Improving Metagenomics-based Plant Virus Identification by Improving Reference Databases

Gercken M.[1,2], Kaur S.[1,2], Mazloom R.[3], Belay K.[1,2], Rodriguez Salamanca L.[1], Heath L. S.[3], Vinatzer B. A.[1]

[1] School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA 24061, U.S.A.; [2] Graduate Program in Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, VA 24061, U.S.A.; [3] Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, U.S.A.

Presenter: Mitchell Gercken
Email: mitchellgercken@vt.edu
PI: Boris A. Vinatzer
Email: vinatzer@vt.edu

Introduction

- The use of metagenomics makes it possible to identify viruses directly from RNA of infected plants.
- Current viral database used for metagenomic classification lack equal distribution and/or equal representation of each sequence present → ***Causes emerging pathogens to be misclassified***.
- Life Identification Numbers (LINs) offer an approach to organize genomes based on Average Nucleotide Identity (ANI) thresholds.
- Here we provide a refined LIN representative-based Kraken 2 database comprised on ~71,000 plant viral sequences.

Materials and Methods

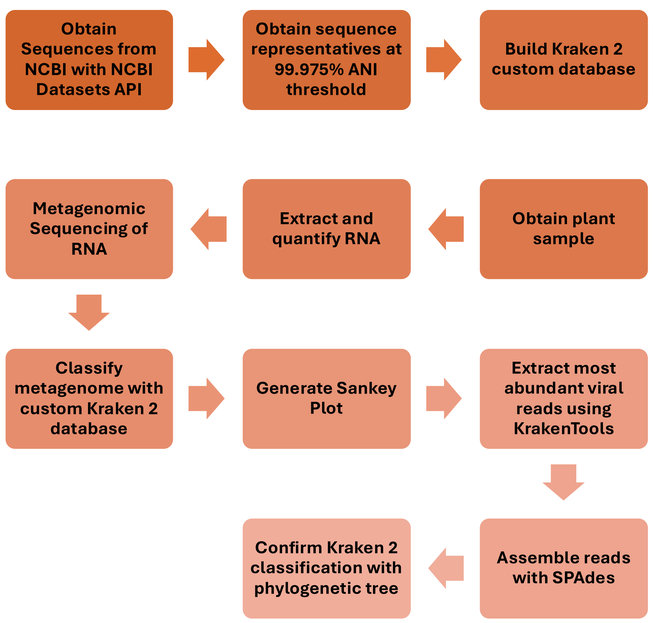

Figure 1. Flowchart of methods.

Databases	Official viral RefSeq	Contaminated Custom	Refined Custom
Number of Sequences	18,654 (2,527 plant viruses)	30,860 plant viruses	71,012 plant viruses

Table 1. Number of plant viral sequences in each database.

Phylogenetic Tree Generation

- Most abundant classified viral reads were extracted from the plant metagenome and assembled into contigs using SPAdes (version v4.1.0).
- NCBI BLASTn was used to confirm the classification of each assembled contig.
- Sequence alignment conducted with MAFFT (version v7.526).
- IQ-tree (version 2.4.0) was used to generate core genome tree.
- RStudio (version 2025.05.0+496) was used to generate tree visualizations.

Results

Official viral RefSeq Database

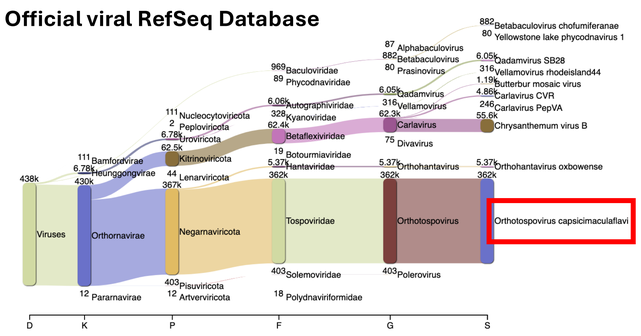

Figure 2. Sankey plot of metagenome sample using Kraken 2 report file with the viral RefSeq database.

Contaminated Custom Database – 99.975% ANI

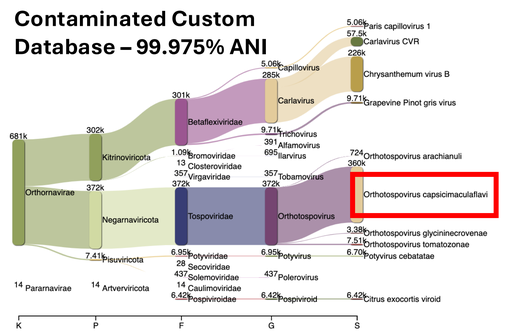

Figure 3. Sankey plot of metagenome sample using Kraken 2 report file with the contaminated custom database.

Refined Custom Database – 99.975% ANI

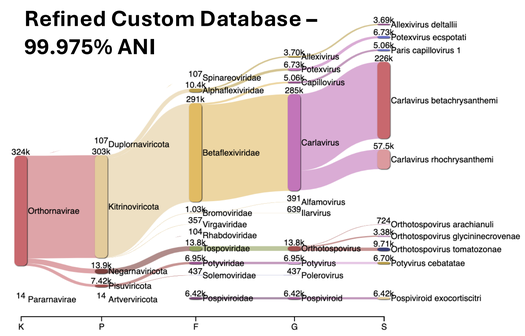

Figure 4. Sankey plot of metagenome sample using Kraken 2 report file with the refined custom database.

