LazypipeX: 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) is pivotal for detecting known and novel viruses in clinical, animal, and wildlife samples, but its efficacy relies on robust bioinformatics pipelines. We present LazypipeX with multiple advanced features including flexible annotation strategies that can be adjusted to match different datasets and applications. 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 BLASTN 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 one of these “annotation chains” we were able to discover and characterise multiple novel whole-genome viral sequences that were missed by Lazypipe1, Lazypipe2 and CZ ID analysis. In LazypipeX we also implemented and benchmarked an improved background filtering and Integrative Genome Viewer reports for visualising coverage and placement of contigs relative to reference genomes. LazypipeX represents a highly efficient and versatile tool for virome analysis, offering customizable and transparent workflows that can facilitate virus discovery and identification in diverse mNGS applications.

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

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 human and animal clinical

samples, wildlife and environment (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, Höper and Beer 2015), Kaiju (Menzel, Ng and Krogh 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 Vries 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, 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; Alburkat *et al.* 2024; Erdin *et al.* 2024; Ogola *et al.* 2024).

Here we describe LazypipeX 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 on synthetic and 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 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 LazypipeX we also implemented (and benchmarked) an improved background filtering and Integrative Genome Viewer reports for visualising coverage and placement of contigs relative to reference genomes.

2 MATERIAL AND METHODS

2.1 Improved background filtering

LazypipeX supports application of multiple background filters to remove reads and contigs from non-targeted host or contaminating eukaryotic species. Removing eukaryotic backgrounds allows fast annotation of background reads and a significantly smaller reference database for contig annotation since eukaryotic sequences can be excluded from the reference. 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 acids introduced during sample collection and library preparation, it is recommended to filter *Homo sapiens* background in addition to (possibly other) host background.

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.

LazypipeX background filtering is implemented using *bwa mem* (Li 2013) 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

LazypipeX 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 (Altschul *et al.* 1997; Buchfink, Xie and Huson 2015; Somervuo and Holm 2015)) and HMMscan (Eddy 2011) searches against default or custom reference databases.

Default databases for Minimap2, BLASTN, BLASTP and DIAMOND aligners were constructed using custom Perl scripts. Specifically, databases for Minimap2 and BLASTN were created by downloading RefSeq (release 228) and NT (version 2024/01/01) sequence and accession to taxid data. Taxid lists for archaea, bacteria and viruses were created with *get_species_taxids.sh* from NCBI BLAST+ 2.16.0 package (Camacho *et al.* 2009) and mapped to lists of sequence ids. Archaeal, bacterial and virus sequences were selected to virus (Vi) and archaea-bacteria-virus (ABV) databases using SeqKit (Shen *et al.* 2016). BLASTN indexes were created with NCBI BLAST+ 2.16.0 makeblastdb (Camacho *et al.* 2009). Minimap2 databases were implemented as fasta files combined with accession-to-taxid tsv files. Databases for BLASTP and DIAMOND were created by downloading the latest UniRef100 sequence and accession to taxid data (version 2024/11/27) and by using seed sequences for the cluster-to-taxid mapping. Archaeal, bacterial and virus sequences were selected to Vi and ABV databases using NCBI BLAST+ get_species_taxids.sh (Camacho *et al.* 2009) and SeqKit (Shen *et al.* 2016). BLASTP index was compiled with NCBI BLAST+ 2.16.0 makeblastdb (Camacho *et al.* 2009). DIAMOND index was compiled with DIAMOND version 2.1.6 (Buchfink, Xie and Huson 2015). HMMscan databases were compiled by downloading all viral HMM profiles from PFAM (Mistry *et al.* 2021) and the NeoRdRp2.1 database (Sakaguchi, Nakano and Nakagawa 2024). HMM databases were indexed with hmmpress from HMMER package v3.4 (Eddy 2011). SANSparallel was run on the UniProtKB database via the online server (Somervuo and Holm 2015). All databases described in this section were made freely available for download with stable links and *install_db.pl* utility included with LazypipeX distribution.

