buffer RLT, 45 µl Ambion Plant RNA isolation aid and 4.5 µl 2-mercaptoethanol. Thereafter the RNA extraction followed the manufacturer's instructions to the RNeasy kit, and the RNA was eluted in RNAse-free water. The RNA was supplied to the Glasgow Polyomics facility (UK) for quality control, ribosomal RNA depletion, library preparation (paired-end 200 bp) and high-throughput sequencing using an Illumina NextSeq instrument (RD_PE2x75_33M). The raw data files for

sample D5 and D6 comprised 64 M and 62 M reads respectively (available from the European Nucleotide Archive (27) under accessions ERR2784286 and ERR2784287 respectively).

2.3.2 Kodoja analysis of raspberry RNA-seq datasets

The Kodoja workflow was run on the two raspberry RNA-seq datasets, using the draft genome of black raspberry (*Rubus occidentalis*)(28) as the host in the Kraken database build.

2.3.3 Assembly and alignment for confirmation

To confirm the predicted viruses, kodoja_retrieve was used to extract reads assigned to each virus species; and contigs were assembled and aligned to a reference database as described in section 2.2.1.

2.3.4 PCR confirmation of virus sequences

To confirm that the viruses identified by Kodoja were present, new samples of total RNA were extracted from the frozen leaves of sample D5 and D6 using the Thompson buffer method as described previously (29) and eluted in RNAse-free water. For detection of *Raspberry leaf mottle virus* (RLMV) the plant RNA was converted to cDNA using SuperScript III (Invitrogen) reverse transcriptase and random hexamer primer following the manufacturer's instructions. For other RNA viruses (*Raspberry leaf blotch virus* (RLBV) and *Beet ringspot virus* (BRSV)) the extracted plant RNA was added directly to a 25 µl illustra Ready-to-Go RT-PCR bead (GE Healthcare) reaction together with virus-specific PCR primers (Table 1). To detect the DNA plant virus *Rubus yellow net virus* (RYNV), six 1 cm diameter frozen D5 and D6 leaf discs were extracted using the DNeasy Plant Mini Kit (Qiagen) according to the manufacturer's instructions. RYNV was detected in the eluted DNA by amplification in a 25 µl illustra Ready-to-Go PCR bead (GE Healthcare) reaction with virus-specific primers (Table 1). Positive controls for virus-detection were RNAs extracted from raspberry plants previously demonstrated to carry specific viruses.

3.0 Results

3.1 Benchmarking of Kodoja workflow on RNA-seq from grapevine

The Kodoja workflow was applied to three publicly available RNA-seq libraries generated from grapevine (24) and analysed for virus sequences (25). The viral sequences detected [with stringent level assignments for viruses and non-stringent for viroids (as viroids do not have protein assignments in RefSeq)] in each sample are summarized in Table 2. Kodoja identified 6, 12 and 9 virus sequences in samples GV1, GV2 and GV3 respectively. For each sample, Kodoja identified all the viral sequences reported in the previous study (25), and in addition, identified 8 viral sequences not reported in the previous study; *Grapevine leafroll-associated virus* 1 (GLRaV1), *Apple mosaic virus* (ApMV), *Grapevine yellow speckle viroid 2* (GYSVd2), *Grapevine rupetris*

vein feathering virus (GRVFV), *Parietaria mottle virus* (PMoV), *Grapevine asteroid mosaic-associated virus* (GAMaV) and *Grapevine rootstock stem lesion associated virus* (GRSLaV) (Table 2). One explanation for the identification of additional virus sequences, could be their submission to GenBank after the date of the previous study (2015). However, 6 of the additional sequences have GenBank submission dates prior to 2011 and only GRVFV and GAMaV have submission dates after 2014 (GAMaV: 2016 and GRVFV: 2017). GAMaV was identified by Kodoja in GV3, and GRVFV was identified in GV2 and GV3. Only two viruses reported in the previous study were not identified by Kodoja: *Grapevine Pinot Gris virus* (GPGV) in GV1, and the *Oat blue dwarf virus* (OBDV) in GV3.

Overall, 85.2% (23/27) of virus species identified by Kodoja were confirmed by the contig assembly and Blast alignment process (Table 2). This included viruses that had not been identified in the previous study (including GRVFV in GV2 and GV3, and GAMaV in GV3). Contig mapping to reference genomes for these two viruses, showed that multiple and extensive regions of the virus genomes were present in the dataset (Figure 3). However, two viruses identified by Kodoja were classified as different species by contig assembly and Blast alignment. GYSVd2 was a viroid identified in all three samples by Kodoja (Table 2), but was classified as *Grapevine yellow speckle viroid 1* (GYSVd1) by the confirmation process. GRSLaV sequences were identified in GV2 by both Kodoja and Jo et al., 2015 but the sequences were classified as *Grapevine leafroll-associated virus 2* (GLRaV2) by the confirmation process.

