Lazypipe3: Customizable Virome Analysis Pipeline Enabling Fast and Sensitive Virus Discovery from NGS data

Ilya Weinstein[1,2*], Olli Vapalahti[1,2,3], Ravi Kant[1,2,4#] and Teemu Smura[2,3#]

[1] Department of Veterinary Biosciences, University of Helsinki, 00014 Helsinki, Finland

[2] Department of Virology, University of Helsinki, 00014 Helsinki, Finland

[3] Department of Virology and Immunology, University of Helsinki and Helsinki University Hospital, 00014 Helsinki, Finland

[4] Department of Tropical Parasitology, Institute of Maritime and Tropical Medicine, Medical University of Gdansk, 81-519 Gdynia, Poland

* To whom correspondence should be addressed. Email: Ilja.Weinstein@helsinki.fi

# Equal contributions

ABSTRACT

Metagenomic next-generation sequencing (mNGS) has revolutionized virus detection across clinical, animal, and wildlife samples, enabling the discovery of both known and novel viruses. However, efficient identification of viromes from mNGS data depends heavily on bioinformatics pipelines. To contribute to publicly available pipelines, we introduce here Lazypipe3 with multiple new features including flexible annotation strategies that can be adjusted to match different datasets and research goals. Using a synthetic benchmark, we show that a simple two-round annotation strategy designed for speed and low false positive rate required in virus diagnostics, can reduce execution time to a fraction of BLAST search without any loss in accuracy. Additionally, using real data from mosquito samples we show that annotation strategies based on combinations of fast-to-accurate homology searches reduce execution time from 5 to 20-fold compared to BLASTN/BLASTP baseline. Using results from one of these “annotation chains” we were able to discover and characterise multiple novel whole-genome viral sequences that were missed by previous Lazypipe1 and BLASTX analysis. In Lazypipe3 we also implemented and benchmarked an improved background filtering and novel interactive reports for visualising coverage and placement of contigs relative to reference genomes. Lazypipe3 represents a highly efficient and versatile tool for virome analysis, offering customizable and transparent workflows that can facilitate virus discovery and diagnostics in diverse mNGS applications.

Key words: metagenomics; virus; discovery; virome; mNGS; bioinformatics; pipeline

1 INTRODUCTION

Rapid development in metagenomic next generation sequencing (mNGS) technology has enabled virologists to assess the true diversity of viruses in various samples, included clinical, animal and wildlife (Bassi *et al.*, 2022). This is particularly true for viruses that do not propagate in cell culture and novel or emerging viruses that fail to be detected with conventional PCR- or antigen-based methods. mNGS is a high-throughput, unbiased technology that can identify both known and novel pathogens and has great potential for virus discovery, surveillance and broad-spectrum clinical diagnostics (Alawi *et al.*, 2019; Kalantar *et al.*, 2020; Bassi *et al.*, 2022).

Virome identification with mNGS is extensively dependent on bioinformatic analysis and it is increasingly recognised that applied bioinformatic methods will significantly influence the results (Kalantar *et al.*, 2020; de Vries *et al.*, 2021). The bioinformatics analysis of mNGS data is an actively evolving field, with a wide range of tools available for different steps of the analysis — including read preprocessing, assembly, and annotation — as summarised by Bassi et al. and Ibañez-Lligoña et al. (Bassi *et al.*, 2022; Ibañez-Lligoña *et al.*, 2023). Since application of stepwise tools require specific knowledge in bioinformatics, pipelines supporting complete workflows have been developed. These can be divided into three types: pipelines for characterising viral diversity and abundancy, pipelines for virus diagnostics or identification and pipelines for virus discovery. Here we will focus on pipelines for virus diagnostics and discovery. Pipelines developed for, or applicable to, viral diagnostics focus on high sensitivity and specificity since these will translate into accurate clinical diagnoses and treatment. Open source pipelines that have been applied in virus diagnostics include metaMix (Morfopoulou and Plagnol, 2015), RIEMS (Scheuch *et al.*, 2015), Kaiju (Menzel *et al.*, 2016), VIP (Li *et al.*, 2016), DAMIAN (Alawi *et al.*, 2019), Jovian (Zwagemaker *et al.*, 2021) and CZ ID (Kalantar *et al.*, 2020) with an excellent comparative evaluation presented by de Vires et al. (de Vries *et al.*, 2021). Pipelines for virus discovery focus on sensitivity and, as suggested by several authors, benefit from integrating protein alignments and/or protein domain annotation in order to identify highly divergent emerging and novel viruses although at a cost of heavier computation (Zhao *et al.*, 2017; Alawi *et al.*, 2019; Kalantar *et al.*, 2020; Pérot *et al.*, 2022). Based on this notion most pipelines for virus diagnostics that include protein level alignments, or protein domain annotation, are also applicable to virus discovery. There is also a number of open-source pipelines that were specifically designed to support discovery of novel pathogens from mNGS data such as VirusSeeker (Zhao *et al.*, 2017), DAMIAN (Alawi *et al.*, 2019), CZ ID (Kalantar *et al.*, 2020) and Microseek (Pérot *et al.*, 2022). We have previously published mNGS pipelines, Lazypipe and Lazypipe2, primarily designed for virus discovery (Plyusnin *et al.*, 2023), which have been successfully applied to characterize viromes and identify novel viruses in clinical samples (Lehikoinen *et al.*, 2024), ticks (Kuivanen *et al.*, 2019; Zakham *et al.*, 2021; Alfaro-Alarcón *et al.*, 2022), mosquitos (Truong Nguyen et al., 2022; Korhonen et al., 2024), snakes (Alfaro-Alarcón *et al.*, 2022), farm animals (Virtanen *et al.*, 2021, 2023) and wildlife (Forbes *et al.*, 2019; Mull *et al.*, 2023; Erdin *et al.*, 2024; Alburkat *et al.*, 2024; Ogola *et al.*, 2024).

Here we describe Lazypipe3 with multiple new features including a novel concept of annotation strategies that can incorporate multiple nucleotide, protein and hidden Markov model (HMM) based searches into contig annotations, and that can be adjusted by the user depending on the data and analysis goals. Furthermore, we benchmarked several annotation strategies designed for virus diagnostics on synthetic data and annotation strategies designed for virus discovery on real data. With these benchmarks, we showed that a conceptually simple two-round strategy — where the first round uses a fast Minimap2 search (Li, 2018) against a virus-only database, followed by a second round using BLASTN (Altschul *et al.*, 1990) against a larger databases to filter false positives — can reduce execution time to just 2.6%–12% of that required by a one-round BLASTN search, without any loss in accuracy. We also showed that combining varying fast-to-accurate homology searches for analysis of real mosquito dataset reduced execution time from 5 to 20-fold compared to the BLASTN/BLASTP baseline. In real terms this translated into average time savings of 19 out of 20 hours for a collection of ~75Mbp sequence libraries analysed on a high-performance cluster. Examining results from one of these chained annotations we were able to recover and characterise multiple novel whole-genome viral sequences that were not recovered by Lazypipe1 and manual BLASTX analysis (Truong Nguyen *et al.*, 2022). Additionally, we demonstrate how combining HMMscan (Eddy, 2011) with other searches allowed us to identify and characterize a very divergent rhabdo-like genome from a horse fly sample. In Lazypipe3 we also implemented (and benchmarked) an improved background filtering and novel interactive reports for visualising coverage and placement of contigs relative to reference genomes.

2 MATERIAL AND METHODS

2.1 Improved background filtering

Lazypipe3 supports application of multiple background filters to remove reads and contigs from non-targeted host or contaminating species. Removing eukaryotic backgrounds allows fast annotation of background reads and a significantly smaller reference database for contig annotation. Filters are applied to both reads prior to assembling and to contigs after assembling. Multiple background filters can be applied in tandem when analysing pooled samples (e.g. arthropods) or samples contaminated by multiple sources. As many samples are likely to contain human nucleic acid background or contamination introduced during sample collection and library preparation, we recommend applying a *Homo sapiens* background filter to all datasets.

The pipeline is supported with several precompiled background filters covering approximately 20 species including human, mosquitos, ticks, bats, domestic and other mammals. These are available from a public server and can be downloaded on demand. Non-default background filters can be added by specifying a filter name and a fasta file in the configuration file.

Lazypipe3 background filtering is implemented using *bwa mem* alignment of reads to genomic and genomic mRNA sequences. Resulting sam alignments are then filtered by alignment score (AS,

default) and/or mapping quality (mapq) with Sambamba (Tarasov *et al.*, 2015) and SeqKit (Shen *et al.*, 2016).

2.2 Flexible and transparent annotation strategies

Lazypipe3 introduces the concept of "annotation strategy" and allows the user to select different strategies depending on the analysis goals and data types. Annotation strategy can be constructed on demand using several supported homology searches that can be arranged in chains to improve sensitivity or in rounds to improve precision and speed. The supported search engines include nucleotide (Minimap2, BLASTN), protein (SANSparallel, BLASTP/BLASTX and DIAMOND blastp/blastx (Somervuo and Holm, 2015; Altschul *et al.*, 1997; Buchfink *et al.*, 2015)) and HMMscan (Eddy, 2011) searches against precompiled or custom reference databases. Here, we tested several annotation strategies dividing them into 2-round strategies for virus identification and chained strategies for virus discovery.

In the two-round annotation the pipeline first runs a fast search against a virus-only database, e.g. Minimap2 on NCBI NT viruses or DIAMOND on UniRef viruses and then prunes any false positives in the second round with a BLAST search on a more extensive database. Choosing BLAST for the pruning stage seems appropriate as BLAST is arguably the most reliable tool for identification of viral sequences (Tangherlini *et al.*, 2016) and is integrated in many mNGS pipelines (Zhao *et al.*, 2017; Kalantar *et al.*, 2020; Vilsker *et al.*, 2019; Morfopoulou and Plagnol, 2015). The two-round strategy aims to speed up the pipeline by reducing the search space covered by the slower search in the 2nd round, while keeping false positive rate in check. For example, assuming 10% of viral contigs in the assembled data, 20% of viruses in the reference database, and an approximately 10-times faster search in the 1st round, should yield on the order of a 10-fold improvement in the execution time.

In the chained annotation the pipeline starts with a fast nucleotide search such as Minimap2. Contigs that have no hits in the first search are channelled to the second search that is expected to yield higher sensitivity either due to the search engine or the database. The process is then iterated to searches that have increasingly higher sensitivity, and in most cases lower precision and lower speed. By chaining several searches this strategy aims to improve the overall sensitivity.

All benchmarked strategies listed here are predefined in the pipeline configuration file and can be easily accessed from the command line.

2.2.1 Two-round strategies for virus diagnostics

This section describes two-round strategies designed for virus diagnostics, or more broadly virus identification, with focus on speed and low false positive rate.

Vi.NT and Vi.RefSeq: In the 1st round Minimap2 all contigs against NT or RefSeq viruses. In the 2nd round BLASTN pre-labelled viral contigs against NT or RefSeq archaea, bacteria and viruses (ABV).

Vi.NT.nohflt: This strategy is designed for samples that have no clear background (e.g. environment samples) and analysis that includes annotation of retroviruses. Background filtering is omitted. In the 1st round Minimap2 all contigs against the entire NT. In the 2nd round BLASTN pre-labelled viral contigs against the entire NT.

Vi.NT.fast and Vi.RefSeq.fast: In the 1st round Minimap2 all contigs against NT or RefSeq viruses. Omit the 2nd run. The goal is to provide very fast annotations at the expense of possibly higher false positive rates.

Vi.NT.extend: In the 1st round Minimap2 all contigs against NT viruses, then contigs with no hits with SANSparallel on UniProtKB. In the 2nd round BLASTN pre-labelled viral contigs against NT ABV and BLASTP against UniRef100 ABV. Chaining nucleotide and protein searches in the 1st round aims to accurately place both well-known and divergent sequences while keeping false positives in check with the more accurate BLAST searches in the 2nd round.

2.2.2 Chained strategies for virus discovery

This section describes chained annotation strategies designed for virus discovery with focus on reducing false negative rate while keeping the execution time feasible.

Vi.chain0: This is a baseline strategy using BLASTN and BLASTP, which are arguably the most reliable tools for identification of viral sequences (Tangherlini *et al.*, 2016). Applied searches include BLASTN against NT ABV, then BLASTP against UniRef100 ABV.

Vi.chain1: In this strategy Minimap2 is introduced in front of BLASTN. Applied searches include Minimap2 against NT ABV, then BLASTN against NT ABV, then BLASTP against UniRef100 ABV.

Vi.chain2: In this strategy BLASTN is omitted from the chain. Applied searches include Minimap2 against NT ABV, then BLASTP against UniRef100 ABV.

Vi.chain3: In this strategy BLASTP is substituted with a faster protein search, DIAMOND blastp (Buchfink *et al.*, 2015). Applied searches include Minimap2 against NT ABV, then Diamond blastp in very-sensitive mode against NT ABV. We also tested DIAMOND blastp in ultra-sensitive mode with no change in sensitivity and DIAMOND blastx in both modes with clearly lower sensitivity (data not shown).

Vi.chain4: In this strategy BLASTP is substituted with SANSparallel: Minimap2 against NT ABV, then SANSparallel against UniProtKB.

2.2.3 Supported reference databases

Lazypipe3 is supported with several precompiled and updated reference databases available from a public server (Table 1). These can be installed to local volumes with a simple command. The default nucleotide database is the NCBI NT, which was also the most common reference database employed by 8 out 13 pipelines examined by de Vries et al (de Vries *et al.*, 2021) and by CZ ID (Kalantar *et al.*,

2020). For the default protein database we chose UniRef100, similarly to GenomeDetective which uses UniRef90 (Vilsker *et al.*, 2019). Other options would be NCBI NR, e.g. DAMIAN (Alawi *et al.*, 2019) and CZ ID (Kalantar *et al.*, 2020), and RefSeq proteins, e.g. metaMix (Morfopoulou and Plagnol, 2015), but we choose UniRef100 as in our view it is both representative, with over 1.2 Gbp of viral sequences in 2024/11/27 release, and at the same time an order of magnitude smaller than NR.

Table 1. List of reference databases supported by Lazypipe3.

Database	Description
minimap.nt.abv, minimap.nt.vi	Fasta and accession-to-taxid files for annotations with Minimap2. NCBI NT archaea-bacteria-viruses and NT viruses.
blastn.nt.abv and blastn.vi	BLASTN index for NT archaea-bacteria-viruses and NT viruses.
minimap.refseq.abv and minimap.refseq.vi	Fasta and accession-to-taxid files for annotations with Minimap2. RefSeq archaea-bacteria-viruses and RefSeq viruses.
blastn.refseq.abv and blastn.refseq.vi	BLASTN index for RefSeq archaea-bacteria-viruses and RefSeq viruses
blastp.uniref100.abv and blastp.uniref100.vi	BLASTP index for UniRef100 archaea-bacteria-viruses and viruses
Diamond.uniref100.abv and diamond.uniref100.vi	Diamond index for UniRef100 archaea-bacteria-viruses and viruses
Hmmscan.pfam.vi	Viral HMM profiles from PFAM database (Mistry et al., 2021)
Hmmscan.rdrp	NeoRdRp2 database (Sakaguchi et al., 2024)

2.3 Other new features

Lazypipe3 adds *host.source* and *genome.composition* information from International Committee on Taxonomy of Viruses Virus Metadata Resource (ICTV VMR) to contig annotation tables whenever such information is available for the identified genus or family.

Other new features include support for annotations with DIAMOND blastp/blastx (Buchfink *et al.*, 2015) and HMMscan (Eddy, 2011). Diamond indices for archaea-bacteria-viruses and viruses in UniRef100, as well as viral HMM profiles from PFAM database (Mistry *et al.*, 2021) and NeoRdRp2 database (Sakaguchi *et al.*, 2024) were added to the list of supported databases.

Lazypipe3 can optionally output Integrative Genome Viewer reports (Robinson *et al.*, 2023) for identified viral contigs. These reports are produced by first binning contigs by the assigned species. For each species top hits from the annotation table are then retrieved and aligned to contigs with Minimap2. BAM files are then converted to an interactive IGV reports visualising genome coverage and variants. These reports are convenient for identifying contig overlaps in fragmented assemblies, visualising coverage of key genes, such as polymerases, as well as sequence similarity and variants.

2.4 Benchmarking

2.4.1 Benchmarking background filtering

We evaluated Lazypipe3 background filtering on two datasets. For the first one we used the human simulated metagenome from the MetaShot project (Fosso *et al.*, 2017). Reads were labelled as human reads or reads from other taxa based on sequence accessions included in the id field. Thresholds for alignment score (AS) and mapping quality (mapq) were optimized by stepwise iteration over a range of values. For the second dataset we used Illumina MiSeq sequence reads from mink faeces (unpublished). Read filtering was evaluated against read mapping to the two main findings, *Mamastrovirus* and *Circovirus,* as labelled with Lazypipe3 *Vi.NT.nohflt* two-round strategy with no background filtering.

2.4.2 Benchmarking strategies for virus diagnostics on simulated data

Two-round strategies were benchmarked against BLASTN baseline: annotation with a single round of BLASTN against NT archaea, bacteria and viruses. Comparison was done on the synthetic metagenome from the MetaShort project (Fosso *et al.*, 2017). This benchmark is a 20.5 M PE 2 × 150 Illumina library covering approximately 80 viral and 70 bacterial pathogens imbedded in a background of human reads (Fosso *et al.*, 2017). Background was filtered with human reference assembly (GCF_000001405.40), other pipeline options were kept at default. All runs were performed on a Linux/Unix CPU supercluster with 30 cores each running at 2.1 GHz.

2.4.3 Benchmarking virus discovery on real data

To test the idea of combining searches for virus discovery we run chained annotations on Illumina MiSeq dataset from 90 pools of *Ochlerotatus mosquitoes* (NCBI BioProject PRJNA852425) that was previously analysed with Lazypipe1 and downstream BLASTX (Truon Nguyen et al, 2022). Background was filtered with four chromosome level assemblies (*Aedes aegypti* acc. GCF_002204515.2, *Aedes albopictus* acc. GCF_035046485.1, *Culex pipiens pallens* acc. GCF_016801865.2, *Culex laticinctus* acc. CA_964187845.1) and the human reference assembly (GCF_000001405.40). Other pipeline options were kept at default.

