Supplementary Materials

Hardware-Aware RNA-seq Diagnostics: Cassava Virus Detection via Cloud and AI

Elisson Silva, Paolo Margaria, Rosana Blawid, Eder Jorge de Oliveira, Stephan Winter, and Stefan Blawid

---

To ensure the detection of known cassava infecting viruses, viral reference data where augmented by the nucleotide sequences listed in Tab. SI in the investigations that do not use the complete NCBI nt data base.

TABLE SI

Reference genomes of cassava infecting viruses. Searching for viral sequences (date 12/09/2024) with the host's taxonomic ID 3983 (*Manihot esculenta*, Cassava) gives 19 RefSeq genomes, 971 complete and 2447 incomplete nucleotide sequences. Genome entries highlighted in the table with a * identify genomes with nucleotide sequences that do specifiy *Manihot esculenta* as host although the RefSeq not. Some cassava infecting virus species do not have a RefSeq genome.

RefSeq Assembly	Virus	Genus	Family
`GCF_000849345.1`	Cassava common mosaic virus*	*Potexvirus*	*Alphaflexiviridae*
`GCF_002116295.1`	Cassava virus X	*Potexvirus*	*Alphaflexiviridae*
`GCF_018580175.1`	Cassava Colombian symptomless virus	*Potexvirus*	*Alphaflexiviridae*
`GCF_000898675.1`	Cassava mosaic Madagascar alphasatellite	*Colecusatellite*	*Alphasatellitidae*
`GCF_000885215.1`	Cassava virus C	*Ourmiavirus*	*Botourmiaviridae*
`GCF_000929575.1`	Cassava Ivorian bacilliform virus	*Anulavirus*	*Bromoviridae*
`GCF_000838625.1`	Cassava vein mosaic virus*	*Cavemovirus*	*Caulimoviridae*
`GCF_000857205.1`	African cassava mosaic virus	*Begomovirus*	*Geminiviridae*
`GCF_004786835.1`	African cassava mosaic Burkina Faso virus	*Begomovirus*	Geminiviridae
`GCF_000859785.1`	East African cassava mosaic virus*	*Begomovirus*	*Geminiviridae*
`GCF_000858965.1`	East African cassava mosaic Cameroon virus*	*Begomovirus*	*Geminiviridae*
`GCF_000882315.1`	East African cassava mosaic Kenya virus	*Begomovirus*	*Geminiviridae*
`GCF_000912855.1`	East African cassava mosaic Malawi virus	*Begomovirus*	*Geminiviridae*
`GCF_000845125.1`	East African cassava mosaic Zansibar virus	*Begomovirus*	*Geminiviridae*
`GCF_000838365.1`	South African cassava mosaic virus*	*Begomovirus*	*Geminiviridae*
`GCF_000896655.1`	West African Asystasia virus 1*	*Begomovirus*	*Geminiviridae*
`GCF_000895455.1`	Cassava mosaic Madagascar virus	*Begomovirus*	*Geminiviridae*
`GCF_000846665.1`	Indian cassava mosaic virus*	*Begomovirus*	*Geminiviridae*
`GCF_000858725.1`	Sri Lankan cassava mosaic virus*	*Begomovirus*	*Geminiviridae*
`GCF_000907355.1`	Ageratum yellow vein China virus*	*Begomovirus*	*Geminiviridae*
`GCF_000882835.1`	Euphorbia yellow mosaic virus*	*Begomovirus*	*Geminiviridae*
`GCF_000838665.1`	Potato yellow mosaic virus*	*Begomovirus*	*Geminiviridae*
`GCF_000889935.1`	Soybean chlorotic blotch virus*	*Begomovirus*	*Geminiviridae*
`GCF_000859325.1`	Tobacco curly shoot virus*	*Begomovirus*	*Geminiviridae*
`GCF_000858105.1`	Begomovirus-associated DNA-II*		
`GCF_000857145.1`	Begomovirus-associated DNA-III*		
`GCF_000919515.1`	Cassava associated gemycircularvirus 1	*Gemycircularvirus*	*Genomoviridae*
`GCF_000884835.1`	Cassava brown streak virus	*Ipomovirus*	*Potyviridae*
`GCF_000888855.1`	Ugandan cassava brown streak virus	*Ipomovirus*	*Potyviridae*
`GCF_029888435.1`	Cassava torrado-like virus	*Torradovirus*	*Secoviridae*
`GCF_000866525.1`	Ageratum yellow vein China betasatellite*	*Betasatellite*	*Tolecusatellitidae*
`GCF_000861305.1`	Malvastrum yellow vein betasatellite*	*Betasatellite*	*Tolecusatellitidae*
`GCF_002008795.1`	Cassava satellite virus		
14 sequences	Cassava alphaflexivirus		*Alphaflexiviridae*
`MW929099`, `MW929100`	Cassava ampelovirus 1	*Ampelovirus*	*Closteroviridae*
13 sequences	Manihot esculenta associated ampelovirus 1	*Ampelovirus*	*Closteroviridae*
`ON924323`, `ON924324`	Cassava Congo Sequivirus	*Sequivirus*	*Secoviridae*
4 sequences	Cassava torrado-like virus 2	*Torradovirus*	*Secoviridae*
20 sequences	Cassava polero-like virus	*Polerovirus*	*Solemoviridae*
7 sequences	Cassava frogskin virus		
57 sequences	Cassava frogskin associated virus		