3.2 Application of Kodoja for the detection of viruses in Raspberry

Kodoja was then applied to the identification of virus sequences in two field-grown raspberry plants with virus-like symptoms (Figure 2). Classifying reads with stringent assignments only, Kodoja identified six viruses in D5 and five viruses in D6 (Table 3). This included *Raspberry leaf blotch virus* (RLBV), *Rubus yellow net virus* (RYNV) and *Cherry leaf roll virus* (CLRV) detected in both samples; and *Beet ringspot virus* (BRSV) detected in D5 only and *Raspberry leaf mottle virus* (RLMV) detected in D6 only. The contig assembly and Blast confirmation process showed that all the assembled contigs corresponded to the viruses identified by Kodoja (Table 3). Contig mapping to reference genomes for selected viruses, showed that multiple and extensive regions of RYNV, RLMV and RLBV genomes were present in the datasets, but only a very short region of the BRSV genome was detected (Figure 4).

In a further confirmation step, RT-PCR was done with a previously used virus-specific primer pair for each of RLMV, RLBV and RYNV (Table 1; primers designed using previously published sequences). These primers detected RLMV in D6 only as predicted by Kodoja (Table 3). However, these primers did not detect RLBV or RYNV in either D5 or D6 as predicted by Kodoja (Table 3). Hence, samples D5 and D6 were tested with three additional RLBV primer pairs [1491/1492, 1495/1496, 2113/2114 (Table 1)] that target three different RLBV RNAs (RLBV has eight viral RNAs in total) based on the sequences assembled from D6. A very faint amplification band was

obtained with primer pair 1491/1492, suggesting a low level of RLBV RNA was present in this sample. However, none of the other RLBV-specific primer pairs produced a positive result for RLBV in D6. None of the 4 RLBV primer pairs gave amplification bands for D5, despite Kodoja predicting the virus was present and despite these primers producing a strong amplification of RLBV from a positive control plant. It should be mentioned that contamination of material submitted for deep sequencing can occur, particularly when preparation work is done in laboratories lacking designated clean rooms. This could be an alternative explanation for the failure to confirm the presence of RLBV by RT-PCR from the D5 and D6 samples.

An additional primer pair was then designed for RYNV [3470/3471(Table 1)] based on the sequence assembled from D6 and was tested on samples D5 and D6. This RT-PCR gave an amplification band for both D5 and D6 but produced non-specific amplification with a RYNV positive control plant (Figure 5A). In an additional test, a new BRSV RNA2-specific primer pair [3472/3473 (Table 1)] was designed based on the sequences assembled from D5. This primer pair detected BRSV in both D5 and D6, even though Kodoja only predicted the presence of BRSV in D5 (Figure 5B).

Discussion

We have developed and applied a new computational workflow (Kodoja) for the identification of plant virus sequences in RNA-seq data. The testing of Kodoja on 3 existing RNA-seq datasets from grapevine showed it had increased sensitivity compared to an analysis comprising the traditional tools of contig building and Blast alignment. The previous analysis identified a total of 19 (non-unique) viruses across the 3 samples (25), but Kodoja identified 27. This increased sensitivity comes from the use of *k-mer* profiling, rather than contig assembly. The ability of Kodoja, to identify virus sequences present at lower levels than are detectable using contig building methods, means that viruses could be detected in plants before symptoms appear. This sensitivity was also exemplified when Kodoja was applied to raspberry. RYNV was reported with just 44 reads meeting the stringent classification criteria, and the presence of this virus sequence was confirmed by the RT-PCR.

The work to benchmark Kodoja using existing datasets gave insights into the difficulty of viral sequence identification in mixed (comprising both plant and potentially viral, bacterial and fungal nucleotides) RNA-seq datasets. One key complexity, when a workflow does not include contig building, is the miss-classification of viruses, which arises due to the small evolutionary distances existing between some viral taxa. This was the case when Kodoja identified GYSVd1 as GYSVd2 and GLRaV2 as GRSLaV. GYSVd1 and GYSVda are viroids, that have a single stranded circular RNA genome that does not code for protein. Hence, the Kodoja assignment for this viroid was made only at the nucleotide level, and this could explain its incorrect classification. In addition, GLRaV2 was incorrectly classified as GRSLaV. GLRaV2 is known to be the closest related virus to GRSLaV within the *Closteroviridae* family (30), with between 71-79% sequence identify across 9 ORFs and this could explain why the *k-mer*