Additionally, we analysed a previously unpublished Illumina MiSeq library sequenced from a Finnish horse fly (*Tabanidae sp.*) to test how HMMscan can augment annotation chains for virus discovery. Background was filtered with a chromosome level assembly from a related species (*Chrysops caecutiens*, acc. GCA_963971475.1) and the human reference assembly (GCF_000001405.40). For annotation we used Minimap2 against NT ABV, then BLASTN against NT ABV, then BLASTP against UniRef100 ABV, then HMMscan against viral HMM profiles collected from Pfam (Mistry *et al.*, 2021) and NeoRdRp v2.1 (Sakaguchi *et al.*, 2024). Other pipeline options were kept at default.

All runs were performed on a Linux/Unix CPU supercluster with 30 cores each running at 2.1 GHz.

2.5 Characterization of viral findings

Annotated array plots visualising ORF structure and closest homologs were constructed for selected viral findings as follows. Open reading frames (ORFs) were identified with orfipy (Singh and Wurtele, 2021), setting minimal ORF length to 300, excluding partial 5' ORFs, including partial 3' ORFs and using the standard genetic code. Smaller ORFs residing completely inside larger ORFs regardless of the strand were excluded. Extracted ORFs were annotated with the top hit from DIAMOND blastp v2.1.10.164 (Buchfink *et al.*, 2021) against UniRef100 viral segment (database retrieved 27/11/2024). ORFs were classified into groups using reciprocal best BLASTP links for pairs of contigs and manual editing. These groups were then labelled either by gene or protein name most commonly appearing in the assigned ORF annotations. ORF structure was visualized as an array plot using geneviewer (Velden, 2025).

Boostrapped maximum likelihood (ML) trees were constructed for selected viral findings as follows. For rhabdo-like sequences the reference set was collected by mapping BLASTX hits (BLASTX against clustered NCBI NR with *max-target-seq* set to 250 and *bitscore* cutoff to 400) to at most 100 nucleotide accessions and complementing these with all *Deltarhabdoviridae* isolates from International Committee of Taxonomy of Viruses database (ICTV VMR) (Lefkowitz *et al.*, 2018). From these protein sequences corresponding to the L-gene were extracted excluding short sequences (<1700 aa) and sequences with no BLASTP homology to contig L-genes, aligned with MAFFT v7.505 (Katoh and Standley, 2013) with options *--retree 2 –maxiter 100* and trimmed with trimAl 1.2rev59 (Capella-Gutiérrez *et al.*, 2009) with option *-automated1*. For negevirus-like contigs reference set was collected by mapping BLASTX hits to at most 100 nucleotide accessions. From these protein sequences corresponding to Methyltransferace_RNA_helicase_RdRp gene were extracted excluding <1000 nt sequences and sequences with no BLASTP homology, aligned and trimmed following the same protocol as for the rhabdo-like contigs. For totivirus-like contigs reference set was collected by mapping BLASTX hits to at most 100 nucleotide accessions, adding all *Artitivirus* isolates from ICTV VMR and adding all *Totiviridae* sequences published previously (Truong Nguyen *et al.*, 2022). From these protein sequences corresponding to both ORFs were extracted excluding <1000 aa ORF1 sequences and sequences with no BLASTP homology, aligned and trimmed as for the rhabdo-like contigs. Alignments for ORF1 and ORF2 were catenated. For flavi-like contigs reference set was collected by mapping BLASTX hits with bitscore cutoff at 200 to at most 200 nucleotide accessions complementing these with all *Orthoflavivirus* isolates from ICTV VMR as well as all *Flaviviridae* sequences priviously published for this dataset (Truong Nguyen *et al.*, 2022). Polyprotein sequence was extracted excluding <1000 nt sequences and sequences with no BLASTP homology, aligned and trimmed as for other groups.

ML trees were created for all protein alignments with IQTREE v2.1.3 (Minh *et al.*, 2020) with options *–ufboot 1000 --nmax 1000* and using ModelFinder (Kalyaanamoorthy *et al.*, 2017) to find the best-fitted model. Generated trees were further edited, re-rooted on midpoint and annotated in TreeViewer (Bianchini and Sánchez-Baracaldo, 2024).

3 RESULTS

3.1 Background filtering

Results for background filtering on simulated data are presented in Figure 1. Filtering by mapping quality (mapq) was less effective than filtering by alignment score (AS). For the default threshold we selected AS100. At this level 99.998% of human reads were filtered, while none of viral reads that were not retroviruses. The default filter removed 96.375% of human endogenous retroviruses (ERV), which is likely due to ERV sequences in the human reference genome. We note that ERVs can still be analysed by skipping background filtering and annotating against a database that include eukaryotic sequences.

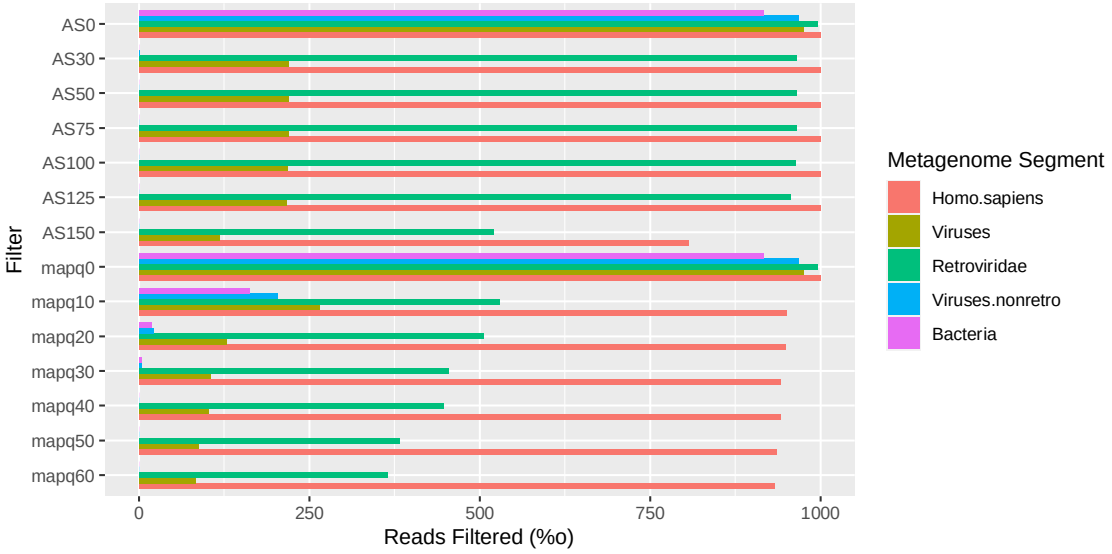
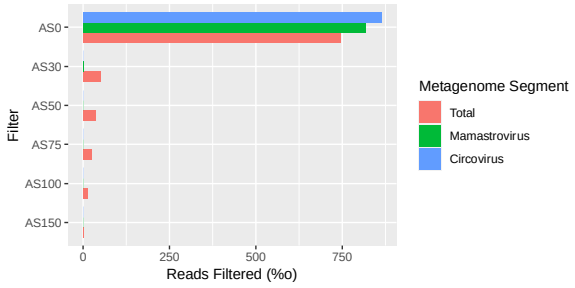


Figure 1 Benchmarking background filtering on simulated data from Fosso et al. (Fosso et al.*, 2017). Filters with AS30 to AS125 removed ~100.0% of human background and ~95.6-96.5% of retroviral reads, while reads of other viruses and and bacteria were not markedly affected. AS, alignment score threshold, mapq, mapping quality threshold.*

Results for background filtering on real data with a mink (*Noegale vison*) background are presented in Supp. Figure 1. At the AS100 1.4% of all reads were filtered but only tree *Mamastrovirus* reads and none of *Circovirus* reads.



Supp. Figure 2 Benchmarking background filtering on real data with mink background and two eukaryotic viruses. AS, alignment score threshold.

3.2 Virus detection from simulated data

We compared various metrics for taxon identification at species level (viruses) as well as execution time and maximum memory consumption (Maxrss) relative to BLASTN baseline for six different annotation strategies designed for virus diagnostics on simulated data from (Fosso *et al.*, 2017) (Figure 2 and Supp. Table 1). Differences in precision and recall, when annotating against NT, were small, but the runtime improved significantly with two-rounds. The lowest error rate was observed for annotation with no background filtering (*Vi.NT.nohflt*), which only reported two false positive (FP) and two false negative (FN) species. Two-round annotations against NT (*Vi.NT*) and the BLASTN (*BLASTN.NT.ABV*) baseline reported tree FPs and tree FNs. Two-round annotation against RefSeq (*Vi.RefSeq*) reported five FPs and seven FNs. Fast annotations with one round of Minimap2 reported four FPs and three FNs when using NT (*Vi.NT.fast*), and five FPs and seven FNs when using RefSeq (*Vi.RefSeq.fast*). Classification errors are discussed in further details in Supplementary Text1. Annotating in two-rounds (Minimap2 and BLASTN) reduced time to 12% (NT) and 2.6% (RefSeq) compared to the baseline. Annotating with just one round of Minimap2 further reduced time to 3.5% (NT) and 2.4% (RefSeq) compared to the baseline.

Read-level benchmarking confirmed that 2nd round BLAST search, did indeed reduce FP contig/read binning when searching against the NT (Supp. Table 2). However, this effect was absent with the smaller RefSeq database. Chaining nucleotide searches to protein searches in Vi.NT.extend resulted in the increased number of ambiguous mappings (Supp. Table 2) and would only be appropriate for virus discovery.

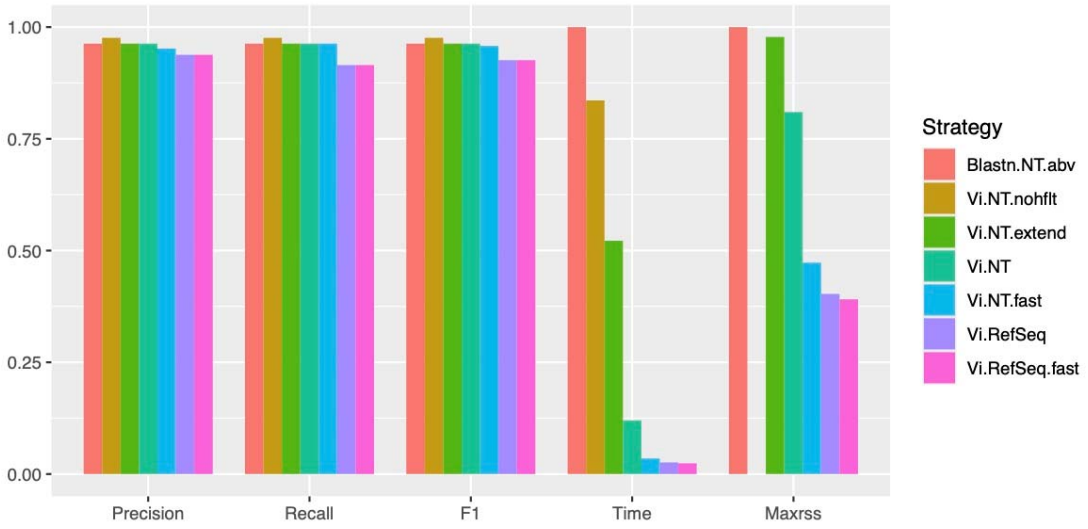


Figure 2 Benchmarking performance of annotation strategies for virus diagnostics on simulated data from Fosso et al. (Fosso et al., 2017). Precision, precision at species level (viruses), recall, recall at species level (viruses), F1, F1 metric at species level (viruses), Time, execution time, Maxrss (maximum memory consumption). All values are given relative to BLASTN.NT.abv baseline. For scale consistency Vi.NT.nohflt Maxrss value (431.6%) was excluded.

3.3 Virus discovery from real data

3.3.1 Benchmarking against BLAST baseline

We compared various metrics for taxon identification at species level (viruses) as well as execution time and maximum memory consumption (Maxrss) relative to the *Vi.chain0* baseline for four chained strategies (Supp. Table 3). Overall performance for the 90 mosquito samples is summarised in Figure 3. Adding Minimap2 before BLASTN and BLASTP resulted in ~1kbp of FN sequences and no FPs (*Vi.chain1*). Omitting baseline BLASTN from the chain resulted in ~11 kbp of FP and ~31 kbp of FN sequences (*Vi.chain2*). Omitting BLASTN and substituting BLASTP with Diamond blastp resulted in ~13 kbp of FP and ~47 kbp of FN sequences (*Vi.chain3*). Omitting BLASTN and substituting BLASTP with SANSparallel resulted in ~9 kbp of FP and ~549 kbp of FN sequences (*Vi.chain4*). Basepair level precision of all compared strategies was at least 99.7%. Basepair level recall was lowest for strategy incorporating SANSparallel (87.0%), and above 98.9% for all other strategies. The slowest chain (*Vi.chain1*) was ~5 times faster and the fastest chain (Vi.chain4) ~20 times faster than the BLAST baseline. The average runtime for mosquito samples (mean 74 Mbp, median 46 Mbp, range 0.2-338 Mbp) ranged from ~5 hours (Vi.chain1) to ~1 hour (Vi.chain3 and Vi.chain4) illustrating time savings of up to 19 out of 20 hours compared to the BLAST baseline.

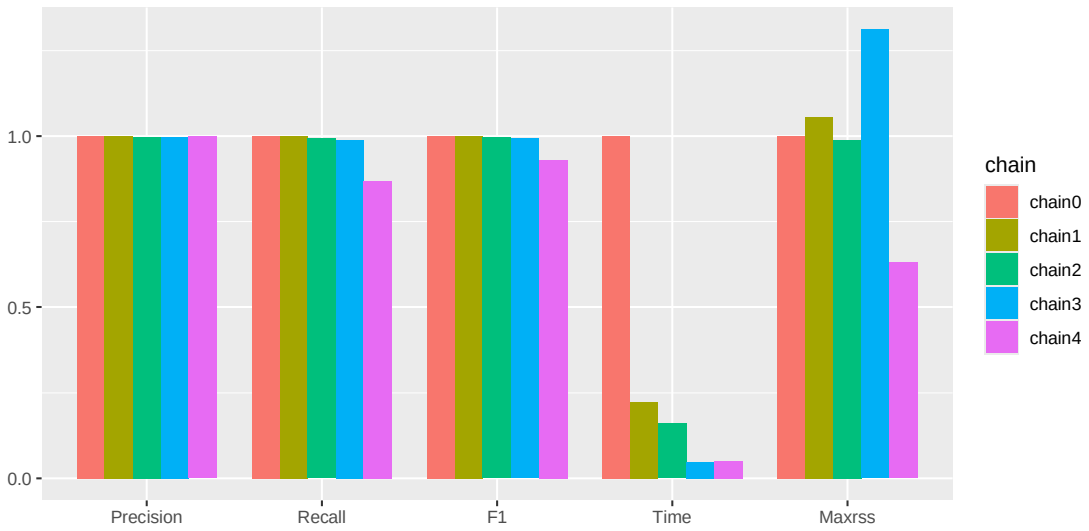


Figure 3 Overall performance of chained strategies on the mosquito dataset. Precision, recall, time and maxrss were calculated relative to Vi.chain0. Precision, precision of identified viral basepairs, Recall, recall of viral bps, F1, F1-metric, Time, relative execution time, Maxrss, relative maximum memory consumption.

3.3.2 Novel viruses

Here we characterise viral discoveries with Lazypipe *Vi.chain1* (Minimap2, BLASTN, BLATP) from the mosquito dataset and compare these to the previously published findings (Truong Nguyen *et al.*, 2022). We also examine a novel rhabdovirus discovered from the horse fly sample with *Vi.chain1* augmented with a HMMscan search.

For the mosquito dataset, *Vi.chain1* identified 1,242 viral contigs exceeding 1 kbp and covering a total of 3,559 kbp. From these Minimap2 identified 78.1 %, BLASTN 1.6% and BLASTP 20.3%. These contigs were assigned to 29 established viral families: *Aliusviridae, Alphaflexiviridae, Aspiviridae, Betaflexiviridae, Chrysoviridae, Chuviridae, Circoviridae, Coronaviridae, Endornaviridae, Flaviviridae, Iflaviridae, Metaviridae, Nodaviridae, Partitiviridae, Permutotetraviridae, Phasmaviridae, Phenuiviridae, Picornaviridae, Qinviridae, Rhabdoviridae, Sedoreoviridae, Solemoviridae, Spinareoviridae, Togaviridae, Tombusviridae, Totiviridae, Virgaviridae, Xinmoviridae* and *Yueviridae*.

Lazypipe3 with *Vi.chain1* identified all 25 viral families and genera reported by (Truong Nguyen *et al.*, 2022) and additional 7 families: *Alphaflexiviridae, Betaflexiviridae, Circoviridae, Coronaviridae, Metaviridae, Nodaviridae and Tombusviridae*. Improved coverage may be due to extending the list of contigs beyond the RdRp gene, a better pipeline and a better coverage of mosquito viruses due to database updates.

Keeping in mind application of chained annotations for virus discovery we next focus on >1 kbp viral contigs identified by BLASTP from the mosquito dataset and by HMMscan from the horse fly sample. These have no close homologs in NCBI NT and are likely candidates for novel viruses.

3.3.2.1 *Tentative Novel Rhabdoviruses*

Altogether chained BLASTP identified seven contigs mapped to *Rhabdoviridae* that were named here according to municipalities they were discovered from (Figure 4). Comparing these sequences against rhabdoviruses previously published for this dataset (Truong Nguyen *et al.*, 2022) confirmed very low nucleotide level homology (qcov <9%) for all sequences except *Ohlsdorf virus isolate FIN/L-2018/02-1*, which had 70.0% identity at 95% qcov to *Ohlsdorf virus* (ON955254.1).