For each of the applied database search we used bitscore and/or e-value cut-offs specified in the main config.yaml file. Minimap2 was run with options "-x asm20 –secondary yes –cs -s 200" followed by selection of top hits without ties. BLASTN was run with options "-evalue 10 -max_target_seqs 5" followed by selection of tophits that exceeded bitscore cutoff of 200. BLASTP was run with options "-evalue 10 -max_target_seqs 5" followed by selection of tophits that exceeded bitscore cutoff of 200. DIAMOND blastp was run with options "--evalue 10 --max_target_seqs 5 --very-sensitive" followed by selection of tophits that exceeded bitscore cutoff of 200. SANSparallel server was accessed using PANNZER v3 (Törönen and Holm 2022) with options "-m SANStopHtaxid --SANS_H 5 -R" followed by selection of tophits that exceeded bitscore cutoff of 200. For BLASTP, DIAMOND and SANSparallel ORFs were extracted using MGA (Noguchi, Taniguchi and Itoh 2008) with default options and excluding ORFs shorter than 72 nucleotides. DIAMOND blastx was run with options "--evalue 10 --min_orf 72 --

query-gencode 1 --max-target-seqs 5 --very-sensitive" followed by selection of tophits without ties that exceeded bitscore cutoff of 200. HMMscan was run with "-E 0.01" followed by selection of tophits with ties.

Here, we tested several annotation strategies dividing them into one- and two-round strategies for virus identification, and chained strategies for virus discovery. All benchmarked strategies listed here are predefined in the pipeline configuration file and can be easily accessed from the command line.

2.2.1 Fast one-round strategies

Minimap.NT$_{Vi}$ and Minimap.RefSeq$_{Vi}$: Minimap2 maps all contigs against NCBI NT or RefSeq viruses (databases *Minimap.NT$_{Vi}$* and *Minimap.RefSeq$_{Vi}$* in Table1). The goal is to provide very fast annotations, e.g. for preliminary analysis, at the expense of possibly higher false positive rates. For one-round and two-round strategies we use a naming convention with database search (e.g. Minimap2) followed by reference database name (e.g. NT) and database segment subscript (e.g. Vi for viruses).

2.2.2 Two-round strategies for virus identification

In two-round annotations the pipeline first runs a fast search against a virus-only database, e.g. Minimap2 on NCBI NT viruses and then prunes any false positives in the second round with a slower BLASTN search on a more extensive database. The two-round strategy aims to speed up the pipeline by reducing the search space covered by the slower search, while keeping false positive rate in check. To explain the rationale behind this approach, let's assume that execution time scales linearly to the size of the query and reference data. Then, execution time for annotation of all contigs with the slower and more sensitive search against the complete database can be expressed as:

1) $T_1 \sim S(Q_{all})\, S(R_{all})\, V_{slow}$

And execution time for annotating only virus contigs in two-rounds can be expressed as:

2) $T_2 \sim S(Q_{all})\, S(R_{vi})\, V_{fast} + S(Q_{vi})\, S(R_{all})\, V_{slow}$

Where, $S(Q_{all})$, is the size of the assembly, $S(Q_{vi})$, is the size of virus part of the assembly, $S(R_{all})$, is the size of the reference database, $S(R_{vi})$, is the size of the virus part of the reference, V_{slow} and V_{fast}, are the speed of the slower and the faster search.