399 analysis made an incorrect classification. The raspberry analysis showed that Kodoja
400 reports viruses even if they are present at reads at low levels.
401
402 The detection of known viruses using Kodoja is dependent upon the virus dataset used
403 to generate the *k-mer* databases for Kraken (16) and Kaiju (17). The size of the
404 databases will greatly influence both the sensitivity and the speed of the workflow. We
405 have used a dataset derived from RefSeq (v89)(20) which comprises 7946 non-
406 redundant viral genome sequences. This means that one virus is represented by a
407 single reference sequence and variants are excluded. Hence, the workflow is in some
408 ways restrictive, and could potentially leave some sequences unclassified or miss-
409 classified if they are derived from diverse sequence variants. An alternative to RefSeq
410 would be Genbank (31), which comprises 2.7 million redundant viral sequences
411 (v228) and includes virus variants. However, creating *k-mer* databases for such a large
412 dataset would be prohibitively expensive in terms of time taken for the database build
413 and running the kodoja_search module. A trade-off between run time and sensitivity
414 could potentially be achieved by using a new database Reference Viral database
415 (RVDB)(32). This database includes a clustered set of virus sequences, extracted
416 from Genbank (31) which comprises 561,676 representatives. This clustered
417 database was designed to retain viral diversity and reduce redundancy (32). It would
418 be possible to use this dataset to generate *k-mer* databases for Kodoja that would
419 increase sensitivity further, without completely compromising speed.
420
421 The application of the Kodoja workflow to RNA-seq data from raspberry demonstrated
422 that field-grown raspberry plants can frequently be infected with multiple viruses, and
423 that relying on visual symptoms to identify viruses is often not possible. In addition,
424 this work clearly demonstrated the limitations of primer-based methods for virus
425 detection (RT-PCR and PCR). The innate variability in the nucleotide sequence of
426 plant viruses means that it is very difficult/impossible to design diagnostic primers that
427 can detect many/all isolates of the same virus. For RYNV, the primer pair 1752/1753
428 gave strong amplification of the isolate carried within our positive control plant but
429 could not amplify the virus in D5 and D6.
430
431 The prediction of BRSV, a nepovirus, in D5 and the creation of a new PCR-primer pair
432 based on the D5 sequences, represents a step forward in virus testing for raspberry.
433 Nepoviruses are soil-borne, nematode-transmitted, viruses that are recognized as
434 important pathogens of many crops, including raspberries (33). Historically, when
435 serological reactions and host ranges were used to characterise viruses, BRSV was
436 thought to be an isolate of *Tomato black ring virus* (TBRV) (34). However, it is now
437 clear that BRSV is a different virus to TBRV (35), and the BRSV test we have designed
438 here will now become part of the battery of molecular tests we use for virus testing of
439 raspberry.
440
441 The sequencing and analysis of small RNAs (sRNA-seq) has also proved successful
442 in detecting siRNAs duplexes induced by plant viruses (36), and a specific workflow

has been developed for this purpose (12). Whilst Kodoja is optimized for RNA-seq datasets, we did apply Kodoja to a previously published sRNA-seq dataset from Grapevine (37) (unpublished data), however the success of Kodoja was less clear than for the RNA-seq datasets. Using Kodoja, we detected all viruses reported in the original study, but in addition a further 16 viruses were detected. However, these additional viruses could not be validated as the read counts were low and made contig building impossible. Further optimization and benchmarking would be required for Kodoja to be used effectively on sRNA-seq datasets.

The testing and application of Kodoja, has exemplified its ability to be used successfully for virus identification in RNA-seq datasets. Kodoja is the first workflow to apply a *k-mers* analysis method for virus detection specifically in plants, and in addition it is the first plant virus detection workflow to be made available through BioConda and as a Galaxy application. This accessibility will make it available to a wide range of researchers, working on diverse plant species. Our application of the workflow to raspberry has highlighted its potential to develop new primers to enhance serological testing and such advances will also be possible with other crops.

Author Statements

Funding

This work was supported by the Biotechnology and Biological Sciences Research Council [BB/N023293/1]. The work of LT, SJ, SM and PC was additionally supported by the Scottish Government's Rural and Environment Science and Analytical Services division (RESAS).

Conflict of Interest

No conflicts of interest are declared.

Acknowledgments

We would like to thank the Bioinformatics Unit, School of Medicine, University of St Andrews, North Haugh, St Andrews, UK, KY16 9TF, for helpful discussions.