Phylogenetic Tree (IQ-tree)

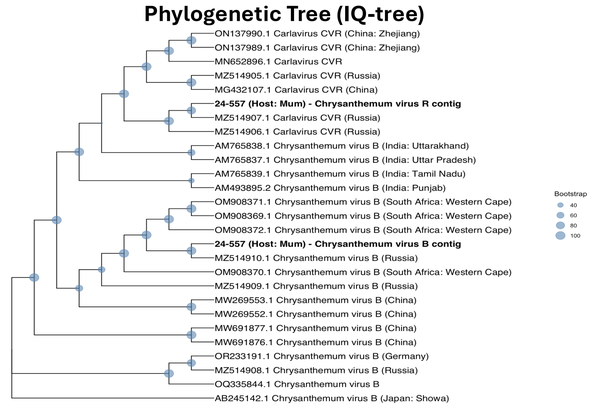

Figure 6. Phylogenetic tree of assembled contigs.

Sourmash Distance Tree

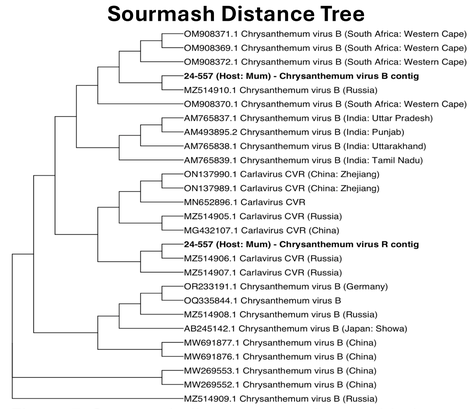

Figure 7. Sourmash distance tree of assembled contigs.

Refined Custom Sourmash Database – 99.975% ANI

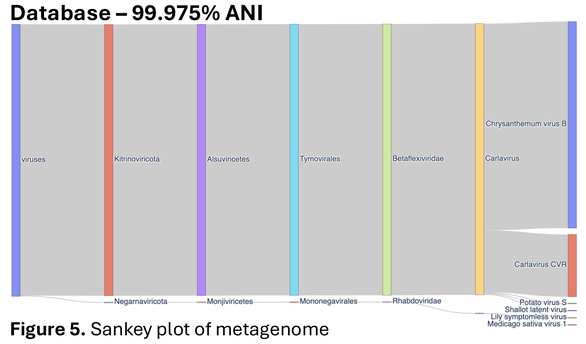

Figure 5. Sankey plot of metagenome sample using Sourmash with the refined custom database.

Tanglegram

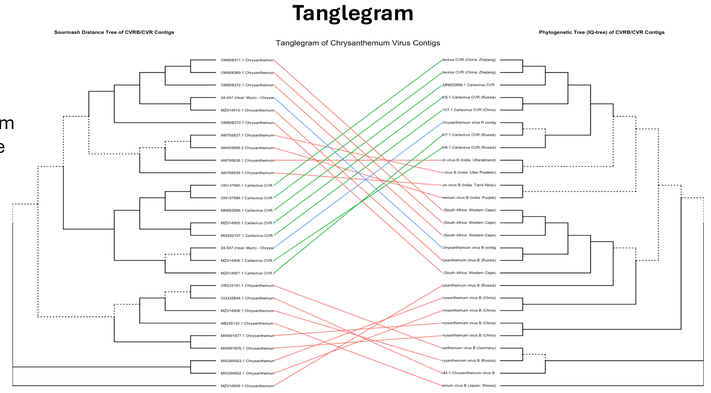

Figure 8. Tanglegram of phylogenetic tree and sourmash distance tree.

Sample Name	Host	Number of most abundant reads	Classification
24-171	Geranium	10,500	Prunus necrotic ringspot virus
24-172	Lungwort	62,200	Gaillardia latent virus
24-243	Butterfly weed	41,500	Cucumber mosaic virus
24-352	Phlox	1,740,000	Alternanthera mosaic virus
24-357	Viola	16,700	Prune dwarf virus
24-440	Tulip	3,000,000	Clover yellow mosaic virus
24-544	Hydrangea	5,760,000	Hydrangea ringspot virus
24-557*	Chrysanthemum	226,000	Chrysanthemum virus B
25-072	Hydrangea	10,100,000	Hydrangea ringspot virus

Table 2. Viral species found in all plant viral metagenomes analyzed (* indicates results shown for this sample).

Conclusions

- The custom Kraken 2 database improves upon currently established viral databases by:
  - Reduces the risk of false positives and negatives.
  - At least a 2-fold increase in the number of viral reads classified by Kraken 2.
  - More complete viral genomes recovered.

Future Research

- Continue to refine the custom database.
- Start introducing animal viruses into the database.
- Attempt to automate pathogen classification with a webserver similar to GenomeRxiv.
- Integrate analysis process into the VT-PLANS webserver → **Janet Lorv: P-158.**

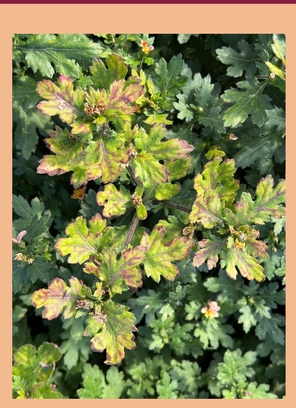

Image 1. Symptoms on chrysanthemum sample 24-557 in the cut flower farmer field.

DBI-2018911