Finally, assuming both searches are executed in a similar environment, the ration of execution times can be approximated by:

3) $\frac{T_2}{T_1} \approx \frac{S(Q_{all})\, S(R_{vi})\, V_{fast} + S(Q_{vi})\, S(R_{all})\, V_{slow}}{S(Q_{all})\, S(R_{all})\, V_{slow}} = \frac{S(R_{vi})\, V_{fast}}{S(R_{all})\, V_{slow}} + \frac{S(Q_{vi})}{S(Q_{all})}$

For example, if V_{fast}/V_{slow} is ~1/10, and if viruses make up ~10% of the query data and 15% of the reference data, then T2/T1 ≈ 0.115. In this case the two-round strategy is expected to yield a ~9-fold speedup. We argue that these assumptions are realistic. For example, in NT released 2024/01/01 virus

176 sequences constituted 15.3% and the proportion of assembled viral data in most metagenomics
177 datasets is below 10% (e.g. in (Erdin *et al.* 2025) library from rRNA depleted tissue RNA this was
178 1.68%), and e.g. Minimap2, although not specifically evaluated against BLASTN, is reported to be tens
179 of times faster than long read genomic mappers.

180 Keeping this rationale in mind we tested two two-round strategies: *Minimap.NT*$_{Vi}$ *-Blastn.NT*$_{ABV}$ and
181 *Minimap.RefSeq*$_{Vi}$*-Blastn.RefSeq*$_{ABV}$.

182 *Minimap.NT*$_{Vi}$ *-Blastn.NT*$_{ABV}$*:* In the 1st round map all contigs with Minimap2 against NT viruses (NT$_{Vi}$,
183 database descriptions are given in Table1). In the 2nd round annotated pre-labelled viral contigs with
184 BLASTN against NT archaea, bacteria and viruses (NT$_{ABV}$). The strategy includes background filtering
185 and assumes that most eukaryotic sequences are removed before the annotation step. This allows to
186 reduce the database for the 2nd round to NT$_{ABV}$.

187 *Minimap.RefSeq*$_{Vi}$*-Blastn.RefSeq*$_{ABV}$*:* In the 1st round map all contigs with Minimap2 against RefSeq
188 virus (RefSeq$_{Vi}$). In the 2nd round annotated pre-labelled viral contigs with BLASTN against RefSeq
189 archaea, bacteria and viruses (RefSeq$_{ABV}$). The strategy includes background filtering.

190 These two strategies were designed for virus identification focusing on speed and low false positive
191 rate. These are suited for samples with clearly defined background sequences such as tissue, swab,
192 plasma and arthropod samples.

193 *Minimap.NT-Blastn.NT:* In the 1st round map all contigs with Minimap2 against complete NT. In the 2nd
194 round annotated pre-labelled viral contigs with BLAST against complete NT. This strategy has no
195 background filtering and is suited for samples with complex backgrounds (e.g. faeces or environment
196 samples).

197 *Table 1. List of reference databases supported by LazypipeX.*

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.NT$_{Vi}$	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)

198

199 *2.2.3 Chained strategies for virus discovery*

200 This section describes chained annotation strategies designed for virus discovery with focus on
201 reducing false negative rate while keeping the execution time feasible. In the chained annotation the
202 pipeline starts with a fast nucleotide search such as Minimap2. Contigs that have no hits in the first
203 search are channelled to the second search that is expected to yield higher sensitivity either due to the
204 search engine or the database. The process is then iterated to searches that have increasingly higher
205 sensitivity, and in most cases lower precision and lower speed. By chaining several searches this
206 strategy aims to improve the overall sensitivity. Strategies *Vi.chain1* to *Vi.chain4* include background
207 filtering and assume that most eukaryotic reads/contigs are removed prior to annotation. This allows to
208 limit applied reference databases to archaea-bacteria-virus segments (NT_{ABV} and $UniRef100_{ABV}$). These
209 are suited for samples with clearly defined eukaryotic background sequences such as tissue, swab,
210 plasma and arthropod samples. Strategy Vi.chain3.env does not filter background and annotates
211 against complete databases. This is suited for samples with complex backgrounds (e.g. faecal or
212 environment samples).

213 *Vi.chain0:* This is a baseline strategy using BLASTN and BLASTP, which represent the more reliable
214 tools for identification of viral sequences (Tangherlini *et al.*, 2016). Contigs are annotated with BLASTN
215 against NT_{ABV} (Table 1), then contigs with no hits are annotated with BLASTP against $UniRef100_{ABV}$.