References

1. Nicaise V. Crop immunity against viruses: outcomes and future challenges. Front Plant Sci. 2014;5:1–18.
2. Valkonen JPT. Viruses: Economical Losses and Biotechnological Potential. In: Potato Biology and Biotechnology: Advances and Perspectives. Elsevier; 2007. p. 619–42.
3. Sasaya T, Nakazono-Nagaoka E, Saika H, Aoki H, Hiraguri A, Netsu O, et al. Transgenic strategies to confer resistance against viruses in rice plants. Front Microbiol. 2013;4:1–11.
4. Lamichhane JR, Venturi V. Synergisms between microbial pathogens in plant disease complexes: a growing trend. Front Plant Sci . 2015;06:1–12.

5. Martin RR, MacFarlane S, Sabanadzovic S, Quito D, Poudel B, Tzanetakis IE. Blackberry Yellow Vein Disease (BYVD) Complex and Associated Viruses. Plant Dis. 2013;97(2):168–82.
6. Mumford R, Boonham N, Tomlinson J, Barker I. Advances in molecular phytodiagnostics - New solutions for old problems. Eur J Plant Pathol. 2006;116(1):1–19.
7. Boonham N, Kreuze J, Winter S, van der Vlugt R, Bergervoet J, Tomlinson J, et al. Methods in virus diagnostics: From ELISA to next generation sequencing. Virus Res . 2014;186:20–31.
8. Wu Q, Luo Y, Lu R, Lau N, Lai EC, Li W-X, et al. Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. Proc Natl Acad Sci U S A. 2010;107(4):1606–11.
9. Jones S, Baizan-Edge A, MacFarlane S, Torrance L. Viral diagnostics in plants using next generation sequencing: Computational analysis in practice. Front Plant Sci. 2017;8:1–10.
10. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med . 2014;370(25):2408–17.
11. Flygare S, Simmon K, Miller C, Qiao Y, Kennedy B, Di Sera T, et al. Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. Genome Biol . 2016;17(1):111.
12. Zheng Y, Gao S, Padmanabhan C, Li R, Galvez M, Gutierrez D, et al. VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. Virology . 2017;500:130–8.
13. Trifonov V, Rabadan R. Frequency Analysis Techniques for Identification of Viral Genetic Data. MBio. 2010;1(3):1–8.
14. Dröge J, Gregor I, Mchardy A. Taxator- tk: Precise taxonomic assignment of aetagenomes by fast approximation of evolutionary neighbourhoods. Bioinformatics. 2014;31(6):817–24.
15. Malhotra S, Sowdhamini R. Genome-wide survey of DNA-binding proteins in Arabidopsis thaliana: analysis of distribution and functions. Nucleic Acids Res . 2013 Aug;41(15):7212–9.
16. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol . 2014;15(3):R46.
17. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun . 2016;7:1–9.
18. Dale R, Grüning B, Sjödin A, Rowe J, Chapman BA, Tomkins-Tinch CH, et al. Bioconda: A sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018;15:475–6.
19. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res . 2016;44:gkw343.
20. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.
21. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, et al. Linking virus genomes with host taxonomy. Viruses. 2016;8(3):10–5.
22. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina