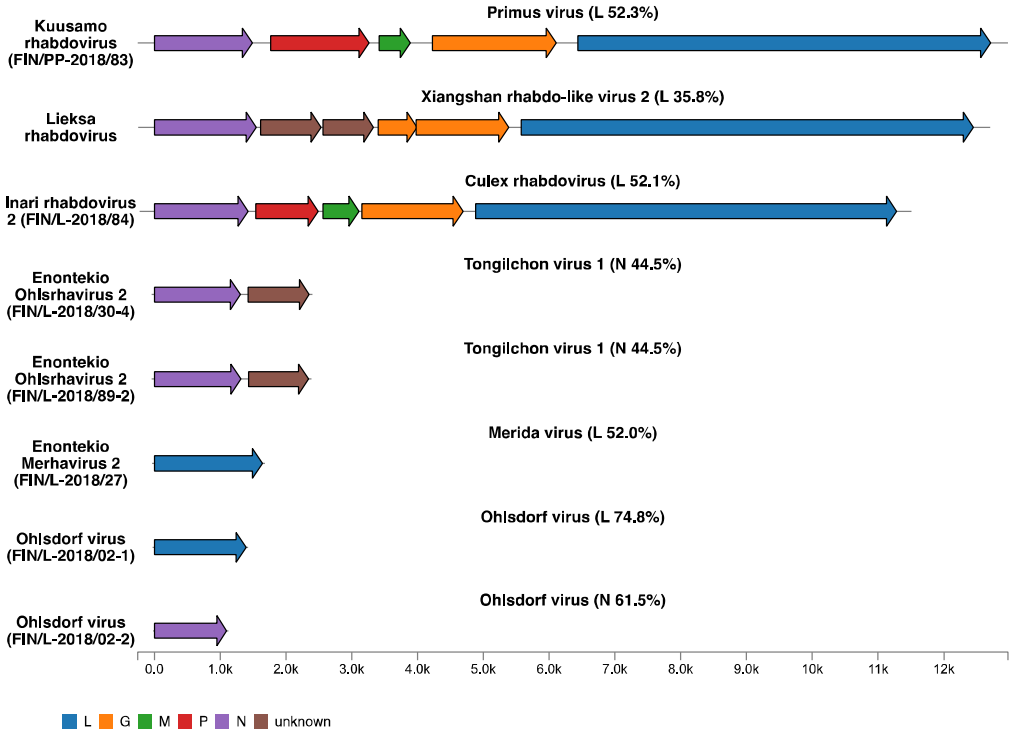


Figure 4 ORF structure of novel Rhabdo-like sequences from mosquito samples and the horse-fly sample. Labels on the left, proposed virus names and variants, labels at the center, closest homologs in UniRef with identity values in parenthesis.

The two largest sequences were named *Kuusamo rhabdovirus* and *Inari rhabdovirus 2* both with five ORFs corresponding to N, P, M, G and L genes in a typical rhabdovirus genome (Figure 4). In *Kuusamo rhabdovirus* ORF1 had 36.5% aa identity to nucleoprotein, ORF2 29.8% identity to phosphoprotein, ORF4 35.9% identity to glycoprotein and ORF5 52.3% identity to RdRp in *Primus virus* (UniRef accessions A0A6H0C864_9VIRU, A0A6H0C6Q8_9VIRU, A0A6H0C7L9_9VIRU and A0A6H0CA79_9VIRU), while ORF3 had no homologs in UniRef. In *Inari rhabdovirus 2* ORF1 had 35.8% identity to *Merida virus* nucleoprotein (A0A894KPN5_9RHAB), ORF2 had no homologs, ORF3 had 26.7% identity to matrix protein in unclassified rhabdovirus (A0A140DDD9_9RHAB), ORF4 had 49.6% identity to *Merida virus* glycoprotein (A0A5Q0TW57_9RHAB) and ORF5 52.1% identity to *Culex rhabdovirus* RdRp (A0A7L7QQ11_9RHAB). Notably, the *Inari rhabdovirus 2* had no significant

nucleotide level similarity to *Inari rhabdovirus* (acc. ON955143.1) previously identified from the same sample.

In *Enontekio Ohlsrhavirus 2* (isolates FIN/L-2018/30-4 and FIN/L-2018/89-2) ORF1 had 44.5% identity to nucleoprotein in *Tongilchon virus 1* (UniRef A0A0S2RRE4_9RHAB), while ORF2 had no homologs in UniRef. These isolates shared ~98% nucleotide identity. *Enontekio merhavirus 2* had a single ORF with 52.0% identity to the L-gene in *Merhavirus merida* (A0A894KD42_9RHAB). *Ohlsdorf virus isolate FIN/L-2018/02-1* had a single ORF with 74.8% identity to the L-gene in *Ohlsdorf virus* (A0A291I327_9RHAB). *Ohlsdorf virus isolate FIN/L-2018/02-2* had a single ORF with 61.5% identity to the N-gene in *Ohlsdorf virus* (A0A291I346_9RHAB). Since these were from the same sample these likely represent a single viral genome.

HMMscan identified a single 12.7 kbp rhabdo-like sequence from the horse fly sample with ORF structure following genomic organisation of a rhabdovirus (Figure 4). We tentatively name this new virus as *Lieksa rhabdovirus* after the municipality where the horse-fly was sampled. In this contig ORF1 had 24.7% aa identity to nucleoprotein (A0A0B5KRF8_9RHAB), ORF4 24.4% aa identity to glycoprotein (A0A7D7F3C3_9RHAB), ORF5 21.9% aa identity to another glycoprotein (A0A8K1YQP7_9RHAB) and ORF6 35.8% aa identity to Replicase (A0A8K1YQP8_9RHAB) of *Water Strider Virus 5*, *Coleopteran rhabdo-related virus OKIAV29* and *Xiangshan rhabdo-like virus 2.* ORF2 and ORF3 had no homologs in UniRef but likely represent P and M genes.

Bootstrapped maximum likelihood (ML) trees were constructed from the L-gene from *Kuusamo rhabdovirus, Inari rhabdovirus 2* and *Lieksa rhabdovirus*. In a phylogenetic tree (Figure 5) *Inari rhabdovirus 2* is placed close to other *Merhaviruses*, including previously published *Enontekio merhavirus*, although the host species is different. This may represent a tentative novel *Merhavirus* or a new variant. *Kuusamo rhabdovirus* is placed close to *Primrhavirus* BLASTX homologs. Again this may be a tentative novel virus or a variant. The *Lieksa rhabdovirus* from the horse fly appears distinct but related to *Strangrhaviruses* in mosquitoes and may represent a novel *Stangrhavirus* (subfamily *Deltarhabdovirinae*) hosted by horse flies (Figure 5).

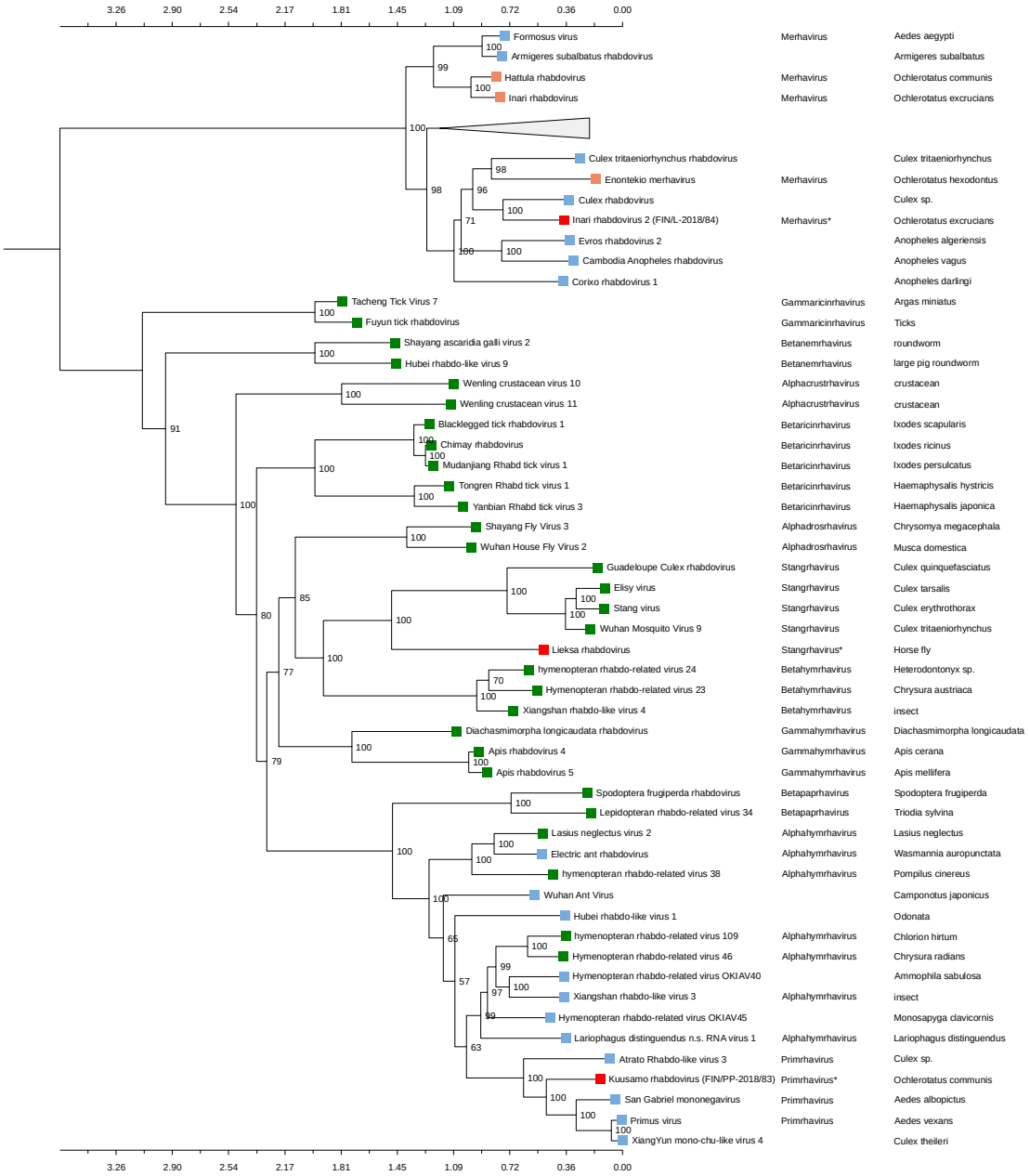


*Figure 5 Phylogeny of novel Rhabdovirus-like sequences identified in mosquito and horse-fly samples. Tip colours: novel sequences in red, BLASTX homologs previously published for the mosquito data in light red, other BLASTX homologs in blue, ICTV Deltarhabdovirinae isolates in green. The first column on the right denotes virus genus, the second column host species. *Proposed name and genus.*

3.3.2.2 Tentative Novel Negeviruses

Chained BLASTP identified four whole-genome negevirus-like sequences ranging from ~11.1-11.4 kbp and 21 shorter sequences. Figure 6 illustrates ORF structure, closest homologs in UniRef and proposed virus names including variants for the 20 largest sequences. In the whole-genome sequences (*Kuusamo, Hanko* and *Kustavi negeviruses*) ORF1 had 41.0-41.8% aa identity to RNA helicase in *Bactrocera dorsalis negev-like virus* (A0AAE9KSF2_9VIRU), ORF2 28.4-29.1% identity to

Virion glycoprotein in *Halyomorpha halys negev-like virus 1* and *Aedes camptorhynchus negev-like virus* (A0AAE9LBF5_9VIRU and A0A1Z2RT28_9VIRU), ORF3 had 44.3 % identity to uncharacterized protein in *Halyomorpha halys negev-like virus 1* (A0AAE9LC09_9VIRU) and ORF4 32.3-34.2% identity to uncharacterized protein in *Culex tritaeniorhynchus Negev-like virus* (A0A6F8PIU4_9VIRU). *Hanko negevirus* isolates *FIN/U-2018/92* and *FIN/VS-2018/100* shared over 99% identity, while *Kuusamo negevirus* and *Kustavi negevirus* were distinct suggesting three different genomes. Comparing against previously published Negeviruses (Truong Nguyen *et al.*, 2022) confirmed very low nucleotide level homology (qcov <=3%) for all novel negevirus-like sequences.

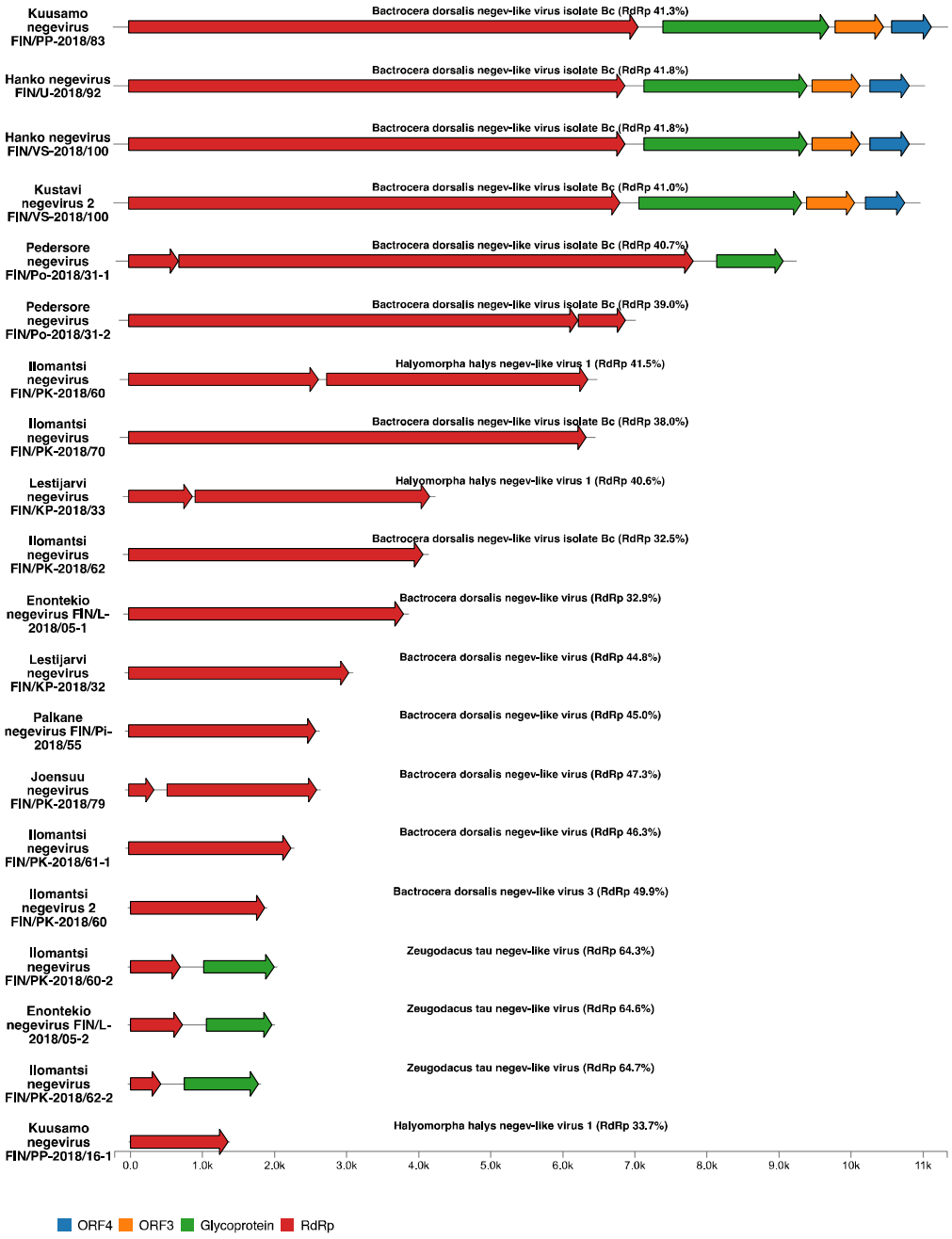


Figure 6 ORF structure of novel Negeviruses from mosquito samples. Labels on the left, proposed virus names and variants, labels at the center, closest homologs in UniRef with identity values in parenthesis.

Bootstrapped ML trees were constructed with the RdRp-gene for Negevirus sequences with RdRp coverage exceeding 1 kbp (Figure 7). In *Ilomantsi negevirus FIN/PK-2018/60* and *Lestijarvi negevirus FIN/KP-2018/33* a signifiant portion of RdRp appeared split in two ORFs and these were treated as a single RdRp aa sequence. Phylogenetic tree showed that all novel sequences formed a monophyletic

group distinct from BLASTX homologs (Figure 7). There were two separate clades, with majority of viruses in the first claded sequenced from *O. caspius* and those in the second claded from *O. communis* (Figure 7).

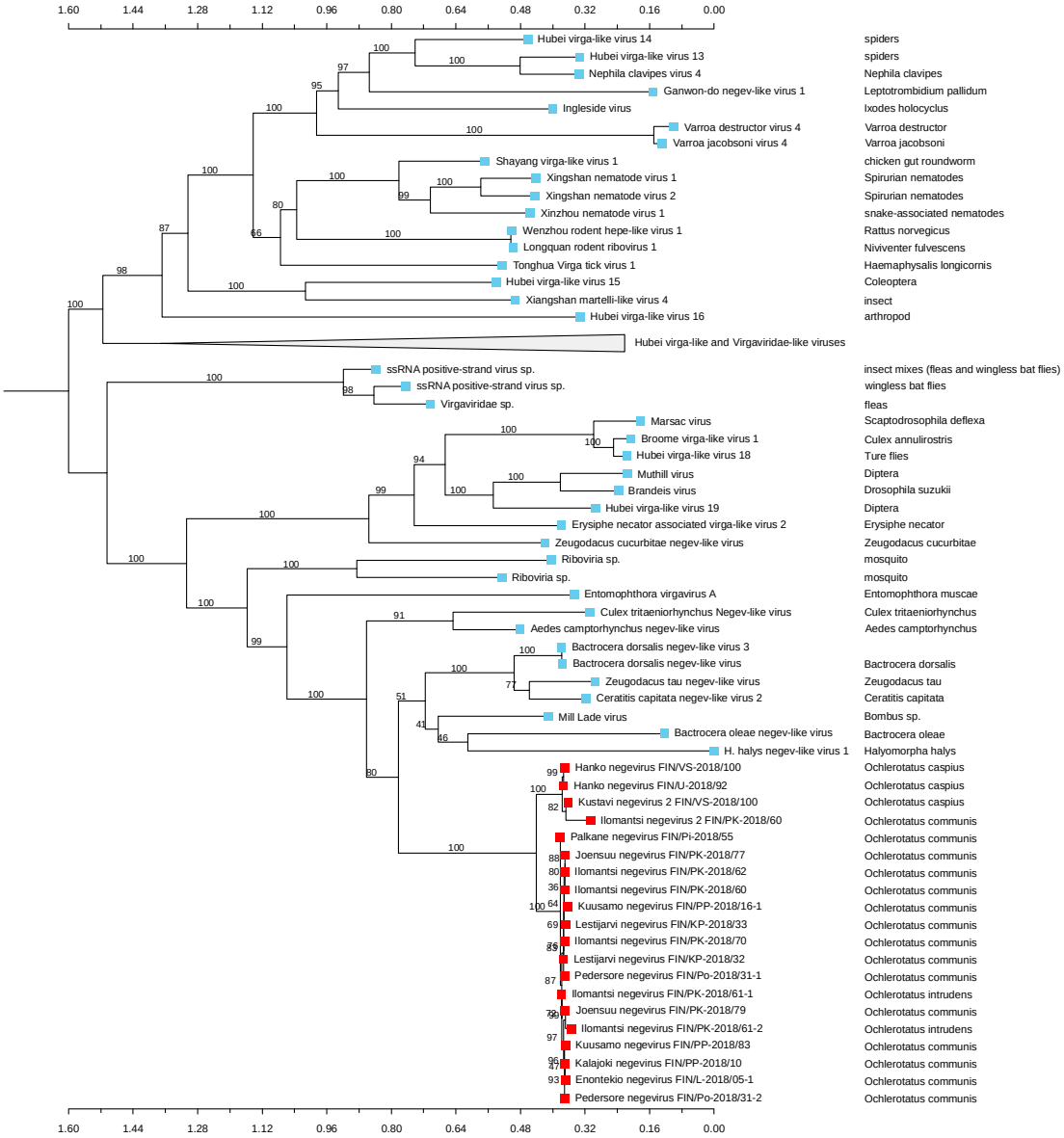


Figure 7 Phylogeny for novel negevirus-like sequences from mosquito samples. Tip colours: novel sequences in red, BLASTX homologs in blue. Host species in the right column.