216 *Vi.chain1:* In this strategy Minimap2 is introduced in front of BLASTN. Contigs are annotated with
217 Minimap2 against NT_{ABV}, then contigs with no hits are annotated with BLASTN against NT_{ABV}, then
218 contigs with no hits with BLASTP against $UniRef100_{ABV}$.

219 *Vi.chain2:* In this strategy BLASTN is omitted from the chain. Contigs are annotated with Minimap2
220 against NT_{ABV}, then contigs with no hits with BLASTP against $UniRef100_{ABV}$.

221 *Vi.chain3:* In this strategy, BLASTP is substituted with a faster protein search, DIAMOND blastp
222 (Buchfink, Xie and Huson 2015). Contigs are annotated with Minimap2 against NT_{ABV}, then contigs with
223 no hits with DIAMOND blastp in very-sensitive mode against $UniRef100_{ABV}$. We also tested DIAMOND
224 blastp in ultra-sensitive mode with no change in sensitivity and DIAMOND blastx in both modes with
225 clearly lower sensitivity (data not shown). ORFs for DIAMOND search are extracted using MGA
226 (Noguchi, Taniguchi and Itoh 2008) with default options and excluding ORFs shorter than 72
227 nucleotides.

228 *Vi.chain4:* In this strategy BLASTP is substituted with SANSparallel. Contigs are annotated with
229 Minimap2 against NT_{ABV}, then contigs with no hits with SANSparallel against UniProtKB (via
230 SANSparallel online server). ORF extraction is identical to *Vi.chain3*.

231 *Vi.chain3.env*: Performs no background filtering and annotates against complete databases. Contigs
232 are annotated with Minimap against NT, then contigs with no hits with Diamond blastp against
233 UniRef100.

234 *2.2.4 Supported reference databases*

235 LazypipeX is supported with several precompiled and updated reference databases available from a
236 public server (Table 1). These can be installed to local volumes with *install_db.pl* utility that is included
237 in the LazypipeX distribution. The default nucleotide database is the NCBI NT, which was also the most
238 common reference database employed by 8 out 13 pipelines examined by de Vries et al (de Vries *et al.*
239 2021) and by CZ ID (Kalantar *et al.* 2020). For the default protein database we chose UniRef100,
240 similarly to GenomeDetective which uses UniRef90 (Vilsker *et al.* 2019). Other options would be NCBI
241 NR, e.g. DAMIAN (Alawi *et al.* 2019) and CZ ID (Kalantar *et al.* 2020), and RefSeq proteins, e.g.
242 metaMix (Morfopoulou and Plagnol 2015), but we choose UniRef100 as in our view it is both
243 representative and at the same time an order of magnitude smaller than NR.

244 2.3 Other new features

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

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

252 LazypipeX can optionally output Integrative Genome Viewer (IGV) reports (Robinson *et al.* 2023) for
253 identified viral contigs. These reports are produced by first binning contigs by the assigned species. For
254 each species top hits from the annotation table are then retrieved and aligned to contigs with Minimap2.
255 BAM files are then converted to an interactive IGV reports visualising genome coverage and variants.
256 These reports are convenient for identifying contig overlaps in fragmented assemblies as well as
257 sequence similarity and variants. For well annotated reference sequences these reports can also
258 provide information on contig coverage of key genes such as polymerases. Example IGV reports for
259 viruses identified in 10 Finnish mosquito pools (accessions SRR19859349-SRR19859358) with
260 *Vi.chain1* annotation strategy are given in Supp. File 1.