sequence data. Bioinformatics . 2014 May 28;1–7.
23. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
24. Da Silva C, Zamperin G, Ferrarini A, Minio A, Dal Molin A, Venturini L, et al. The High Polyphenol Content of Grapevine Cultivar Tannat Berries Is Conferred Primarily by Genes That Are Not Shared with the Reference Genome. Plant Cell . 2013;25(12):4777–88.
25. Jo Y, Choi H, Kyong Cho J, Yoon J-Y, Choi S-K, Kyong Cho W. In silico approach to reveal viral populations in grapevine cultivar Tannat using transcriptome data. Sci Rep . 2015;5(1):15841.
26. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson D a, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol . 2011;29(7):644–52.
27. Silvester N, Alako B, Amid C, Cerdeño-Tarrága A, Clarke L, Cleland I, et al. The European Nucleotide Archive in 2017. Nucleic Acids Res. 2018;46(D1):D36–40.
28. VanBuren R, Bryant D, Bushakra JM, Vining KJ, Edger PP, Rowley ER, et al. The genome of black raspberry (Rubus occidentalis). Plant J. 2016;87(6):535–47.
29. Macfarlane S, Mcgavin W, Tzanetakis I. Virus Testing by PCR and RT-PCR Amplifi cation in Berry Fruit. In: Lacomme C, editor. Plant Pathology: Techniques and Protocols . Springer; 2015. p. 227–48.
30. Alkowni R, Zhang YP, Rowhani A, Uyemoto JK, Minafra A. Biological, molecular, and serological studies of a novel strain of grapevine leafroll-associated virus 2. Virus Genes. 2011;43(1):102–10.
31. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2013;41(D1):36–42.
32. Goodacre N, Aljanahi A, Nandakumar S, Mikailov M, Khan AS. A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus. mSphere. 2018;3(2):1–18.
33. Martin RR, Polashock JJ, Tzanetakis IE. New and emerging viruses of blueberry and cranberry. Viruses. 2012;4(11):2831–52.
34. Harrison B. Relationship between Beet Ringspot, Potato Bouquet and Tomato Black Ring Viruses. J Gen Microbiol. 1958;18:450–60.
35. Kis S, Salamon P, Kis V, Szittya G. Molecular characterization of a beet ringspot nepovirus isolated from Begonia ricinifolia in Hungary. Arch Virol. 2017;162(11):3559–62.
36. Niu X, Sun Y, Chen Z, Li R, Padmanabhan C, Ruan J, et al. Using small RNA-seq data to detect siRNA duplexes induced by plant viruses. Genes (Basel). 2017;8(6):1–8.
37. Barrero RA, Napier KR, Cunnington J, Liefting L, Keenan S, Frampton RA, et al. An internet-based bioinformatics toolkit for plant biosecurity diagnosis and surveillance of viruses and viroids. BMC Bioinformatics . 2017;18(1):26.
38. Federhen S. The NCBI Taxonomy Database. Nucleic Acids Res. 2012;40(D1):D136–43.

Table and Figure Legends

Table 1.
Information on the RT-PCR primer pairs for the confirmation of four raspberry viruses; Raspberry leaf mottle virus (RLMV), Raspberry leaf blotch virus (RLBV), Rubus yellow net virus (RYNV) and Beet ringspot virus (BRSV) predicted to be present in raspberry samples D5 and D6 by Kodoja.

Table 2.
Kodoja results for the three RNA-sequence datasets from grapevine. The species taxonomic identify from the NCBI Taxonomy database (38) is shown in column 4 (Sp TaxID), the number total number of reads that were classified to each virus species is shown in column 5 (Sp seq), the number of reads classified by both Kaiju and Kraken (stringent for viruses) (Sp seq (S)) is shown in column 6. The results of the contig building and BlastN confirmation process are indicated in column 9. Y indicates that the BlastN alignment assigned the sequences to the same species as Kodoja. N indicates that BlastN assigned the sequences to a different species to Kodoja. The detection of the viruses in the original work is indicated in Column 10.

Table 3. Kodoja results for the 2 RNA-seq datasets from raspberry. The column headers are as described for Table 2. The results of the RT-PCR confirmation experiments are indicated in the last 7 columns, with Y indicating the virus was detected with the specified primer pair and N indicating the virus was not detected.

Figure 1.
Flow diagram summarizing the 3 modules of the Kodoja workflow: kodoja_build, kodoja_search and kodoja_retrieve.

Figure 2.
Images of leaves taken from two Glen Dee raspberry plants grown on a commercial farm in Angus, Scotland, UK. (A) Plant D5 showing major vein yellowing and (B) Plant D6 showing leaf blade yellowing.

Figure 3.
Diagrammatic alignments of selected virus contigs to their reference genomes. (A) Alignment for Grapevine rupestris vein feathering virus (GRVFV) from dataset GV2, (B) Alignment for Grapevine asteroid mosaic-associated virus (GAMaV) from GV3, (C) Alignment for GRVFV from dataset GV3.

Figure 4.
Diagrammatic alignments of the selected virus contigs to their reference genomes. (A) Beet ringspot virus (BRSV) from D5, (B) Rubus yellow net virus (RYNV) from D5, (C)

629 Raspberry leaf mottle virus (RLMV) from D6, (D) RYNV from D6 and (E) Raspberry
630 leaf blotch virus from D6.
631
632 **Figure 5**. Virus detection by RT-PCR in raspberry. (A) Raspberry yellow net virus
633 (RYNV) amplified with primers 3470/3471. (B) BRSV amplified with primers
634 3472/3473. Within each panel, lane 1 is kilobase DNA markers (500bp and 250bp
635 markers are indicated), lane 2 is water only amplification, lane 3 is sample D5 RNA,
636 lane 4 is sample D6 RNA, lane 5 is RNA extracted from a known RYNV-infected (A)
637 or BRSV-infected (B) plant.