Bacteria

Given the limited understanding of plant microbiomes, we could not rely on prior work to identify typical bacterial genera for inclusion. Instead, we characterized a random RNA-seq dataset from diseased cassava plants in the DSMZ collection. Taxonomic composition was determined using Kraken2 and Kaiju, with results summarized in Fig.1 showing the predominant bacteria found in cassava microbiomes.

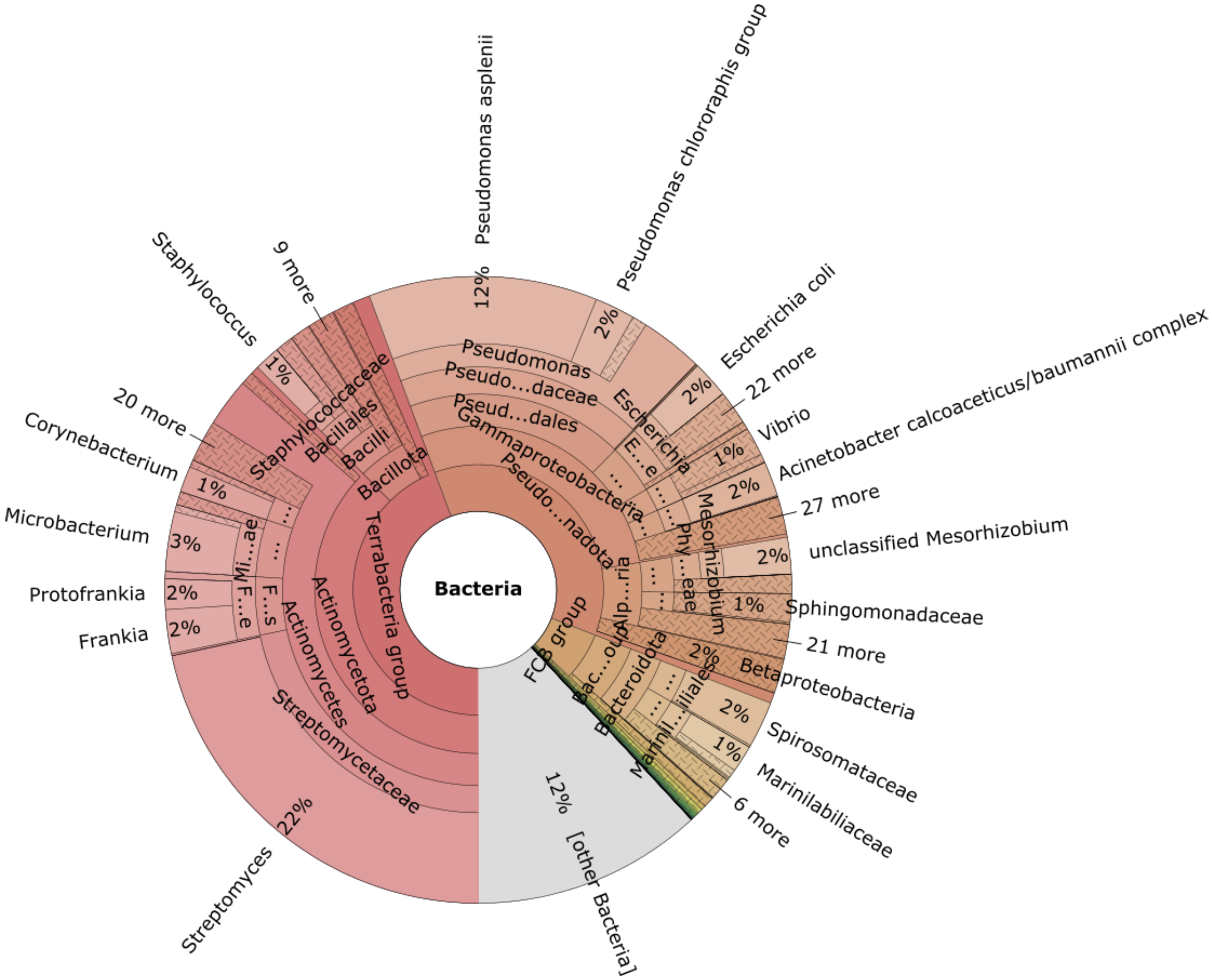
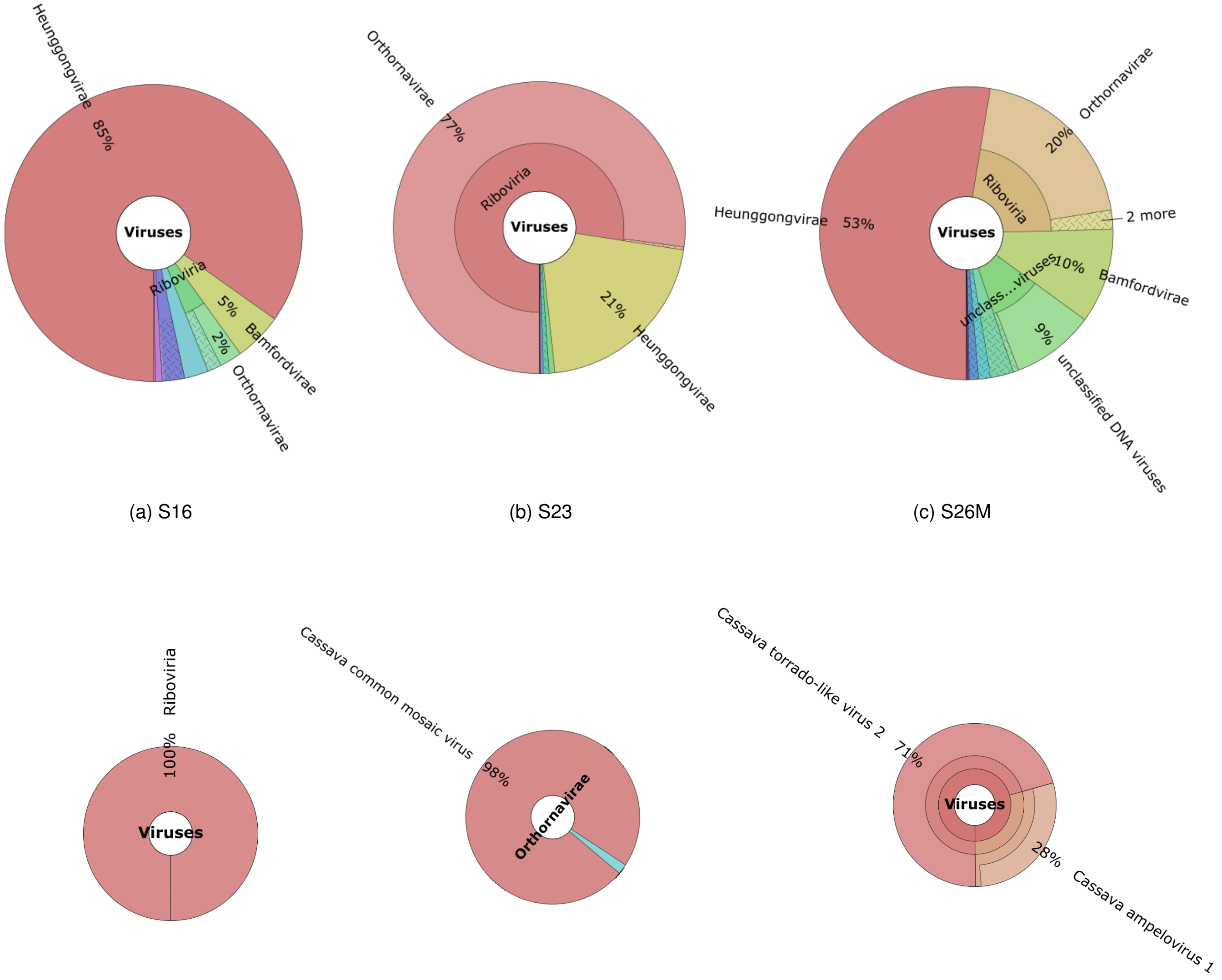