3.3.2.3 Tentative Novel Totivirus

Chained BLASTP identified five whole-genome totivirus-like sequences ranging from ~7.5 to 7.7 kbp and 16 shorter sequences (Figure 8). In the whole-genome sequences ORF1 had 38.1-39.4% aa identity to capsid protein in *Armigeres subalbatus virus SaX06-AK20* (D2CSN1_9VIRU) and ORF2 37.1% identity to RdRp in *Penaeid shrimp infectious myonecrosis virus* (A0A7D0H216_9VIRU). Five

whole-genome sequences clustered together at over 97.5% identity suggesting these represented highly similar genomes in different samples.

Comparison against priviously published *Totivirus* sequences (Truong Nguyen *et al.*, 2022) showed very low nucleotide level homology (qcov<14%) for all but two sequences. These were *Lapland toti-like virus FIN/L-2018/23* with 70.0% identity at 99% qcov to previously published *Hattula totivirus 3* (OP019879.1) and *Hattula toti-like virus FIN/KH-2018/29* with 77.9% identity at 96% qcov to *Hattula totivirus 1* (OP019869.1).

Bootstrapped ML trees were constructed with catenated ORF1 and ORF2 aa sequences for the five complete genome contigs. Most of shorter contigs had low aa homology to the complete sequences and, thus, could not be included in the tree. In the resulting phylogenetic tree we see contigs forming a monophyletic group of closely related sequences quite distinct from BLASTX homologs and totiviruses previously identified from this data (Figure 9). These sequences likely represent variants of single novel Totivirus. The closes BLASTX homolog is *Crogonang virus 29* (acc. OR270267.1) and the closes ICTV exemplar is the *Artitivus sani* (*Armigeres subalbatus virus*, acc. EU715328.1).

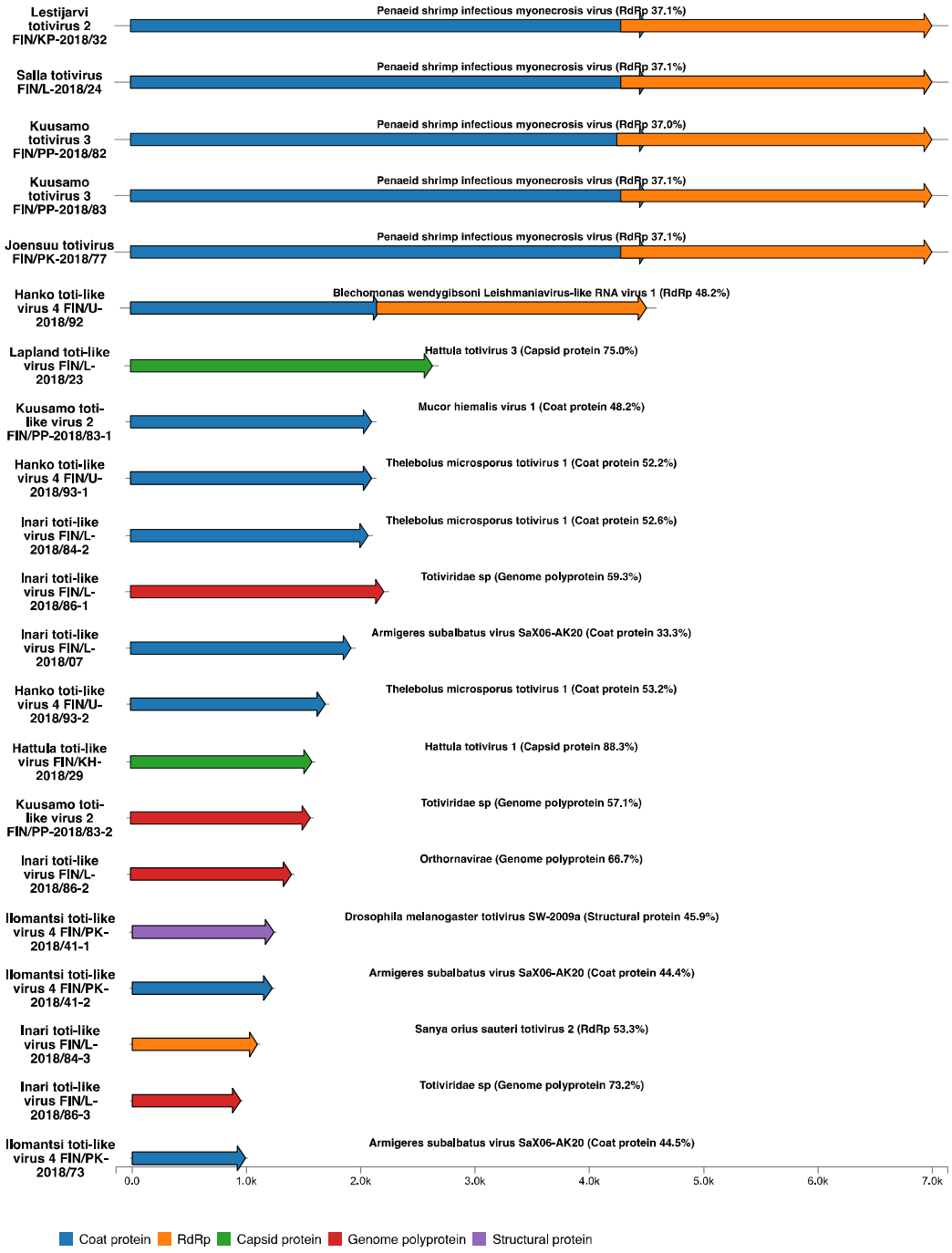


Figure 8 ORF structure of novel Totivirus-like sequences from mosquito samples. Labels on the left, proposed virus names and isolates, labels at the center, closest homologs in UniRef with identity values in parenthesis.

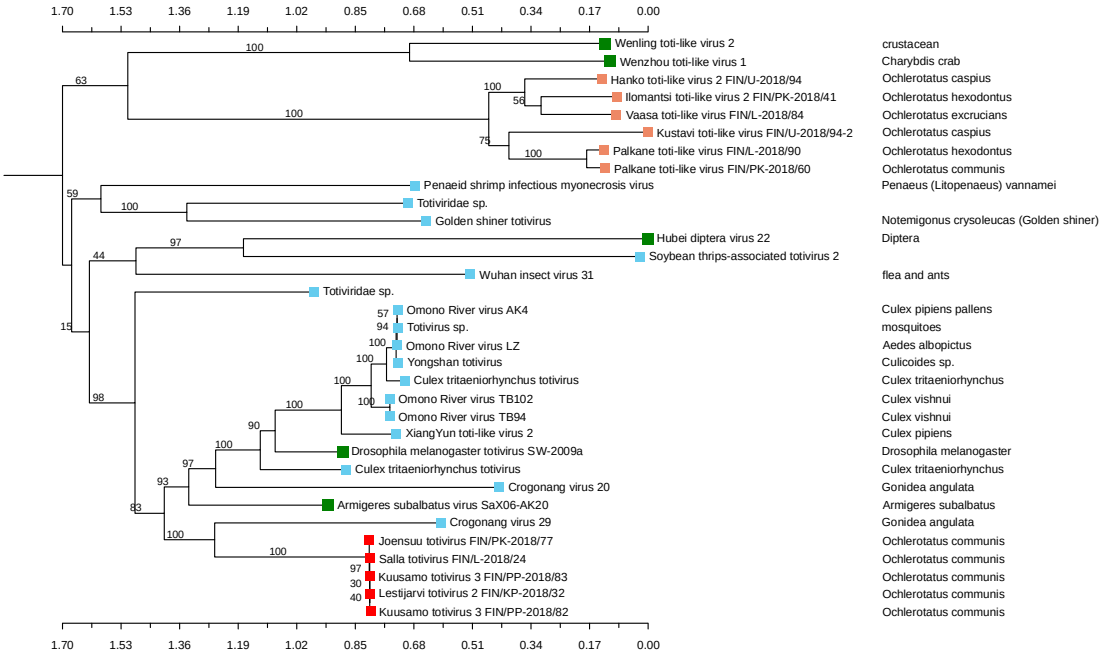


Figure 9 Phylogeny of novel Totivirus-like sequences from moquito samples. Tip colors: novel sequences in red, Totiviridae sequences previously published for this data in light red, BLASTX homologs in blue, ICTV Artivirus exemplars in green. Host species in denoted in the right column.

3.3.2.4 Tentative Novel Flaviviruses

Chained BLASP identified six contigs mapped to *Flaviviridae*. Most of these sequences had no pairwise similarity, except for *Palkane flavivirus FIN/Pi-2018/54* and *Palkane flavivirus FIN/Pi-2018/55* that were 96.3% nt identical. All contigs had ORFs spanning most of contig length with flavivirus polyprotein homologs in UniRef (Figure 10). Protein level identities to the closest UniRef homolog (polyprotein in *Culex tritaeniorhynchus flavi-like virus*, *Hymenopteran flavi-related virus* and *unclassified Flavivirus*) varied between 29.3% and 76.0% (Figure 10). Comparison against previously published *Flaviviridae* sequences showed 72.0% nt identity between *Hameenlinna flavivirus FIN/KH-2018/38-2* and *Hanko virus* (acc. ON949929.1), while other sequences were non-homologous.

Bootstrapped ML tree was constructed for novel *Flaviviridae* sequences as described in the methods section. One of the sequences appeared to be a new variant (*FIN/KH-2018/38-2)* of previously published *Hameenlinna flavivirus variant FIN/KH-2018/38.* Another flavi-like sequence, named *Palkane flavi-like virus variant FIN/Pi-2018/54*, was placed distant from other novel and previous characterized *Flaviviridae* and as an outgroup to *Placeda and Culex tritaeniorhynchus flavi-like viruses*. Four novel sequences, including two variants of *Lestijarvi flavivirus* hosted by *Ochlerotatus diantaeus* and two variants of *Palkane flavivirus* hosted by *O. communis*, formed a monophyletic group very distant from previously published *Flaviviridae* and BLASTX homologs (Figure 10).

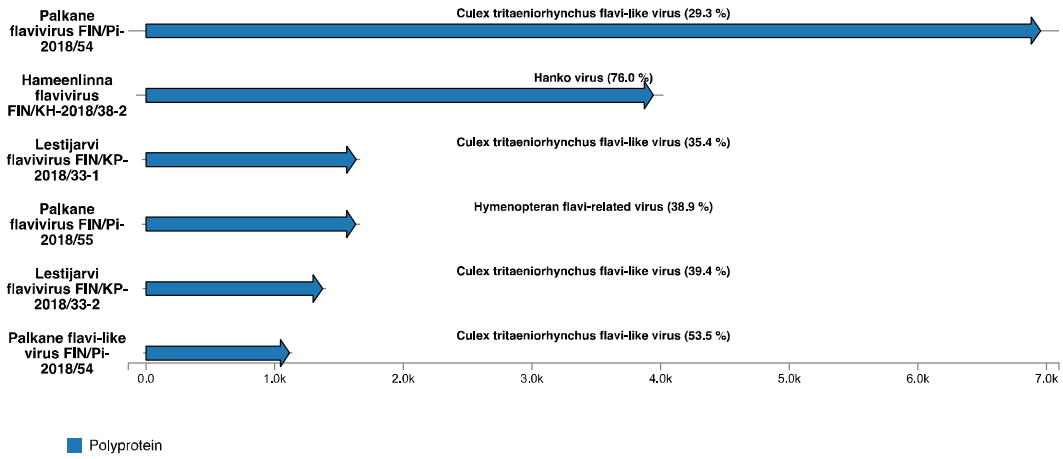


Figure 10 ORF structure of novel Flavi-like sequences from mosquito samples. Labels on the left, proposed virus names and isolates, labels at the center, closest homologs in UniRef with identity values in parenthesis.

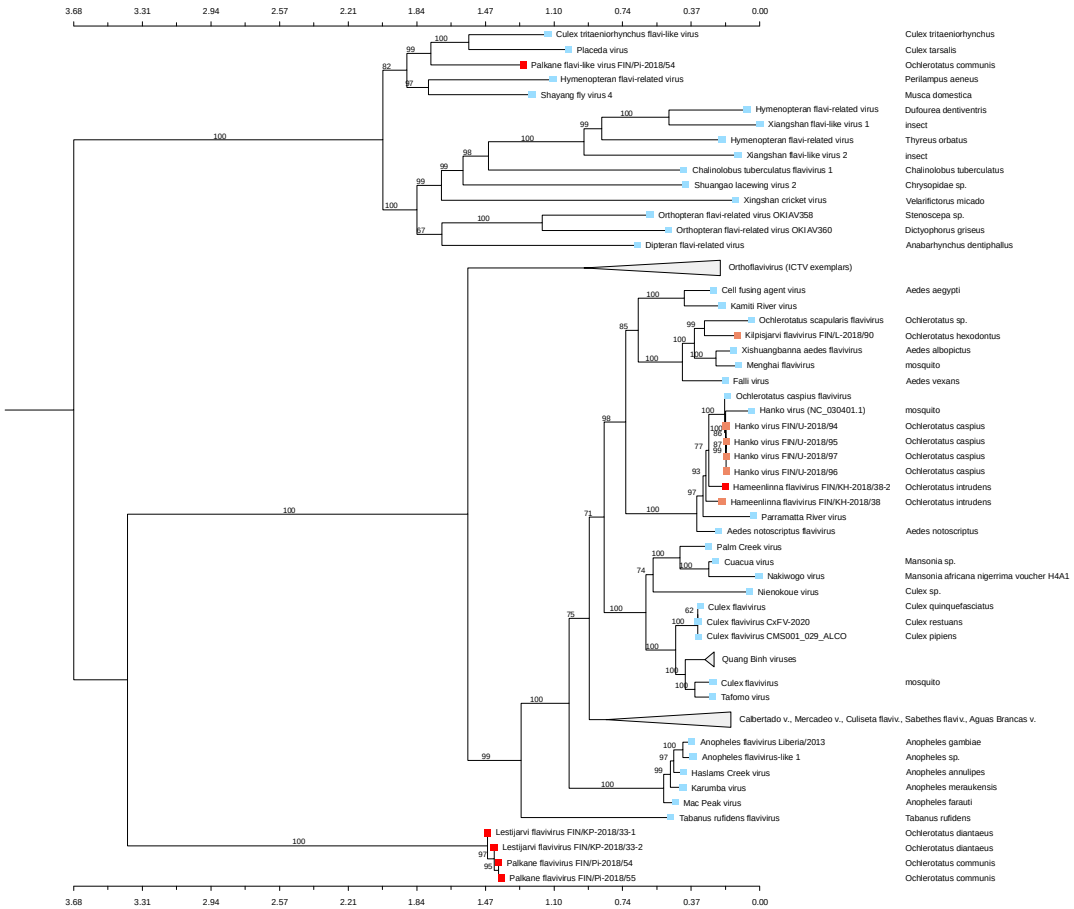


Figure 11 Phylogenetic tree of novel Flavi-like sequences from mosquito samples. Tip colors: novel sequences in red, homologs previously published for this data in light red, BLASTX homologs in blue. Host species in denoted in the right column.

DISCUSSION

Lazypipe3 implements several unique features that make it a flexible, transparent and powerful tool for virus discovery and identification from mNGS data.

Lazypipe3 supports filtering of one to several backgrounds. Filters are made transparent by specifying database accessions and other metadata for each filter in the configuration file. Filtering at the read stage allows for a much faster pipeline with up to 10-fold reduction of execution time for samples with a heavy background as was illustrated on the simulated benchmark (Figure 2). This is due to both reducing the size of the data and using a smaller reference database for the annotation stage. Our benchmarking on real and synthetic data demonstrates high accuracy of the implemented background filtering. A collection of background filters for common target species in vertebrate metagenomics is provided on demand and custom background filters can be added by listing new entries to the configuration file. These options address multiple user cases, such as metagenomics of clinical samples from human patients, animal patients, contamination with human nucleic acids during sample preparation, contamination by other than targeted animal species (e.g. samples from animal farms), analysis of vector pools such as mosquitoes and ticks with multiple species intermixed, and analysis of wild-life specimens when species-level identification is unavailable. Multiple filters can be applied in a tandem. Although many mNGS pipelines support filtering of human background (Alawi *et al.*, 2019; Gleizes *et al.*, 2020; Zwagemaker *et al.*, 2021; Kalantar *et al.*, 2020; Morfopoulou and Plagnol, 2015), and many allow for custom filters (Alawi *et al.*, 2019; Gleizes *et al.*, 2020; Zwagemaker *et al.*, 2021), Lazypipe3 also supports application of multiple filters in tandem and offers precompiled filters, which is generally not supported by other pipelines.

Lazypipe3 as its predecessors performs *de novo* assembling. The reasoning here is that assembling into longer sequences is likely to increase sensitivity and specificity of homology searches. This idea is shared by pipelines that classify assemblies instead or in addition to classifying raw reads (Alawi *et al.*, 2019; Zwagemaker *et al.*, 2021; Morfopoulou and Plagnol, 2015; Scheuch *et al.*, 2015; Kalantar *et al.*, 2020).