Virus	Primer pair number	Sequence
RLMV	991/992	CGAAACTTYTACGGGGAAC/ CCTTTGAAYTCTTTAACATCGT
RLBV	1095/1287	CACCATCAGGAACTTGTAATGTTT/ ATCCAGTAGTGAACTCC
	1491/1492	GGTGAATGAGTTCTATACTAAGAC/ TCGACACTCATCAGAATAATTGCC
	1495/1496	GAATTGCAAGGCAAATCAGC/ GCATTCTGACCATTCCTCAAA
	2113/2114	CAAAGAGTTGCGTCATGTCA/ CCATTCCAGTATTCAACATCTGA
RYNV	1752/1753	TCCAAAACCTCCCAGACCTAAAAC/ ATAATCGCAAAAGGCAAGCCAC
	3470/3471	ATAATCACAAAAAGCTAACCAC/ TCCAGAACCTCCCAGACCTCAAAC
BRSV	3472/3473	GCCACTGTACAGCCCATCTT/ AGAGTAAGATCAGAGGCACGT

Table 1

	Virus species	Acronym	Sp TaxID	Sp seqs	Sp seq (S)	Genus	Genus seqs	BlastN	Jo et al (2015)
GV1	Grapevine rupestris stem pitting-associated virus	GRSPaV	196400	63151	7057	Foveavirus	103	Y	Y
	Grapevine leafroll-associated virus 1	GLRaV1	47985	3	1	Ampelovirus	0	Y	N
	Apple mosaic virus	ApMV	12319	1	1	Ilarvirus	0	Y	N
	Hop stunt viroid	HSVd	12893	1636	0	Hostuviroid	0	Y	Y
	Grapevine yellow speckle viroid 1	GYSVd1	12904	1371	0	Apscaviroid	14	Y	Y
	Grapevine yellow speckle viroid 2	GYSVd2	46342	251	0	Apscaviroid	14	N	N
GV2	Grapevine rupestris stem pitting-associated virus	GRSPaV	196400	305827	54432	Foveavirus	463	Y	Y
	Grapevine Pinot gris virus	GPGV	1051792	2116	2026	Trichovirus	22	Y	Y
	Potato virus Y	PVY	12216	2449	1007	Potyvirus	232	Y	Y
	Grapevine rupestris vein feathering virus	GRVFV	204933	566	183	Marafivirus	0	Y	N
	Grapevine leafroll-associated virus 2	GLRaV2	64003	447	154	Closterovirus	34	Y	Y
	Cucumber mosaic virus	CMV	12305	50	40	Cucumovirus	0	Y	Y
	Grapevine rootstock stem lesion associated virus	GRSLaV	167634	109	16	Closterovirus	34	N	Y
	Alfalfa mosaic virus	AMV	12321	19	15	Alfamovirus	0	Y	Y
	Parietaria mottle virus	PMoV	64958	1	1	Ilarvirus	0	Y	N
	Hop stunt viroid	HSVd	12893	1604	0	Hostuviroid	0	Y	Y
	Grapevine yellow speckle viroid 1	GYSVd1	12904	440	0	Apscaviroid	5	Y	Y
	Grapevine yellow speckle viroid 2	GYSVd2	46342	129	0	Apscaviroid	5	N	N
GV3	Grapevine rupestris stem pitting-associated virus	GRSPaV	196400	20645	3781	Foveavirus	22	Y	Y
	Grapevine Pinot gris virus	GPGV	1051792	234	223	Trichovirus	1	Y	Y
	Grapevine asteroid mosaic-associated virus	GAMaV	103724	236	163	Marafivirus	3	Y	N
	Grapevine rupestris vein feathering virus	GRVFV	204933	84	27	Marafivirus	3	Y	N
	Grapevine leafroll-associated virus 2	GLRaV2	64003	29	10	Closterovirus	0	Y	Y
	Grapevine rootstock stem lesion associated virus	GRSLaV	167634	16	2	Closterovirus	0	Y	N
	Hop stunt viroid	HSVd	12893	945	0	Hostuviroid	0	Y	Y
	Grapevine yellow speckle viroid 1	GYSVd1	12904	129	0	Apscaviroid	2	Y	Y
	Grapevine yellow speckle viroid 2	GYSVd2	46342	25	0	Apscaviroid	2	N	N