261 2.4 Benchmarking

262 *2.4.1 Benchmarking background filtering*

263 We evaluated LazypipeX background filtering on two datasets. For the first one we used the human
264 simulated metagenome from the MetaShot project (Fosso *et al.* 2017). Reads were labelled as human
265 reads or reads from other taxa based on sequence accessions included in the id field. Thresholds for
266 alignment score (AS) and mapping quality (mapq) were optimized by stepwise iteration over a range of
267 values. Background reads were filtered with BWA MEM (Li 2013) and SeqKit (Shen *et al.* 2016) using
268 human reference genome (acc. GCF_000001405.40). For the second dataset we used Illumina paired-

end 2 x 150 bp libraries sequenced from two liver samples collected from two Finnish brown rats (*Rattus norvegicus*) of which *Rocahepevirus ratti* was previously detected (Erdin *et al.* 2025). For rat samples background was filtered with *Rattus norvegicus* reference genome (acc. GCF_036323735.1) and background filtering was evaluated against read binning by LazypipeX with BLASTN against NT.

2.4.2 Benchmarking strategies for virus identification on simulated data

We evaluated LazypipeX with one-round and two-round strategies designed for virus identification on the human simulated metagenome (Fosso *et al.* 2017). This dataset is a simulated 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). Specifically, we compared two-round strategy $Minimap.NT_{Vi}$-$Blastn.NT_{ABV}$ against a one-round $Blastn.NT_{ABV}$, two-round strategy $Minimap.NT_{Vi}$-*Blastn.NT* against one-round *Blastn.NT (*for definitions see sections 2.2.1 and 2.2.2). The corresponding two-round and one-round strategies were compared to investigate expected time savings and possible reduction in accuracy for the two-round approach. We also run $Minimap.RefSeq_{Vi}$-$Blastn.RefSeq_{ABV}$ to test annotations on a smaller RefSeq database, and $Minimap.NT_{Vi}$ and $Minimap.RefSeq_{Vi}$ to see how a single round of Minimap annotations would compare to other strategies. In addition, we included results with DIAMOND blastp against complete NR (labelled as *Diamondp.NR*) as this represents a common approach in viral metagenomics. DIAMOND blastp was run as an annotation strategy within LazypipeX as described in section 2.2.

To compare against older releases of Lazypipe we included results with Lazypipe version 2.1 with background filtering against human reference assembly (accession GCF_000001405.40) and Minimap annotation (--ann minimap) against NT. Lazypipe2 with Minimap against NT had best performance in previous evaluations (Plyusnin *et al.* 2023). Finally, we included results from CZ ID web pipeline, version Illumina v8.3 with NCBI index dated 2024/02/06 and human host subtraction (Kalantar *et al.* 2020). For all local pipeline runs we used RefSeq release 228, NT dated 2024/01/01, UniRef100 dated 2024/11/27 and NR dated 2025/08/07, and for all local runs, except $Minimap.NT_{Vi}$-*Blastn.NT* and *Blastn.NT,* we filtered background with human reference assembly (accession GCF_000001405.40).

To properly perform benchmarking sequence accessions present in human simulated metagenome (Fosso *et al.* 2017) were linked to NCBI taxonomy dated 2025/09/10. Also, for all compared pipelines reported taxids were linked to NCBI taxonomy dated 2025/09/10 using taxonkit (Shen and Ren 2021). For CZ ID read counts were selected from NT count field, when this was non-zero, and from NR count field when NT count was zero. For all pipelines taxa with less then 10 reads and taxa labelled as bacteriophages were ignored. All benchmarking was 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 annotation strategies Vi.chain0, Vi.chain1, Vi.chain2, Vi.chain3, Vi.chain4 and Vi.chain3.env on Illumina MiSeq dataset from

305 90 pools of *Ochlerotatus mosquitoes* (NCBI BioProject PRJNA852425) that was previously analysed
306 with Lazypipe1 and downstream BLASTX (Truong Nguyen *et al.* 2022). Background was filtered with
307 four mosquito chromosome level assemblies publicly available at the moment of analysis (*Aedes*
308 *aegypti* acc. GCF_002204515.2, *Aedes albopictus* acc. GCF_035046485.1, *Culex pipiens pallens* acc.
309 GCF_016801865.2, *Culex laticinctus* acc. CA_964187845.1) and the human reference assembly
310 (GCF_000001405.40). Other pipeline options were kept at default. To compare against previous
311 pipeline releases the same data was analysed with Lazypipe2 --ann minimap and identical background
312 filtering.