Lazypipe3 supports flexible annotations that incorporate both nucleotide and protein alignments, which is an uncommon feature. For example, in the recent comparison by de Vries et al (de Vries *et al.*, 2021) 8 out of 13 pipelines/classifiers were based on nucleotide alignments, two on protein alignments (Buchfink *et al.*, 2015; Morfopoulou and Plagnol, 2015) and only one public pipeline (DAMIAN, Alawi *et al.*, 2019), from this survey and CZ ID (Kalantar *et al.*, 2020), not covered by the survey, used both. Lazypipe also supports annotations with Hidden Markov Models (HMMs). From the 13 pipelines examined by de Vries (de Vries *et al.*, 2021) only a single pipeline, DAMIAN (Alawi *et al.*, 2019), incorporated HMMs into the annotation algorithm. Finally, unlike most mNGS pipelines Lazypipe3 allows to flexibly adjust and change the annotation pipeline within the main pipeline to match different data sets and user cases. Annotation procedure can be modified by chaining different homology searches against different databases to increase sensitivity, and by arranging searches in

two rounds to increase specificity. This also allows to iteratively adjust the annotation procedure depending on intermediate results. Users can experiment with eight different homology searches (Minimap2, SANSparallel, BLASTN, BLASTP, BLASTX, Diamond blastp/blastx and HMMscan). This level of flexibility is generally not available in mNGS pipelines. To summarise, with Lazypipe3 we have brough forward the concept of annotation strategies, which offers a flexible and transparent way to define search engines and reference databases applied in annotation. We showed how different annotation strategies can be reused and benchmarked to improve pipeline precision and time performance on synthetic and real data.

Different two-round annotation strategies for virus diagnostics were compared on as synthetic data against a baseline with a single BLASTN. As we showed annotations against NT performed slightly better than annotations against RefSeq likely due to better coverage of the sequence space. We tested the idea of speeding up virus identification by scanning for viral candidate sequences in the first round using a small virus-only database and then filtering false positives in the second round with BLASTN against a larger database that covers also archaea and bacteria. As expected, this two-round search reduced execution time to 12% on NT and to 2.6% on RefSeq compared to the one-round baseline without any reduction in species-level precision/recall. Annotating with just one round of Minimap2 against NCBI NT and RefSeq viruses further reduced execution time to 3.5% and 2.4%. Although there was almost no detectable increase in FPs for the applied synthetic data, according to our experience annotating only on the viral database is likely to report misclassified bacterial sequences when analysing real datasets. The two-round *Vi.NT* reported just three false negative (FN) and three false positive (FP) viral species. FNs included two retroviruses, likely due to filtering of these sequences with the human background, and *Mopeia Lassa virus reassortant 29* which was misclassified to a nearly identical *Mammarenavirus lassaense.* Filtering of retroviral reads could be addressed in future releases by masking retroviral sequences in background filters. From FPs one was likely due to accession-to-taxonomy update while the remaining FPs were misclassification of closely related viral taxa.

Lazypipe3 two-round strategy is not unique its attempt to perform faster alignments prior to the more accurate but slower alignments. The idea was inspired by VirusSeeker pipeline that starts by fast identification of candidate viral sequences against virus-only nucleotide and protein databases and then removes false positives with alignments against NCBI NT and NR databases (Zhao *et al.*, 2017). Also, in CZ ID reads are first mapped with fast nucleotide and protein aligners prior to aligning contigs with BLAST against targeted databases (Kalantar *et al.*, 2020). In GenomeDetective viral reads are identified with fast DIAMOND search on UniRef90 prior to assembling and running BLASTN/BLASTX on RefSeq. Lazypipe complements these approaches by offering flexibility to choose different search engines and databases for both scanning and refinement stages.

Different chained strategies for virus discovery were compared on a real dataset of 90 mosquito samples against the BLASTN/BLASTP baseline. Introducing fast Minimap2 search prior to BLAST searches reduced execution time ~5X and substituting BLAST searches with faster

Minimap2/Diamond or Minimap2/SANSparallel searches reduced execution time ~20X. Baselevel precision and recall of virus discovery chains 1 (Minimap2/BLASTN/BLASTP), 2 (Minimap2/BLASTP) and 3 (Minimap2/DIAMOND) were very high (precision above 99.7% and recall above 98.9%), while chain 4 (Minimap2/SANSparallel) had lower recall (87.0%). Distant viral findings with chain 1 on the mosquito data were further characterised and demonstrated that multiple novel viral genomes were recovered that were not identified in the previous analysis of this dataset, including whole genomes from *Rhabdoviridae, Negeviridae*, *Totiviridae* and *Flaviviridae*. Furthermore, by adding the more sensitive HMMscan to *Vi.chain1* we were able to identify a novel Rhabdo-like virus from a horse fly sample, illustrating how HMMscan can be used with Lazypipe3 to discover very distant novel viruses. These results demonstrate the usability of our concept of chained annotation strategies for exploration of real datasets and annotation options.

Future perspectives

To further enhance the accessibility and usability of Lazypipe3, we are planning to develop a graphical user interface (GUI). This will allow users without a bioinformatics background or command-line experience to run virus discovery analyses effortlessly. Designed with biologists, clinicians, and field researchers in mind, the GUI will enable them to perform complex analyses with minimal technical expertise. Additionally, we plan to extend Lazypipe3 for use in field settings, particularly for integration with the Oxford Nanopore MinION platform. This version will be optimized for both offline and online modes, enabling real-time analysis of mNGS data in remote locations.

4 Author Contributions

Conceptualization of the project: I.W., R.K., T.S. and O.V.; Conceptualization of the pipeline: I.W., R.K. and T.S.; Implementation, updates and code distribution: I.W.; Reference Database Curation: I.W.; Benchmarking: I.W.; Testing and Reporting: I.W., R.K. and T.S.; Writing — Original Draft Preparation: I.W.; Writing — Review & Editing: O.V., R.K. and T.S.; Supervision of the study: R.K., T.S. and O.V.; Funding Acquisition: R.K. and O.V.

5 Acknowledgements

The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

6 Conflicts of Interest

The authors have no conflict of interest to declare.

7 Data Availability Statement

Lazypipe3 user manual and source code are available at https://bitbucket.org/plyusnin/lazypipe/ and https://www.helsinki.fi/en/projects/lazypipe. The produced virus sequences are publicly available in the NCBI GenBank database (for accession codes see Supp. Table 4).

8 Funding

This research was funded by the Academy of Finland (grant number 339510), PREPARE-TID (EU grant number 101137132) DURABLE (EU HERA), VEO – European Union's Horizon 2020 (grant number 874735), the Jane and Aatos Erkko Foundation, and Helsinki University Hospital Funds. Funding bodies were not directly involved in designing or implementing research described in this manuscript.

9 Supplementary Materials

Supplementary Text 1

Supplementary Figure 1

Supplementary Tables 1, 2 and 3

10 References

Alawi,M. *et al.* (2019) DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples. *Sci. Rep.*, **9**.

Alburkat,H. *et al.* (2024) Evolution and genetic characterization of Seoul virus in wild rats Rattus norvegicus from an urban park in Lyon, France 2020–2022. *PLoS Negl. Trop. Dis.*, **18**, e0012142.

Alfaro-Alarcón,A. *et al.* (2022) Boid Inclusion Body Disease Is Also a Disease of Wild Boa Constrictors. *Microbiol. Spectr.*, **10**, e01705-22.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bassi,C. *et al.* (2022) Novel Virus Identification through Metagenomics: A Systematic Review. *Life*, **12**, 2048.

Bianchini,G. and Sánchez Baracaldo,P. (2024) TreeViewer: Flexible, modular software to visualise and manipulate phylogenetic trees. *Ecol. Evol.*, **14**, e10873.

Buchfink,B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.

Buchfink,B. *et al.* (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*, **18**, 366–368.

Capella-Gutiérrez,S. *et al.* (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLOS Comput. Biol.*, **7**, e1002195.

Erdin,M. *et al.* (2024) Detection of divergent Orthohantavirus tulaense provides insight into wide host range and viral evolutionary patterns. *Npj Viruses*, **2**, 1–13.

Forbes,K.M. *et al.* (2019) Bombali Virus in Mops condylurus Bat, Kenya. *Emerg. Infect. Dis.*, **25**, 955–957.

Fosso,B. *et al.* (2017) MetaShot: an accurate workflow for taxon classification of host-associated microbiome from shotgun metagenomic data. *Bioinformatics*, **33**, 1730–1732.

Gleizes,A. *et al.* (2020) Virosaurus A Reference to Explore and Capture Virus Genetic Diversity. *Viruses*, **12**, 1248.

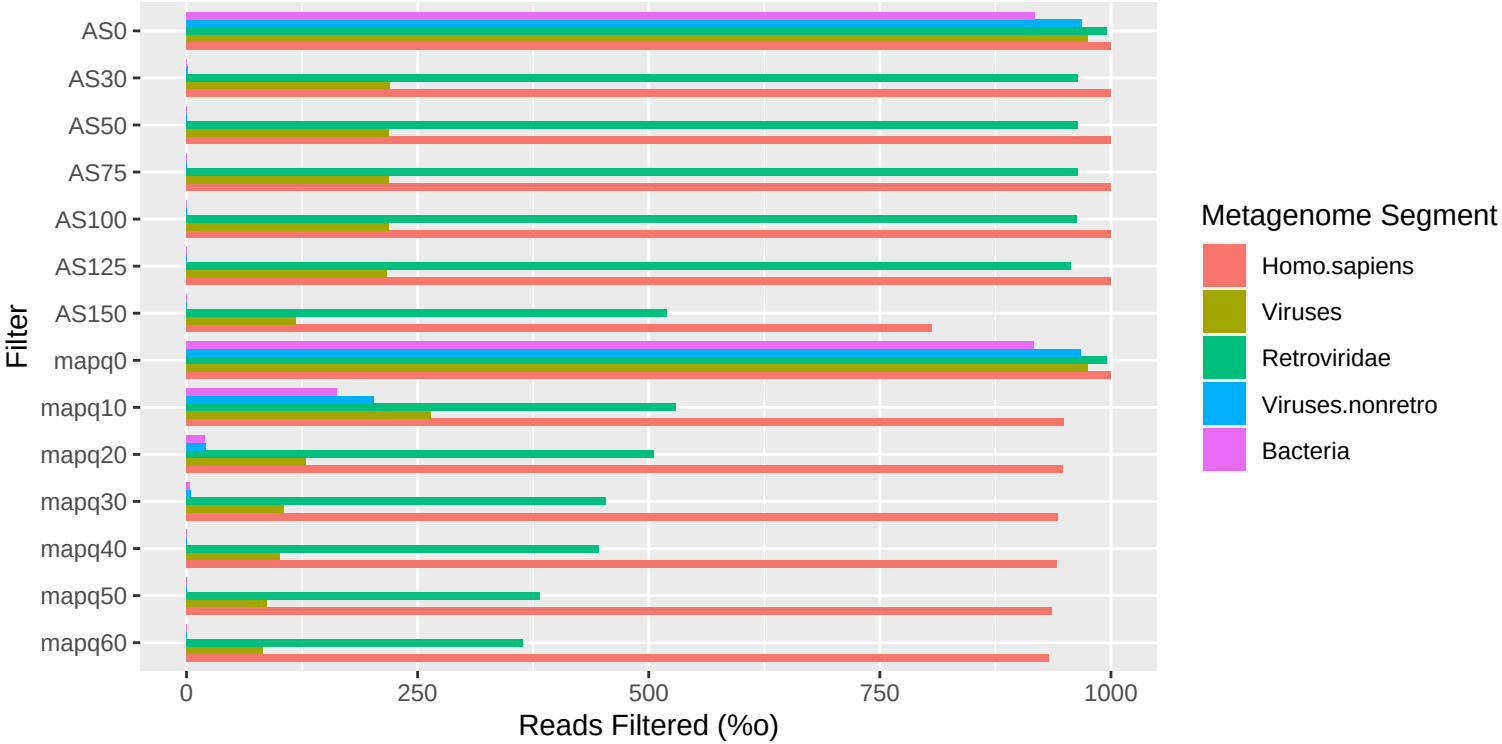
Ibañez-Lligoña,M. *et al.* (2023) Bioinformatic Tools for NGS-Based Metagenomics to Improve the Clinical Diagnosis of Emerging, Re-Emerging and New Viruses. *Viruses*, **15**, 587.

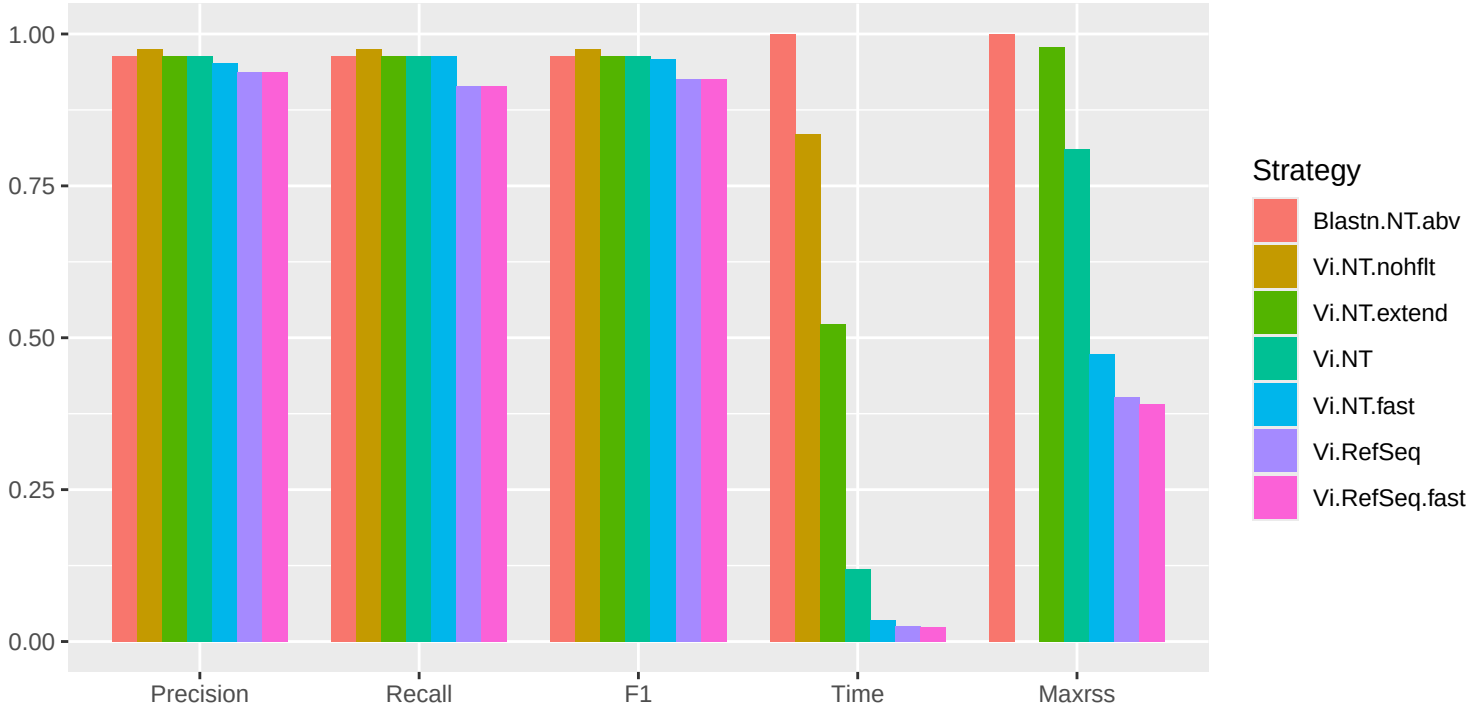
Kalantar,K.L. *et al.* (2020) IDseq—An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *GigaScience*, **9**, giaa111.

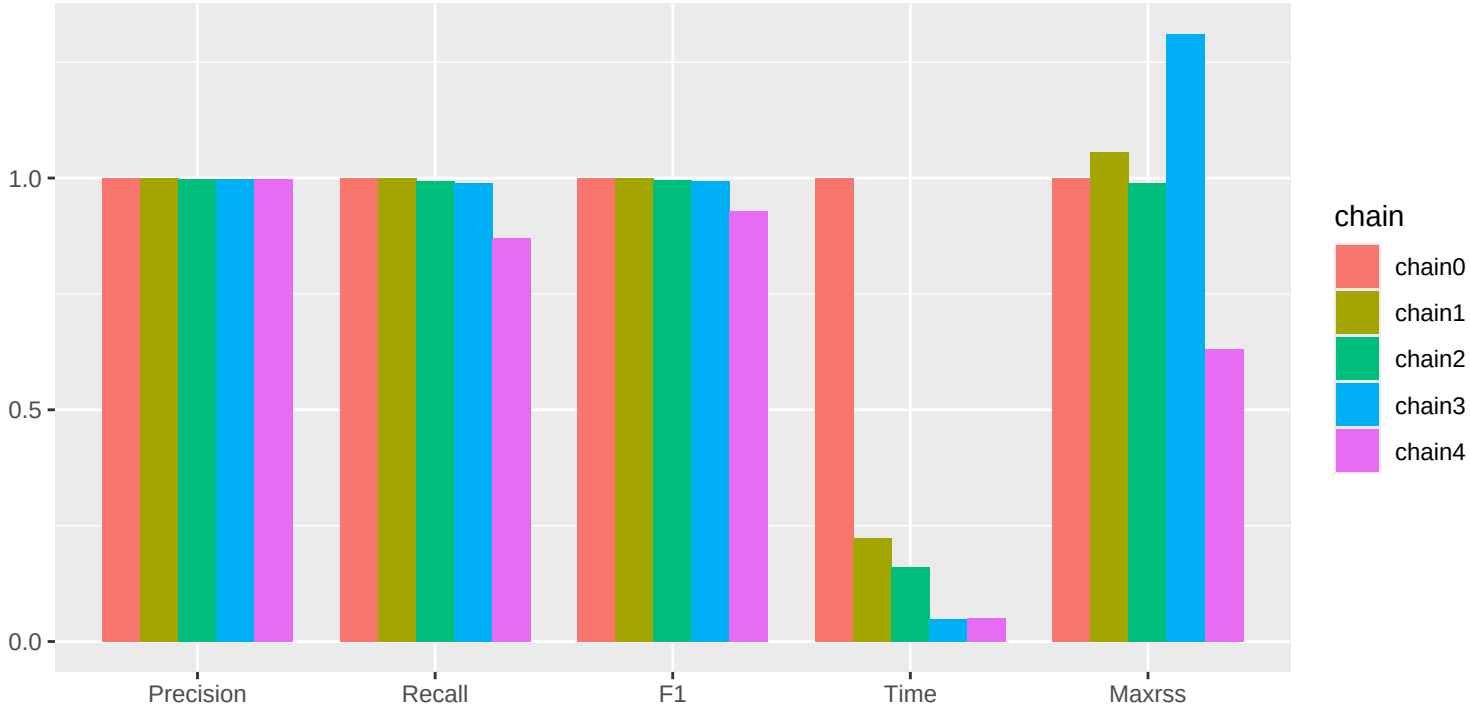
Kalyaanamoorthy,S. *et al.* (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*, **14**, 587–589.