Table 2

	Virus species	Acronym	Sp TaxID	Sp seqs	Sp seqs (S)	Genus	Genus seqs	BlastN	RT-PCR Primers						
									RLMV 911/912	RLBV 1095/1094	RLBV 1495/1496	RLBV 2113/2114	RYNV 1752/1753	RYNV 3470/3470	BRSV 3472/3473
D5	Beet ringspot virus	BRSV	191547	80	36	Nepovirus	10	Y							Y
	Rubus yellow net virus	RYNV	198310	287	24	Badnavirus	0	Y					N	Y	
	Raspberry leaf blotch virus	RLBV	1980431	16	16	Emaravirus	0	Y		N	N	N			
	Cherry leaf roll virus	CLRV	12615	11	6	Nepovirus	10	Y							
	Tomato black ring virus	TBRV	12275	7	2	Nepovirus	10	Y							
	Pelargonium leaf curl virus	PLCV	35280	1	1	Tombusvirus	0	Y							
D6	Raspberry leaf mottle virus	RLMV	326941	15011	912	Closterovirus	51	Y	Y						
	Raspberry leaf blotch virus	RLBV	1980431	225	186	Emaravirus	0	Y		N	Y	N			
	Rubus yellow net virus	RYNV	198310	629	44	Badnavirus	2	Y					N	Y	
	Cherry leaf roll virus	CLRV	12615	20	10	Nepovirus	0	Y							
	Tobacco mosaic virus	TMV	12242	2	1	Tobamovirus	1	Y							

Table 3

Figure 1.

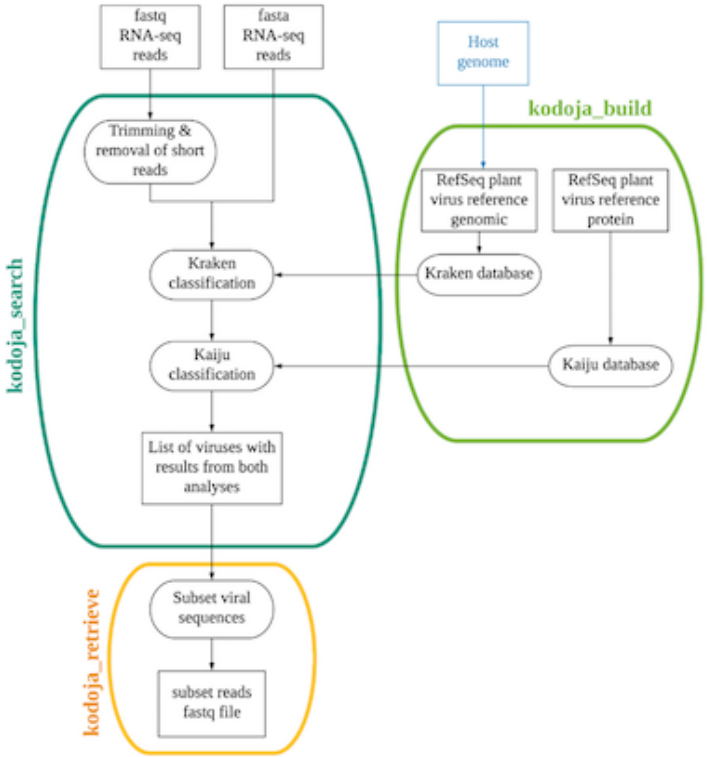
fastq RNA-seq reads
fasta RNA-seq reads
Host genome
kodoja_build
Trimming & removal of short reads
RefSeq plant virus reference genomic
RefSeq plant virus reference protein
Kraken classification
Kraken database
kodoja_search
Kaiju classification
Kaiju database
List of viruses with results from both analyses
Subset viral sequences
kodoja_retrieve
subset reads fastq file

Figure 2.

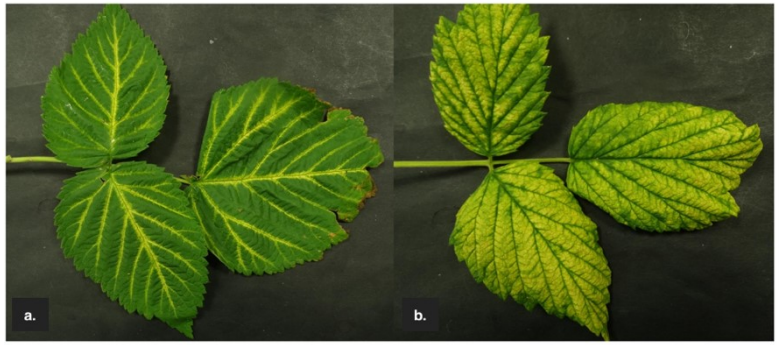

Figure 3.