313 Additionally, we analysed a previously unpublished Illumina MiSeq library sequenced from a Finnish
314 horse fly (*Tabanidae sp.*) to test how HMMscan can augment annotation chains for virus discovery.
315 Background was filtered with a chromosome level assembly from a related species (*Chrysops*
316 *caecutiens*, acc. GCA_963971475.1) and the human reference assembly (GCF_000001405.40). For
317 annotation we applied a strategy that is equivalent to *Vi.chain1* with added HMMscan steps. Specifically,
318 we used Minimap2 against NT_{ABV}, then BLASTN against NT_{ABV}, then BLASTP against $UniRef100_{ABV}$,
319 then HMMscan against viral HMM profiles collected from Pfam (Mistry *et al.* 2021) and NeoRdRp v2.1
320 (Sakaguchi, Nakano and Nakagawa 2024). Other pipeline options were kept at default.

321 We also compared LazypipeX virus discovery against web CZ ID on three samples with novel findings
322 by LazypipeX: the horse fly sample and two mosquito samples (accessions SRR19859382 and
323 SRR19859383). These were analysed with CZ ID version Illumina v8.3, NCBI index 2024/02/06 and
324 "mosquito" host subtraction for the mosquito samples.

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

327 2.5 Characterization of viral findings

328 Annotated array plots visualising ORF structure and closest homologs were constructed for selected
329 viral findings as follows. Open reading frames (ORFs) were identified with orfipy (Singh and Wurtele
330 2021), setting minimal ORF length to 300, excluding partial 5' ORFs, including partial 3' ORFs and using
331 the standard genetic code. Smaller ORFs residing completely inside larger ORFs regardless of the
332 strand were excluded. Extracted ORFs were annotated with the top hit from DIAMOND blastp
333 v2.1.10.164 (Buchfink, Reuter and Drost 2021) against UniRef100 viral segment (database retrieved
334 27/11/2024). ORFs were classified into groups using reciprocal best BLASTP links for pairs of contigs
335 and manual editing. These groups were then labelled either by gene or protein name most commonly
336 appearing in the assigned ORF annotations. ORF structure was visualized as an array plot using
337 geneviewer (Velden 2025).

338 Boostrapped maximum likelihood (ML) trees were constructed for selected viral findings and
339 corresponding references as follows. For the "*Inari merhavirus*" refenreces included all ICTV
340 *Alpharhabdovirinae* isolates, *Rhabdoviridae* sequences previously identified from this dataset (Truong