Fig. S1. Kaiju classification of reads obtained from a cassava sample showing the bacteria genera Acinetobacter (tax id 469), Streptomyces (tax id 1883) and Pseudomonas (tax id 286) and the FCB group (tax id 1783270) accounting for ca. 48% of all bacterial reads in the given sample.

Misclassification Using Reduced Reference Base

Viral Kraken2 identified several additional non-plant viruses, such as phages. This misclassification apparently occurred because portions of the cassava genome were mistaken for viral sequences using a virus sequence-database only. The encountered presumably “viral” reads in the *Manihot esculenta* genome were mainly annotated as tailed bacteriophages from the class *Caudoviricetes*, giant viruses from the class *Megaviricetes*, and negative-strand RNA viruses from the order *Bunyavirales*. The latter also includes plant-infecting viruses and might bias virus discovery.



(a) S16 (b) S23 (c) S26M

(d) S16 (e) S23 (f) S26M

Fig. S2. Kraken2 classification of the trimmed reads from samples S16, S23 and S26M. In the upper row (a,b,c,) only the viral reference database is used during analysis, in the lower row (d,e,f) the full NCBI nt database. The use of the small viral database leads to the prediction of an artificial virus diversity that is not real (large number of false positives) including bacteriophages belonging to the kingdom *Heunggongvirae* (a,b,c). Moreover, in the kingdom *Orthornavirae* most of the identified viruses are of different phyla than the known cassava infecting viruses. On the other hand, employing the NCBI nt database leads to the correct classification of reads originating from (e) cassava common mosaic virus (*Alphaflexiviridae*) and Cassava torrado-like virus (*Secoviridae*) for S23 and (f) cassava torrado-like virus 2 (*Secoviridae*) and cassava ampelovirus 1 (*Closteroviridae*) for S26M (f). The sample S16 is virus free since only four viral reads are identified (d).

Viromes of Selected Samples

The semi-automatic pipeline employed at DSMZ revealed a complex virome for sample S26N and a novel *Bromoviridae* virus in S6G. Tab. II summarizes all discovered sequences.

TABLE SII
ASSEMBLED CONTIGS FROM S26N AND S6G GENERATED BY MANUAL CURATION.

Contig	Length (nt)	#Mapped Reads	Mean Coverage
S26N_cassava torrado-like virus 2 RNA1_Seq.1	7451	4,726	90.2
S26N_cassava torrado-like virus 2 RNA1_Seq.2	7518	9,852	187.3
S26N_cassava torrado-like virus 2 RNA2_Seq.1	4638	17,230	528.3
S26N_cassava torrado-like virus 2 RNA2_Seq.2	4635	5,334	162.8
S26N_cassava ampelovirus 1	13,572	9,515	101.8
S26N_cassava ampelovirus 2	14,150	7,544	76.9
S26N_cassava ampelovirus 3_Seq.1	13,875	2,531	26.6
S26N_cassava ampelovirus 3_Seq.2	13,767	1,475	15.5
S26N_cassava ampelovirus 3_Seq.3	13,547	1,053	11.2
S26N_cassava satellite virus	1,163	5,337	544.1
S6G_cassava-associated bromo-like virus RNA 1	3,233	1,593	68.5
S6G_cassava-associated bromo-like virus RNA 2	2,551	1,689	89.3
S6G_cassava-associated bromo-like virus RNA 3	2,321	1,734	105.5