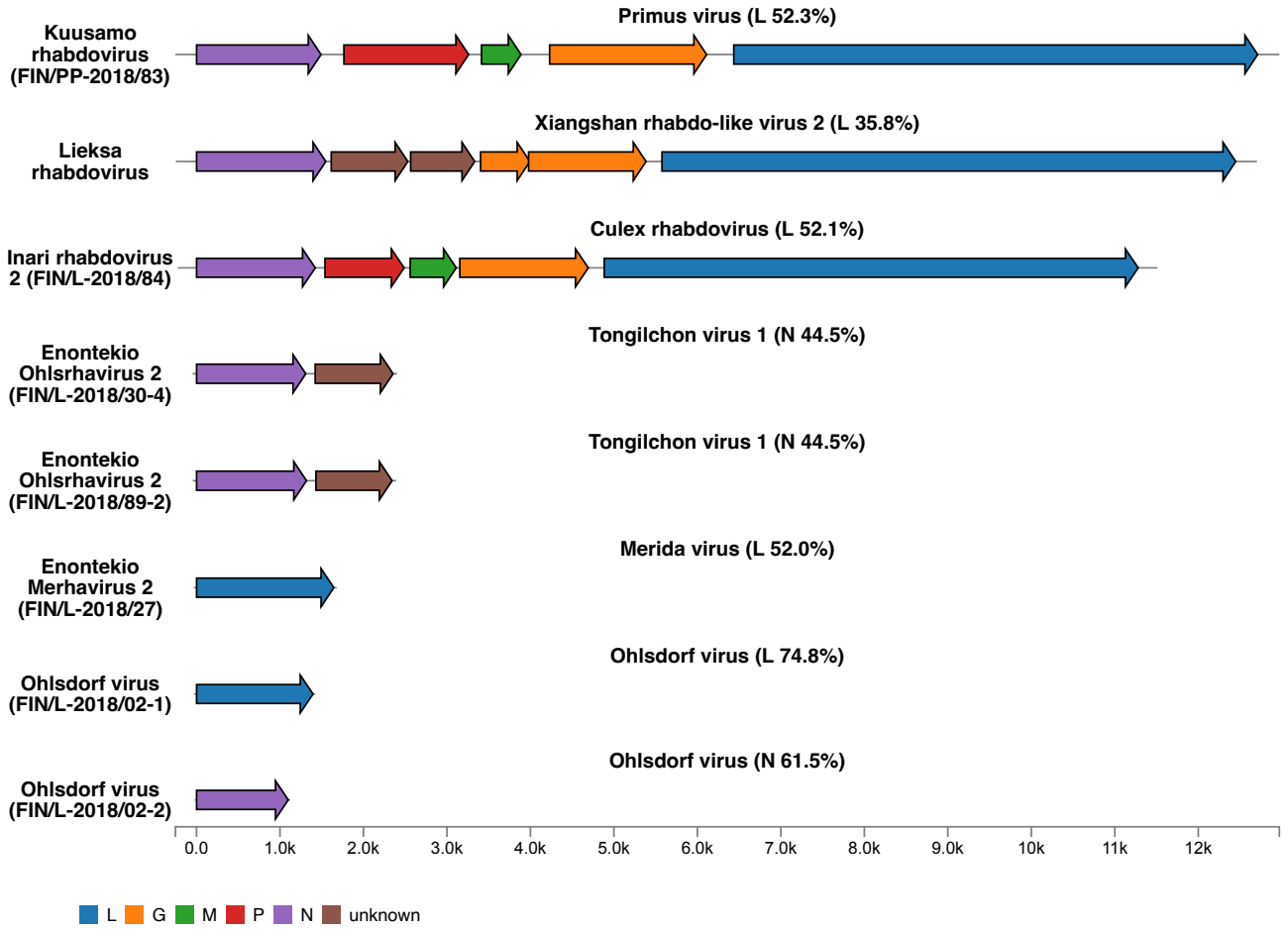
Katoh,K. and Standley,D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.*, **30**, 772–780.

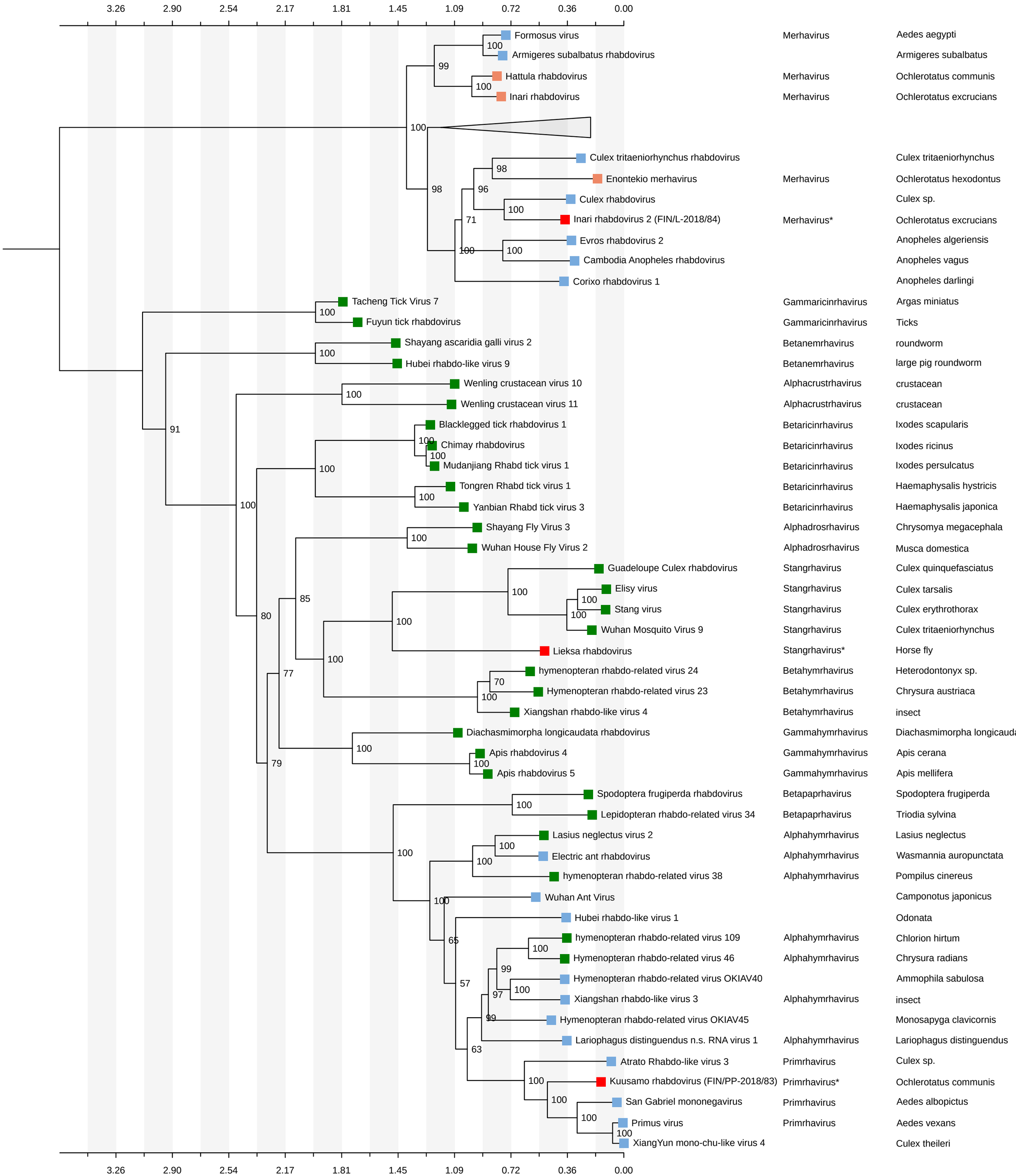
Korhonen,E.M. *et al.* (2024) Complete coding sequence of an Aedes flavivirus strain isolated from Aedes albopictus collected in Northern Italy. *Microbiol. Resour. Announc.*, **13**, e00146-24.
Kuivanen,S. *et al.* (2019) Detection of novel tick-borne pathogen, Alongshan virus, in Ixodes ricinus ticks, south-eastern Finland, 2019. *Eurosurveillance*, **24**.
Lefkowitz,E.J. *et al.* (2018) Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.*, **46**, D708–D717.
Lehikoinen,J. *et al.* (2024) Epstein-Barr Virus in the Cerebrospinal Fluid and Blood Compartments of Patients With Multiple Sclerosis and Controls. *Neurol. Neuroimmunol. Neuroinflammation*, **11**, e200226.
Li,Y. *et al.* (2016) VIP: an integrated pipeline for metagenomics of virus identification and discovery. *Sci. Rep.*, **6**, 23774.
Menzel,P. *et al.* (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.*, **7**, 11257.
Minh,B.Q. *et al.* (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.*, **37**, 1530–1534.
Mistry,J. *et al.* (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
Morfopoulou,S. and Plagnol,V. (2015) Bayesian mixture analysis for metagenomic community profiling. *Bioinformatics*, **31**, 2930–2938.
Mull,N. *et al.* (2023) Novel Ozark Orthohantavirus in Hispid Cotton Rats (Sigmodon hispidus), Arkansas, USA. *Emerg. Infect. Dis.*, **29**, 2576–2578.
Ogola,J.G. *et al.* (2024) Detection and genetic characterization of alphacoronaviruses in co-roosting bat species, southeastern Kenya. 2024.12.23.24319537.
Pérot,P. *et al.* (2022) Microseek: A Protein-Based Metagenomic Pipeline for Virus Diagnostic and Discovery. *Viruses*, **14**, 1990.
Plyusnin,I. *et al.* (2023) Enhanced Viral Metagenomics with Lazypipe 2. *Viruses*, **15**, 431.
Robinson,J.T. *et al.* (2023) igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics*, **39**, btac830.
Sakaguchi,S. *et al.* (2024) NeoRdRp2 with improved seed data, annotations, and scoring. *Front. Virol.*, **4**.
Scheuch,M. *et al.* (2015) RIEMS: A software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. *BMC Bioinformatics*, **16**.
Shen,W. *et al.* (2016) SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE*, **11**, e0163962.
Singh,U. and Wurtele,E.S. (2021) orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics*, **37**, 3019–3020.
Somervuo,P. and Holm,L. (2015) SANSparallel: interactive homology search against Uniprot. *Nucleic Acids Res.*, **43**, W24–W29.
Tangherlini,M. *et al.* (2016) Assessing viral taxonomic composition in benthic marine ecosystems: reliability and efficiency of different bioinformatic tools for viral metagenomic analyses. *Sci. Rep.*, **6**, 28428.
Tarasov,A. *et al.* (2015) Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, **31**, 2032–2034.
Truong Nguyen,P.T. *et al.* (2022) Characterisation of the RNA Virome of Nine Ochlerotatus Species in Finland. *Viruses*, **14**, 1489.
Velden,N. van der (2025) geneviewer: Gene Cluster Visualizations.
Vilsker,M. *et al.* (2019) Genome Detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics*, **35**, 871–873.
Virtanen,J. *et al.* (2021) Diversity and transmission of Aleutian mink disease virus in feral and farmed American mink and native mustelids. *Virus Evol.*, **7**, veab075.
Virtanen,J. *et al.* (2023) Equine dermatitis outbreak associated with parapoxvirus. 2023.09.01.555671.
de Vries,J.J.C. *et al.* (2021) Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. *J. Clin. Virol.*, **141**, 104908.
Zakham,F. *et al.* (2021) Viral RNA Metagenomics of Hyalomma Ticks Collected from Dromedary Camels in Makkah Province, Saudi Arabia. *Viruses*, **13**, 1396.
Zhao,G. *et al.* (2017) VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology*, **503**, 21–30.
Zwagemaker,F. *et al.* (2021) DennisSchmitz/Jovian: 1.2.07.

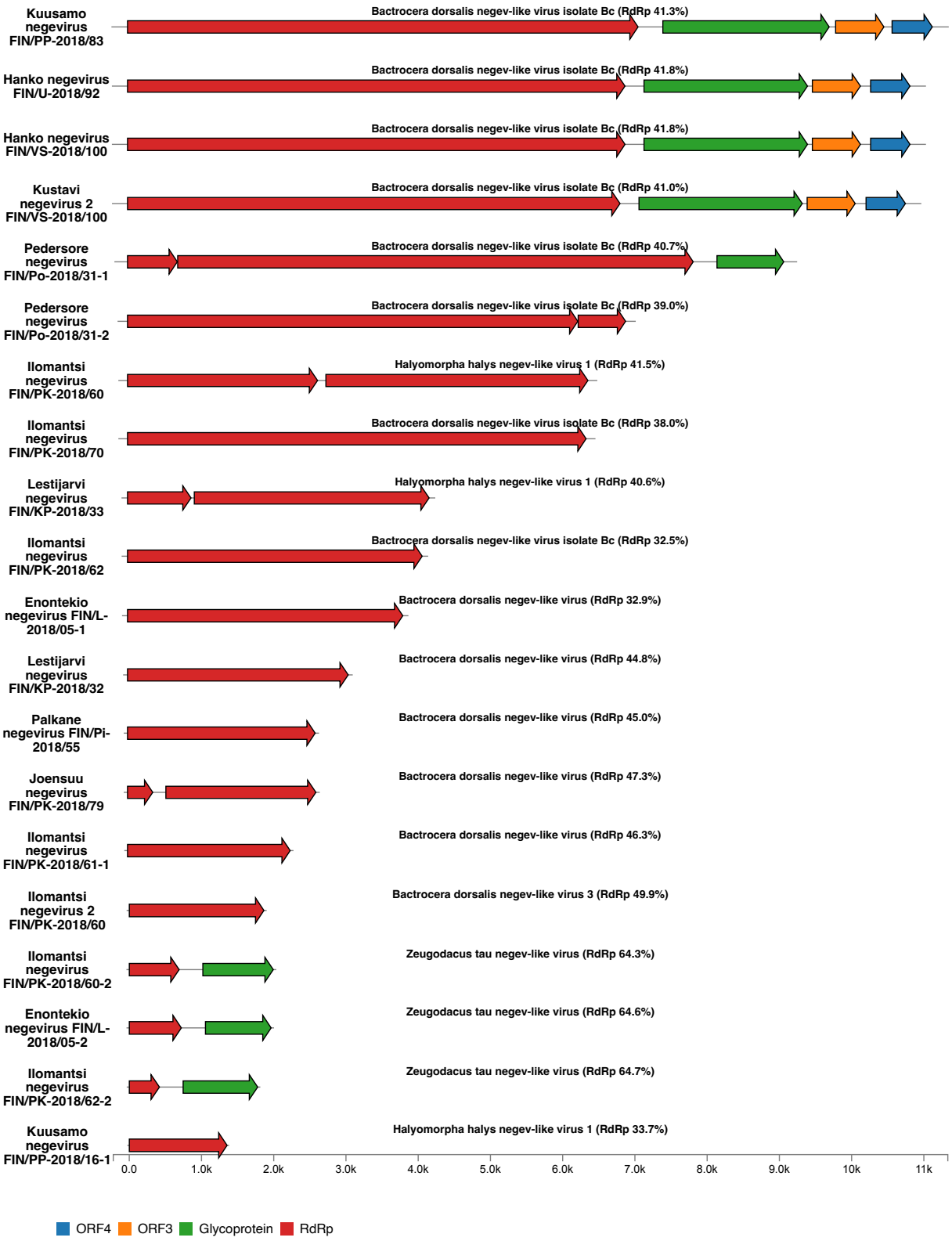

Filter
AS0
AS30
AS50
AS75
AS100
AS125
AS150
mapq0
mapq10
mapq20
mapq30
mapq40
mapq50
mapq60
0
250
500
750
1000
Reads Filtered (%o)
Metagenome Segment
Homo.sapiens
Viruses
Retroviridae
Viruses.nonretro
Bacteria