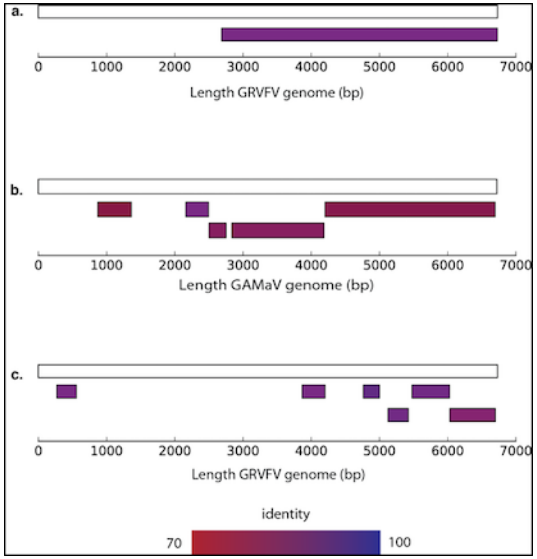
a.
0
1000
2000
3000
4000
5000
6000
7000
Length GRVFV genome (bp)
b.
0
1000
2000
3000
4000
5000
6000
7000
Length GAMaV genome (bp)
c.
0
1000
2000
3000
4000
5000
6000
7000
Length GRVFV genome (bp)
identity
70
100

Figure 4.

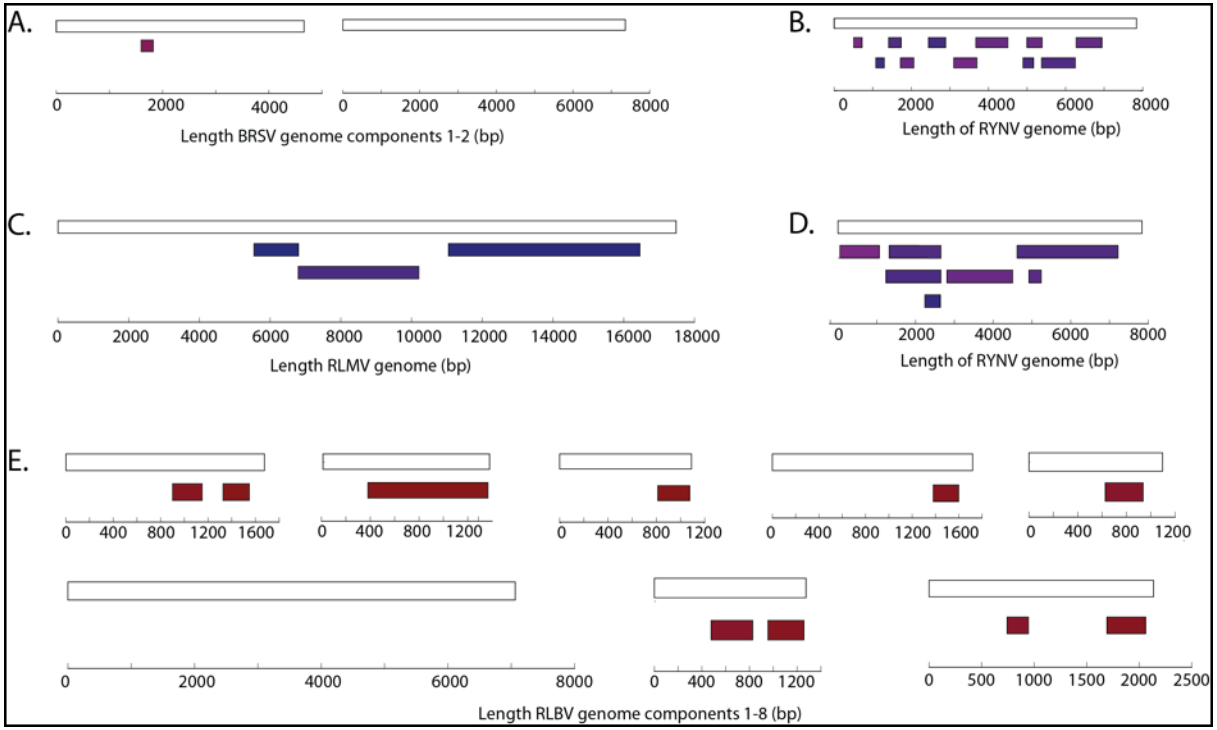
A.
0
2000
4000
0
2000
4000
6000
8000
Length BRSV genome components 1-2 (bp)
B.
0
2000
4000
6000
8000
Length of RYNV genome (bp)
C.
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
Length RLMV genome (bp)
D.
0
2000
4000
6000
8000
Length of RYNV genome (bp)
E.
0
400
800
1200
1600
0
400
800
1200
0
400
800
1200
0
400
800
1200
1600
0
400
800
1200
0
2000
4000
6000
8000
0
400
800
1200
0
500
1000
1500
2000
2500
Length RLBV genome components 1-8 (bp)

Figure 5.