Nguyen *et al.* 2022), and BLASTX homologs in clustered NR (with bitscore cut-off set to 400). For "*Primrhavirus kuusamo"* and *"Stangrhavirus lieksa"* references included all ICTV *Deltarhabdovirinae* isolates, *Rhabdoviridae* sequences previously identified from this dataset (Truong Nguyen *et al.* 2022), and BLASTX homologs in clustered NR (with bitscore cut-off set to 400). From these, L-gene protein sequences were extracted, excluding short sequences (<1500 aa) and sequences with no BLASTP homology to the new viruses. Sequences were aligned with MAFFT v7.505 (Katoh and Standley 2013) with options *--retree 2 –maxiter 100* and trimmed with trimAl 1.2rev59 (Capella-Gutiérrez, Silla-Martínez and Gabaldón 2009) with *-automated1* option. For negevirus-like contigs references were collected by mapping BLASTX hits to at most 100 nucleotide accessions and complementing these with exemplar isolates of *Negevirus* member species as listed in (Nunes *et al.* 2017). 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. Previously published negevirus- sequences (Truong Nguyen *et al.* 2022) were not included due to poor alignment with the new and reference sequences. For artivirus contigs references were collected by mapping BLASTX hits to at most 100 nucleotide accessions, adding all *Artitivirus* isolates from ICTV 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 rhabdovirus contigs. Alignments for ORF1 and ORF2 were concatenated. For flavivirus contigs references were collected by mapping BLASTX hits with bitscore cutoff at 200 to at most 200 nucleotide accessions complementing these with all *Orthoflavivirus* isolates from ICTV 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 re-rooted on the outgroup for rhabdoviruses (infectious hematopoietic necrosis virus, acc. L40883) or midpoint for other viruses 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 Table 2. Filtering by mapping quality (mapq) was less effective than filtering by alignment score (AS). For the default threshold we selected AS>=100. At this level 99.998% of human reads were filtered, while none of viral reads that were not retroviruses. Notably, 96.375% of human endogenous retroviruses were removed (ERV), which is likely

due to ERV sequences in the human reference genome. To correctly bin ERV sequences we suggest using annotation strategies without background filtering (see discussion).

Table 2 Benchmarking background filtering on simulated data from Fosso et al. (Fosso et al. *2017). Selected filter highlighted. Filter, applied filter with AS representing alignment score and mapq mapping quality threshold, Total, total fraction of reads filtered, Homo sapiens, human reads filtered, Bacteria, Bacterial reads filtered, Viruses, all viral reads filtered, HERV, reads from Human Endogeneous Retroviruses filtered, Viruses excl. Retroviridae, filtered viral reads excluding Retroviridae reads.*

Filter	Total	Homo sapiens	Bacteria	Viruses	HERV	Viruses excl. Retroviridae
AS0	99.587 %	100.000 %	91.696 %	97.529 %	99.544 %	96.804 %
AS30	94.687 %	100.000 %	0.021 %	21.982 %	96.453 %	0.151 %
AS50	94.686 %	99.999 %	0.002 %	21.913 %	96.450 %	0.046 %
AS75	94.685 %	99.999 %	0.001 %	21.887 %	96.444 %	0.008 %
AS100	94.684 %	99.998 %	0.000 %	21.861 %	96.354 %	0.000 %
AS125	94.681 %	99.997 %	0.000 %	21.701 %	95.649 %	0.000 %
AS150	76.305 %	80.631 %	0.000 %	11.793 %	51.980 %	0.000 %
mapq0	99.587 %	100.000 %	91.696 %	97.529 %	99.544 %	96.804 %
mapq10	90.708 %	94.938 %	16.278 %	26.474 %	52.932 %	20.282 %
mapq20	89.851 %	94.855 %	1.954 %	12.879 %	50.600 %	2.122 %
mapq30	89.159 %	94.219 %	0.382 %	10.595 %	45.364 %	0.459 %
mapq40	89.043 %	94.117 %	0.038 %	10.147 %	44.656 %	0.022 %
mapq50	88.564 %	93.623 %	0.005 %	8.680 %	38.240 %	0.005 %
mapq60	88.285 %	93.331 %	0.001 %	8.266 %	36.431 %	0.000 %

Results for background filtering for rat samples are presented in Supp. Table 1. Also for this data filtering by alignments score was the best option with threshold at 100 filtering 63-70% of rodent reads and only a few viral (<0.05%) or bacterial (<0.001) reads.