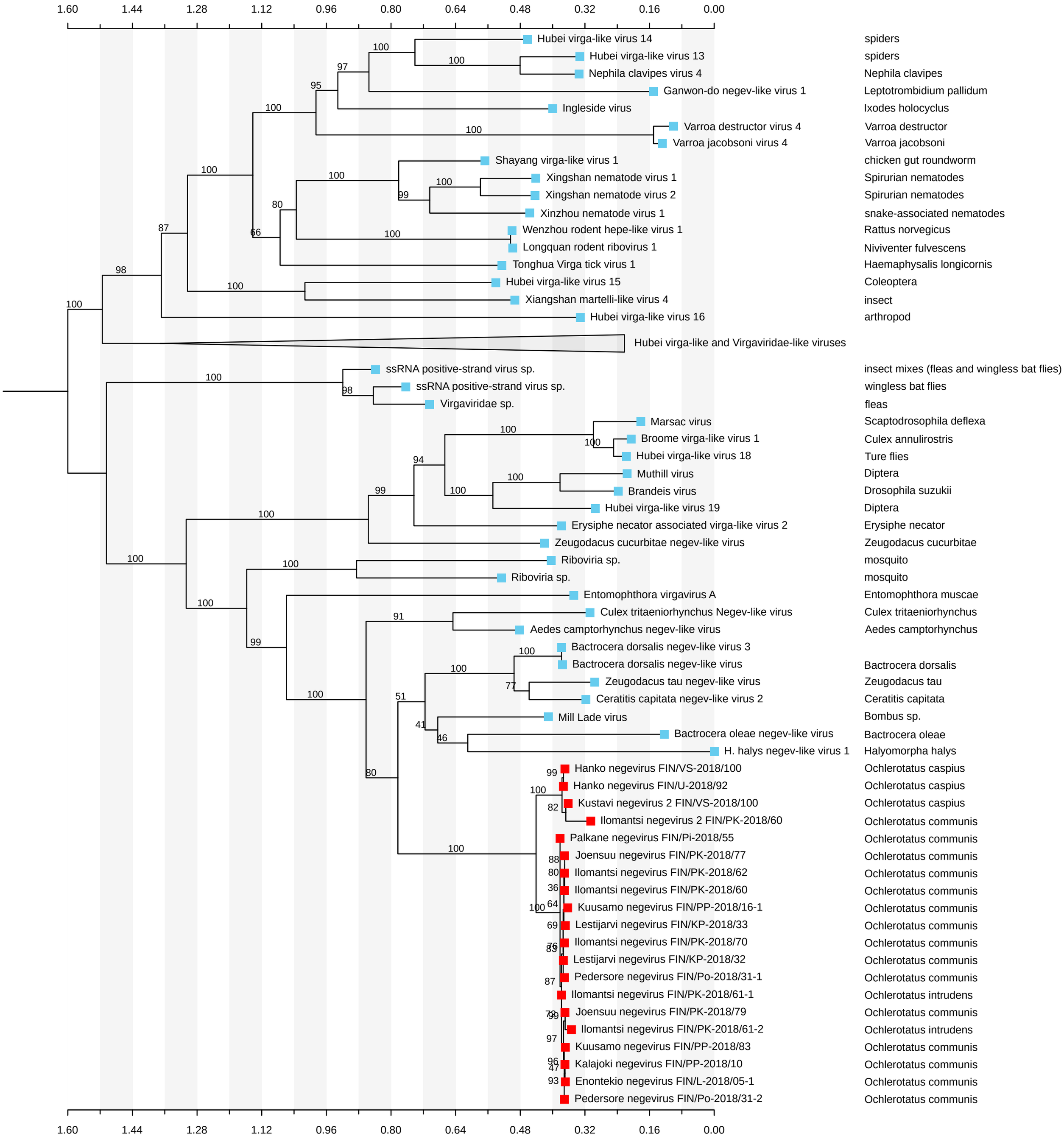

1.00
0.75
0.50
0.25
0.00
Precision
Recall
F1
Time
Maxrss
Strategy
Blastn.NT.abv
Vi.NT.nohflt
Vi.NT.extend
Vi.NT
Vi.NT.fast
Vi.RefSeq
Vi.RefSeq.fast

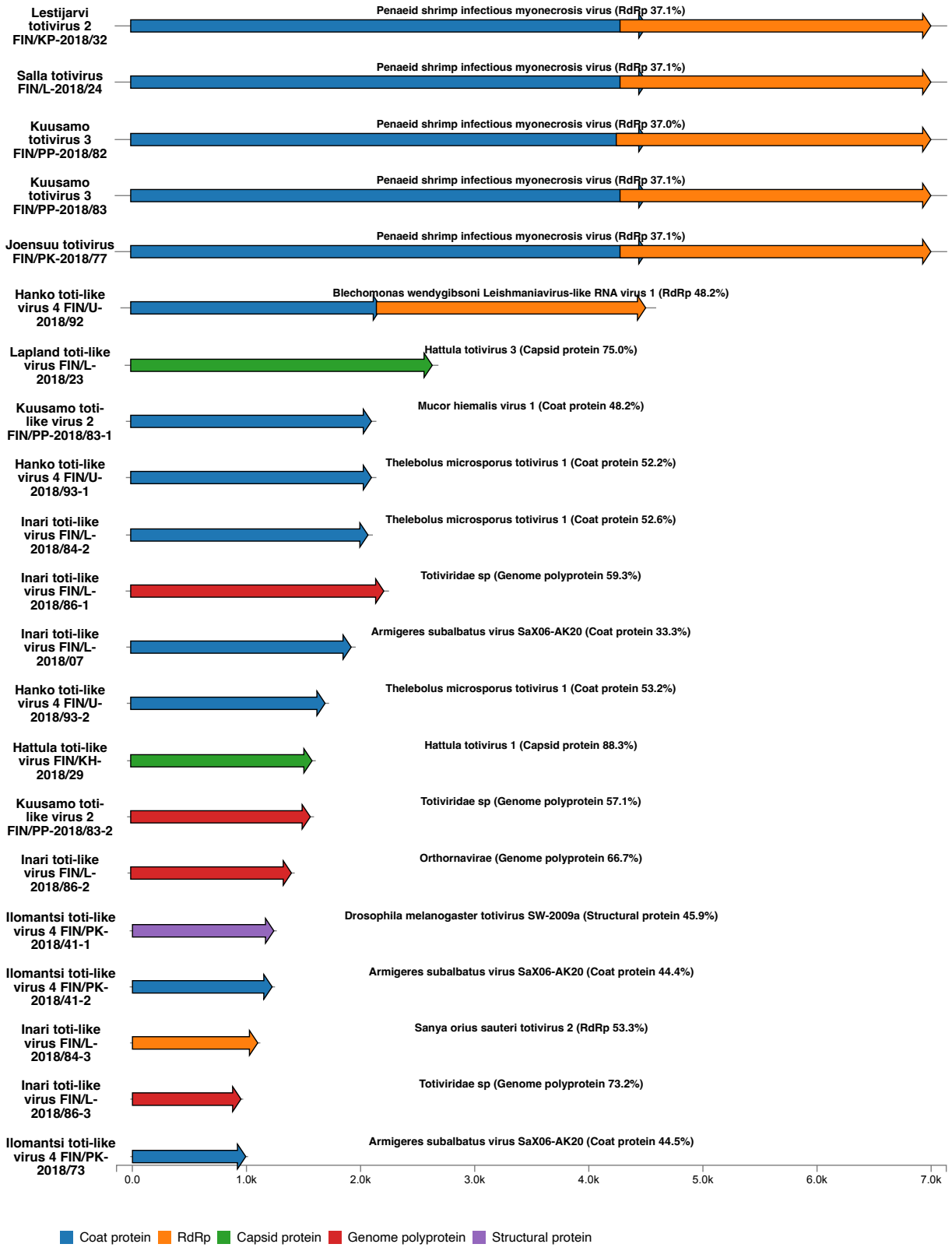

1.0
0.5
0.0
Precision
Recall
F1
Time
Maxrss
chain
chain0
chain1
chain2
chain3
chain4

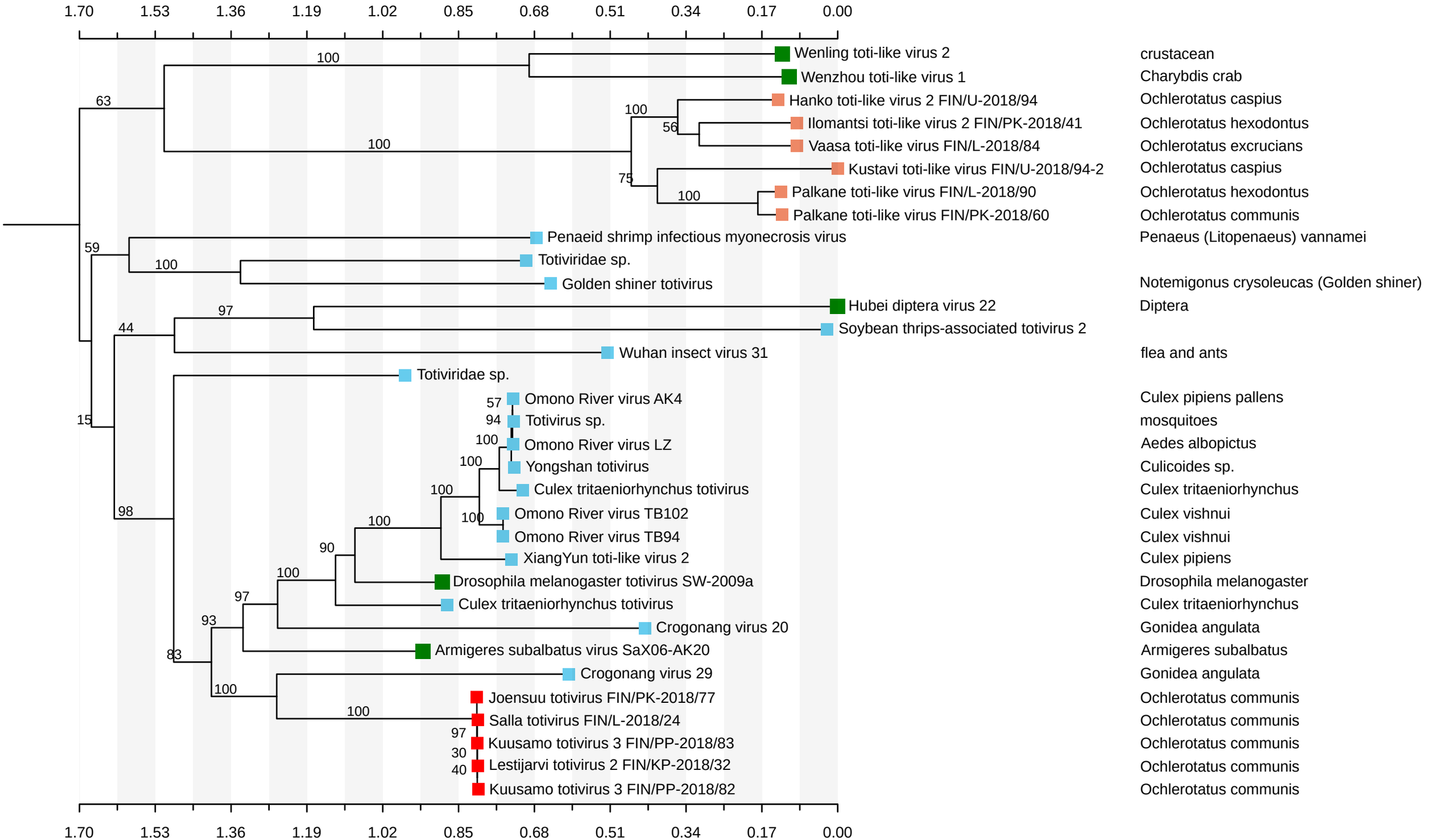

Kuusamo rhabdovirus (FIN/PP-2018/83)
Primus virus (L 52.3%)
Lieksa rhabdovirus
Xiangshan rhabdo-like virus 2 (L 35.8%)
Inari rhabdovirus 2 (FIN/L-2018/84)
Culex rhabdovirus (L 52.1%)
Enontekio Ohlsrhavirus 2 (FIN/L-2018/30-4)
Tongilchon virus 1 (N 44.5%)
Enontekio Ohlsrhavirus 2 (FIN/L-2018/89-2)
Tongilchon virus 1 (N 44.5%)
Enontekio Merhavirus 2 (FIN/L-2018/27)
Merida virus (L 52.0%)
Ohlsdorf virus (FIN/L-2018/02-1)
Ohlsdorf virus (L 74.8%)
Ohlsdorf virus (FIN/L-2018/02-2)
Ohlsdorf virus (N 61.5%)
0.0
1.0k
2.0k
3.0k
4.0k
5.0k
6.0k
7.0k
8.0k
9.0k
10k
11k
12k
L
G
M
P
N
unknown

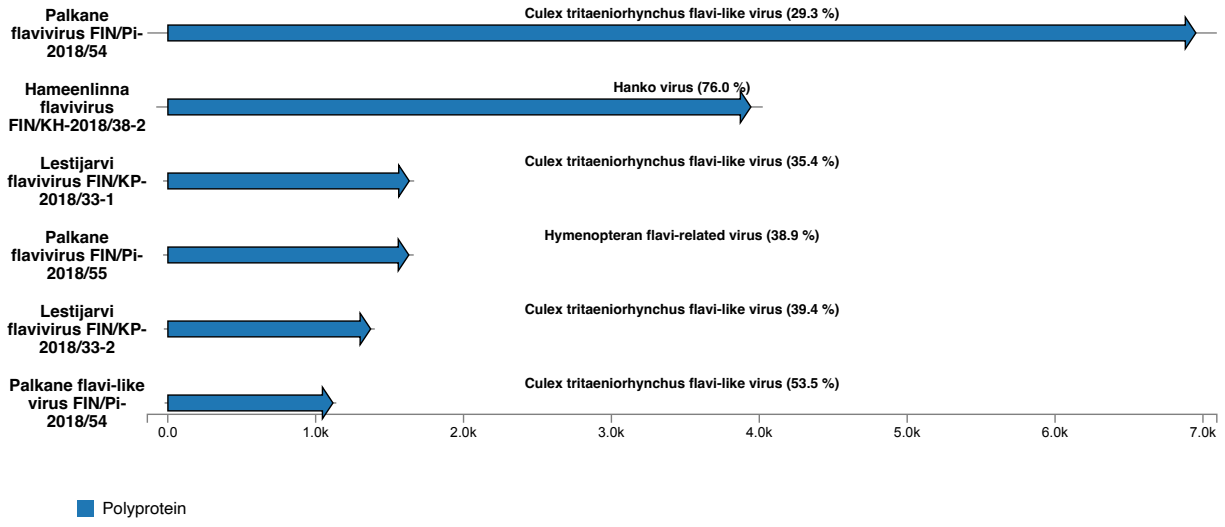

3.26 2.90 2.54 2.17 1.81 1.45 1.09 0.72 0.36 0.00
Formosus virus Merhavirus Aedes aegypti
Armigeres subalbatus rhabdovirus Armigeres subalbatus
Hattula rhabdovirus Merhavirus Ochlerotatus communis
Inari rhabdovirus Merhavirus Ochlerotatus excrucians
Culex tritaeniorhynchus rhabdovirus Culex tritaeniorhynchus
Enontekio merhavirus Merhavirus Ochlerotatus hexodontus
Culex rhabdovirus Culex sp.
Inari rhabdovirus 2 (FIN/L-2018/84) Merhavirus* Ochlerotatus excrucians
Evros rhabdovirus 2 Anopheles algeriensis
Cambodia Anopheles rhabdovirus Anopheles vagus
Corixo rhabdovirus 1 Anopheles darlingi
Tacheng Tick Virus 7 Gammaricinrhavirus Argas miniatus
Fuyun tick rhabdovirus Gammaricinrhavirus Ticks
Shayang ascaridia galli virus 2 Betanemrhavirus roundworm
Hubei rhabdo-like virus 9 Betanemrhavirus large pig roundworm
Wenling crustacean virus 10 Alphacrustrhavirus crustacean
Wenling crustacean virus 11 Alphacrustrhavirus crustacean
Blacklegged tick rhabdovirus 1 Betaricinrhavirus Ixodes scapularis
Chimay rhabdovirus Betaricinrhavirus Ixodes ricinus
Mudanjiang Rhabd tick virus 1 Betaricinrhavirus Ixodes persulcatus
Tongren Rhabd tick virus 1 Betaricinrhavirus Haemaphysalis hystricis
Yanbian Rhabd tick virus 3 Betaricinrhavirus Haemaphysalis japonica
Shayang Fly Virus 3 Alphadrosrhavirus Chrysomya megacephala
Wuhan House Fly Virus 2 Alphadrosrhavirus Musca domestica
Guadeloupe Culex rhabdovirus Stangrhavirus Culex quinquefasciatus
Elisy virus Stangrhavirus Culex tarsalis
Stang virus Stangrhavirus Culex erythrothorax
Wuhan Mosquito Virus 9 Stangrhavirus Culex tritaeniorhynchus
Lieksa rhabdovirus Stangrhavirus* Horse fly
hymenopteran rhabdo-related virus 24 Betahymrhavirus Heterodontonyx sp.
Hymenopteran rhabdo-related virus 23 Betahymrhavirus Chrysura austriaca
Xiangshan rhabdo-like virus 4 Betahymrhavirus insect
Diachasmimorpha longicaudata rhabdovirus Gammahymrhavirus Diachasmimorpha longicaudata
Apis rhabdovirus 4 Gammahymrhavirus Apis cerana
Apis rhabdovirus 5 Gammahymrhavirus Apis mellifera
Spodoptera frugiperda rhabdovirus Betapaprhavirus Spodoptera frugiperda
Lepidopteran rhabdo-related virus 34 Betapaprhavirus Triodia sylvina
Lasius neglectus virus 2 Alphahymrhavirus Lasius neglectus
Electric ant rhabdovirus Alphahymrhavirus Wasmannia auropunctata
hymenopteran rhabdo-related virus 38 Alphahymrhavirus Pompilus cinereus
Wuhan Ant Virus Camponotus japonicus
Hubei rhabdo-like virus 1 Odonata
hymenopteran rhabdo-related virus 109 Alphahymrhavirus Chlorion hirtum
Hymenopteran rhabdo-related virus 46 Alphahymrhavirus Chrysura radians
Hymenopteran rhabdo-related virus OKIAV40 Ammophila sabulosa
Xiangshan rhabdo-like virus 3 Alphahymrhavirus insect
Hymenopteran rhabdo-related virus OKIAV45 Monosapyga clavicornis
Lariophagus distinguendus n.s. RNA virus 1 Alphahymrhavirus Lariophagus distinguendus
Atrato Rhabdo-like virus 3 Primrhavirus Culex sp.
Kuusamo rhabdovirus (FIN/PP-2018/83) Primrhavirus* Ochlerotatus communis
San Gabriel mononegavirus Primrhavirus Aedes albopictus
Primus virus Primrhavirus Aedes vexans
XiangYun mono-chu-like virus 4 Culex theileri
3.26 2.90 2.54 2.17 1.81 1.45 1.09 0.72 0.36 0.00

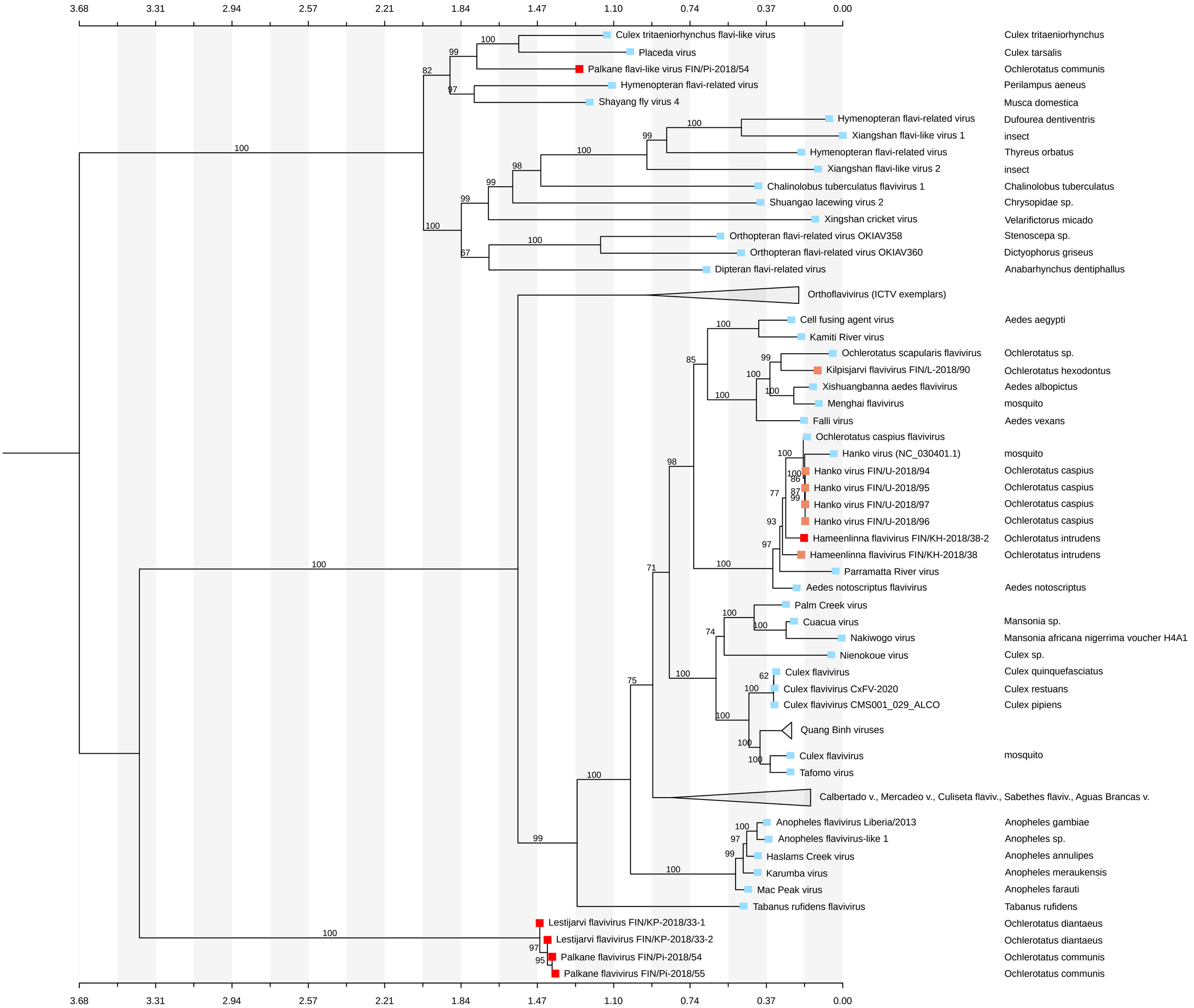

Kuusamo negevirus FIN/PP-2018/83
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 41.3%)
Hanko negevirus FIN/U-2018/92
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 41.8%)
Hanko negevirus FIN/VS-2018/100
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 41.8%)
Kustavi negevirus 2 FIN/VS-2018/100
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 41.0%)
Pedersore negevirus FIN/Po-2018/31-1
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 40.7%)
Pedersore negevirus FIN/Po-2018/31-2
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 39.0%)
Ilomantsi negevirus FIN/PK-2018/60
Halyomorpha halys negev-like virus 1 (RdRp 41.5%)
Ilomantsi negevirus FIN/PK-2018/70
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 38.0%)
Lestijarvi negevirus FIN/KP-2018/33
Halyomorpha halys negev-like virus 1 (RdRp 40.6%)
Ilomantsi negevirus FIN/PK-2018/62
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 32.5%)
Enontekio negevirus FIN/L-2018/05-1
Bactrocera dorsalis negev-like virus (RdRp 32.9%)
Lestijarvi negevirus FIN/KP-2018/32
Bactrocera dorsalis negev-like virus (RdRp 44.8%)
Palkane negevirus FIN/Pi-2018/55
Bactrocera dorsalis negev-like virus (RdRp 45.0%)
Joensuu negevirus FIN/PK-2018/79
Bactrocera dorsalis negev-like virus (RdRp 47.3%)
Ilomantsi negevirus FIN/PK-2018/61-1
Bactrocera dorsalis negev-like virus (RdRp 46.3%)
Ilomantsi negevirus 2 FIN/PK-2018/60
Bactrocera dorsalis negev-like virus 3 (RdRp 49.9%)
Ilomantsi negevirus FIN/PK-2018/60-2
Zeugodacus tau negev-like virus (RdRp 64.3%)
Enontekio negevirus FIN/L-2018/05-2
Zeugodacus tau negev-like virus (RdRp 64.6%)
Ilomantsi negevirus FIN/PK-2018/62-2
Zeugodacus tau negev-like virus (RdRp 64.7%)
Kuusamo negevirus FIN/PP-2018/16-1
Halyomorpha halys negev-like virus 1 (RdRp 33.7%)
0.0
1.0k
2.0k
3.0k
4.0k
5.0k
6.0k
7.0k
8.0k
9.0k
10k
11k
ORF4
ORF3
Glycoprotein
RdRp


1.60
1.44
1.28
1.12
0.96
0.80
0.64
0.48
0.32
0.16
0.00
Hubei virga-like virus 14 spiders
Hubei virga-like virus 13 spiders
Nephila clavipes virus 4 Nephila clavipes
Ganwon-do negev-like virus 1 Leptotrombidium pallidum
Ingleside virus Ixodes holocyclus
Varroa destructor virus 4 Varroa destructor
Varroa jacobsoni virus 4 Varroa jacobsoni
Shayang virga-like virus 1 chicken gut roundworm
Xingshan nematode virus 1 Spirurian nematodes
Xingshan nematode virus 2 Spirurian nematodes
Xinzhou nematode virus 1 snake-associated nematodes
Wenzhou rodent hepe-like virus 1 Rattus norvegicus
Longquan rodent ribovirus 1 Niviventer fulvescens
Tonghua Virga tick virus 1 Haemaphysalis longicornis
Hubei virga-like virus 15 Coleoptera
Xiangshan martelli-like virus 4 insect
Hubei virga-like virus 16 arthropod
Hubei virga-like and Virgaviridae-like viruses
ssRNA positive-strand virus sp. insect mixes (fleas and wingless bat flies)
ssRNA positive-strand virus sp. wingless bat flies
Virgaviridae sp. fleas
Marsac virus Scaptodrosophila deflexa
Broome virga-like virus 1 Culex annulirostris
Hubei virga-like virus 18 Ture flies
Muthill virus Diptera
Brandeis virus Drosophila suzukii
Hubei virga-like virus 19 Diptera
Erysiphe necator associated virga-like virus 2 Erysiphe necator
Zeugodacus cucurbitae negev-like virus Zeugodacus cucurbitae
Riboviria sp. mosquito
Riboviria sp. mosquito
Entomophthora virgavirus A Entomophthora muscae
Culex tritaeniorhynchus Negev-like virus Culex tritaeniorhynchus
Aedes camptorhynchus negev-like virus Aedes camptorhynchus
Bactrocera dorsalis negev-like virus 3
Bactrocera dorsalis negev-like virus Bactrocera dorsalis
Zeugodacus tau negev-like virus Zeugodacus tau
Ceratitis capitata negev-like virus 2 Ceratitis capitata
Mill Lade virus Bombus sp.
Bactrocera oleae negev-like virus Bactrocera oleae
H. halys negev-like virus 1 Halyomorpha halys
Hanko negevirus FIN/VS-2018/100 Ochlerotatus caspius
Hanko negevirus FIN/U-2018/92 Ochlerotatus caspius
Kustavi negevirus 2 FIN/VS-2018/100 Ochlerotatus caspius
Ilomantsi negevirus 2 FIN/PK-2018/60 Ochlerotatus communis
Palkane negevirus FIN/Pi-2018/55 Ochlerotatus communis
Joensuu negevirus FIN/PK-2018/77 Ochlerotatus communis
Ilomantsi negevirus FIN/PK-2018/62 Ochlerotatus communis
Ilomantsi negevirus FIN/PK-2018/60 Ochlerotatus communis
Kuusamo negevirus FIN/PP-2018/16-1 Ochlerotatus communis
Lestijarvi negevirus FIN/KP-2018/33 Ochlerotatus communis
Ilomantsi negevirus FIN/PK-2018/70 Ochlerotatus communis
Lestijarvi negevirus FIN/KP-2018/32 Ochlerotatus communis
Pedersore negevirus FIN/Po-2018/31-1 Ochlerotatus communis
Ilomantsi negevirus FIN/PK-2018/61-1 Ochlerotatus intrudens
Joensuu negevirus FIN/PK-2018/79 Ochlerotatus communis
Ilomantsi negevirus FIN/PK-2018/61-2 Ochlerotatus intrudens
Kuusamo negevirus FIN/PP-2018/83 Ochlerotatus communis
Kalajoki negevirus FIN/PP-2018/10 Ochlerotatus communis
Enontekio negevirus FIN/L-2018/05-1 Ochlerotatus communis
Pedersore negevirus FIN/Po-2018/31-2 Ochlerotatus communis


Lestijarvi totivirus 2 FIN/KP-2018/32
Penaeid shrimp infectious myonecrosis virus (RdRp 37.1%)
Salla totivirus FIN/L-2018/24
Penaeid shrimp infectious myonecrosis virus (RdRp 37.1%)
Kuusamo totivirus 3 FIN/PP-2018/82
Penaeid shrimp infectious myonecrosis virus (RdRp 37.0%)
Kuusamo totivirus 3 FIN/PP-2018/83
Penaeid shrimp infectious myonecrosis virus (RdRp 37.1%)
Joensuu totivirus FIN/PK-2018/77
Penaeid shrimp infectious myonecrosis virus (RdRp 37.1%)
Hanko toti-like virus 4 FIN/U-2018/92
Blechomonas wendygibsoni Leishmaniavirus-like RNA virus 1 (RdRp 48.2%)
Lapland toti-like virus FIN/L-2018/23
Hattula totivirus 3 (Capsid protein 75.0%)
Kuusamo toti-like virus 2 FIN/PP-2018/83-1
Mucor hiemalis virus 1 (Coat protein 48.2%)
Hanko toti-like virus 4 FIN/U-2018/93-1
Thelebolus microsporus totivirus 1 (Coat protein 52.2%)
Inari toti-like virus FIN/L-2018/84-2
Thelebolus microsporus totivirus 1 (Coat protein 52.6%)
Inari toti-like virus FIN/L-2018/86-1
Totiviridae sp (Genome polyprotein 59.3%)
Inari toti-like virus FIN/L-2018/07
Armigeres subalbatus virus SaX06-AK20 (Coat protein 33.3%)
Hanko toti-like virus 4 FIN/U-2018/93-2
Thelebolus microsporus totivirus 1 (Coat protein 53.2%)
Hattula toti-like virus FIN/KH-2018/29
Hattula totivirus 1 (Capsid protein 88.3%)
Kuusamo toti-like virus 2 FIN/PP-2018/83-2
Totiviridae sp (Genome polyprotein 57.1%)
Inari toti-like virus FIN/L-2018/86-2
Orthornavirae (Genome polyprotein 66.7%)
Ilomantsi toti-like virus 4 FIN/PK-2018/41-1
Drosophila melanogaster totivirus SW-2009a (Structural protein 45.9%)
Ilomantsi toti-like virus 4 FIN/PK-2018/41-2
Armigeres subalbatus virus SaX06-AK20 (Coat protein 44.4%)
Inari toti-like virus FIN/L-2018/84-3
Sanya orius sauteri totivirus 2 (RdRp 53.3%)
Inari toti-like virus FIN/L-2018/86-3
Totiviridae sp (Genome polyprotein 73.2%)
Ilomantsi toti-like virus 4 FIN/PK-2018/73
Armigeres subalbatus virus SaX06-AK20 (Coat protein 44.5%)
0.0
1.0k
2.0k
3.0k
4.0k
5.0k
6.0k
7.0k
Coat protein
RdRp
Capsid protein
Genome polyprotein
Structural protein


1.70
1.53
1.36
1.19
1.02
0.85
0.68
0.51
0.34
0.17
0.00
Wenling toti-like virus 2
crustacean
Wenzhou toti-like virus 1
Charybdis crab
Hanko toti-like virus 2 FIN/U-2018/94
Ochlerotatus caspius
Ilomantsi toti-like virus 2 FIN/PK-2018/41
Ochlerotatus hexodontus
Vaasa toti-like virus FIN/L-2018/84
Ochlerotatus excrucians
Kustavi toti-like virus FIN/U-2018/94-2
Ochlerotatus caspius
Palkane toti-like virus FIN/L-2018/90
Ochlerotatus hexodontus
Palkane toti-like virus FIN/PK-2018/60
Ochlerotatus communis
Penaeid shrimp infectious myonecrosis virus
Penaeus (Litopenaeus) vannamei
Totiviridae sp.
Golden shiner totivirus
Notemigonus crysoleucas (Golden shiner)
Hubei diptera virus 22
Diptera
Soybean thrips-associated totivirus 2
Wuhan insect virus 31
flea and ants
Totiviridae sp.
Omono River virus AK4
Culex pipiens pallens
Totivirus sp.
mosquitoes
Omono River virus LZ
Aedes albopictus
Yongshan totivirus
Culicoides sp.
Culex tritaeniorhynchus totivirus
Culex tritaeniorhynchus
Omono River virus TB102
Culex vishnui
Omono River virus TB94
Culex vishnui
XiangYun toti-like virus 2
Culex pipiens
Drosophila melanogaster totivirus SW-2009a
Drosophila melanogaster
Culex tritaeniorhynchus totivirus
Culex tritaeniorhynchus
Crogonang virus 20
Gonidea angulata
Armigeres subalbatus virus SaX06-AK20
Armigeres subalbatus
Crogonang virus 29
Gonidea angulata
Joensuu totivirus FIN/PK-2018/77
Ochlerotatus communis
Salla totivirus FIN/L-2018/24
Ochlerotatus communis
Kuusamo totivirus 3 FIN/PP-2018/83
Ochlerotatus communis
Lestijarvi totivirus 2 FIN/KP-2018/32
Ochlerotatus communis
Kuusamo totivirus 3 FIN/PP-2018/82
Ochlerotatus communis


Palkane flavivirus FIN/Pi-2018/54
Culex tritaeniorhynchus flavi-like virus (29.3 %)
Hameenlinna flavivirus FIN/KH-2018/38-2
Hanko virus (76.0 %)
Lestijarvi flavivirus FIN/KP-2018/33-1
Culex tritaeniorhynchus flavi-like virus (35.4 %)
Palkane flavivirus FIN/Pi-2018/55
Hymenopteran flavi-related virus (38.9 %)
Lestijarvi flavivirus FIN/KP-2018/33-2
Culex tritaeniorhynchus flavi-like virus (39.4 %)
Palkane flavi-like virus FIN/Pi-2018/54
Culex tritaeniorhynchus flavi-like virus (53.5 %)
0.0
1.0k
2.0k
3.0k
4.0k
5.0k
6.0k
7.0k
Polyprotein


3.68
3.31
2.94
2.57
2.21
1.84
1.47
1.10
0.74
0.37
0.00
Culex tritaeniorhynchus flavi-like virus
Culex tritaeniorhynchus
Placeda virus
Culex tarsalis
Palkane flavi-like virus FIN/Pi-2018/54
Ochlerotatus communis
Hymenopteran flavi-related virus
Perilampus aeneus
Shayang fly virus 4
Musca domestica
Hymenopteran flavi-related virus
Dufourea dentiventris
Xiangshan flavi-like virus 1
insect
Hymenopteran flavi-related virus
Thyreus orbatus
Xiangshan flavi-like virus 2
insect
Chalinolobus tuberculatus flavivirus 1
Chalinolobus tuberculatus
Shuangao lacewing virus 2
Chrysopidae sp.
Xingshan cricket virus
Velarifictorus micado
Orthopteran flavi-related virus OKIAV358
Stenoscepa sp.
Orthopteran flavi-related virus OKIAV360
Dictyophorus griseus
Dipteran flavi-related virus
Anabarhynchus dentiphallus
Orthoflavivirus (ICTV exemplars)
Cell fusing agent virus
Aedes aegypti
Kamiti River virus
Ochlerotatus scapularis flavivirus
Ochlerotatus sp.
Kilpisjarvi flavivirus FIN/L-2018/90
Ochlerotatus hexodontus
Xishuangbanna aedes flavivirus
Aedes albopictus
Menghai flavivirus
mosquito
Falli virus
Aedes vexans
Ochlerotatus caspius flavivirus
Hanko virus (NC_030401.1)
mosquito
Hanko virus FIN/U-2018/94
Ochlerotatus caspius
Hanko virus FIN/U-2018/95
Ochlerotatus caspius
Hanko virus FIN/U-2018/97
Ochlerotatus caspius
Hanko virus FIN/U-2018/96
Ochlerotatus caspius
Hameenlinna flavivirus FIN/KH-2018/38-2
Ochlerotatus intrudens
Hameenlinna flavivirus FIN/KH-2018/38
Ochlerotatus intrudens
Parramatta River virus
Aedes notoscriptus flavivirus
Aedes notoscriptus
Palm Creek virus
Cuacua virus
Mansonia sp.
Nakiwogo virus
Mansonia africana nigerrima voucher H4A1
Nienokoue virus
Culex sp.
Culex flavivirus
Culex quinquefasciatus
Culex flavivirus CxFV-2020
Culex restuans
Culex flavivirus CMS001_029_ALCO
Culex pipiens
Quang Binh viruses
Culex flavivirus
mosquito
Tafomo virus
Calbertado v., Mercadeo v., Culiseta flaviv., Sabethes flaviv., Aguas Brancas v.
Anopheles flavivirus Liberia/2013
Anopheles gambiae
Anopheles flavivirus-like 1
Anopheles sp.
Haslams Creek virus
Anopheles annulipes
Karumba virus
Anopheles meraukensis
Mac Peak virus
Anopheles farauti
Tabanus rufidens flavivirus
Tabanus rufidens
Lestijarvi flavivirus FIN/KP-2018/33-1
Ochlerotatus diantaeus
Lestijarvi flavivirus FIN/KP-2018/33-2
Ochlerotatus diantaeus
Palkane flavivirus FIN/Pi-2018/54
Ochlerotatus communis
Palkane flavivirus FIN/Pi-2018/55
Ochlerotatus communis