Rat sample HEV-226					Rat sample HEV-227				
Filter	Total	Rodents	Bacteria	Viruses	Filter	Total	Rodents	Bacteria	Viruses
AS0	38.645 %	99.977 %	59.382 %	50.457 %	AS0	29.894 %	99.961 %	58.730 %	52.605 %
AS30	4.473 %	99.792 %	0.015 %	0.016 %	AS30	0.614 %	99.934 %	0.003 %	0.050 %
AS50	4.107 %	99.780 %	0.002 %	0.001 %	AS50	0.537 %	99.934 %	0.001 %	0.050 %
AS75	3.311 %	99.764 %	0.001 %	0.000 %	AS75	0.440 %	99.750 %	0.000 %	0.050 %
AS100	2.190 %	70.186 %	0.000 %	0.000 %	AS100	0.264 %	63.096 %	0.000 %	0.050 %
AS125	1.350 %	42.986 %	0.000 %	0.000 %	AS125	0.110 %	25.982 %	0.000 %	0.050 %
AS150	0.689 %	22.063 %	0.000 %	0.000 %	AS150	0.038 %	8.621 %	0.000 %	0.050 %
mapq0	38.645 %	99.977 %	59.382 %	50.457 %	mapq0	29.894 %	99.961 %	58.730 %	52.605 %
mapq10	16.435 %	2.751 %	29.209 %	23.875 %	mapq10	17.711 %	2.602 %	33.633 %	22.796 %
mapq20	15.604 %	2.640 %	27.513 %	21.395 %	mapq20	13.985 %	2.471 %	25.234 %	20.341 %
mapq30	15.562 %	2.215 %	27.495 %	21.206 %	mapq30	13.972 %	2.234 %	25.217 %	20.240 %
mapq40	15.549 %	2.046 %	27.487 %	21.193 %	mapq40	13.969 %	2.037 %	25.216 %	20.240 %
mapq50	0.320 %	1.239 %	0.131 %	4.981 %	mapq50	0.335 %	1.971 %	0.266 %	2.004 %
mapq60	0.089 %	0.797 %	0.002 %	0.001 %	mapq60	0.010 %	0.053 %	0.001 %	0.050 %

Supp. Table 1 Benchmarking background filtering on real data with Rattus norvegicus background (Erdin et al., 2025). Selected filter highlighted. Filter, applied filter with AS representing alignment score and mapq mapping quality threshold, Total, total fraction of reads filtered, Rodents, rodent reads filtered, Bacteria, Bacterial reads filtered, Viruses,Viral reads filtered.

3.2 Virus identification from simulated data

We evaluated LazypipeX with several one-round and two-round strategies, Lazypipe2 and web CZ ID on the human simulated metagenome (Fosso *et al.* 2017) as detailed in section 2.4.2. Precision, recall,

396 F1-score (harmonic mean of precision and recall) in terms of predicted virus taxa at species level,
397 execution time, maximum memory consumption and erroneous predictions are given in Table 3.

398 The lowest error rate was observed for LazypipeX annotations against complete NT without host
399 filtering, with both one-round (*Blastn.NT*) and two-round (*Minimap.NT*$_{Vi}$*-Blastn.NT*) strategies reporting
400 two false positive (FP) and two false negative (FN) species. FNs were *Mopeia Lassa virus reassortant*
401 *29* (ML29) which was misclassified to *Mammarenavirus lassaense*ı and *Orthobornavirus bornaense*
402 with no annotation for the corresponding contig. Notably, ML29 contains S-segment of the *Lassa virus*
403 which caused the misclassification. FPs were *Orthopoxvirus cowpox* due to one misclassified *Vaccinia*
404 *virus* contig, and *African green monkey-like simian foamy virus* due to one misclassified *Human*
405 *spumaretrovirus* contig. The execution time for one-round version was ~117h and for two-round version
406 ~27h showing a 4.3X speedup.

407 The next most accurate strategies were one and two-round annotations against NT$_{Vi}$/NT$_{ABV}$ with
408 background filtering to remove eukaryotic sequences. Both two-round (*Minimap.NT*$_{Vi}$*-Blastn.NT*$_{ABV}$) and
409 one-round (*Blastn.NT*$_{ABV}$) reported two FP and three FN species. FNs include two endogenous
410 retroviruses (*Human endogeneous retrovirus* and *Human endogenous retrovirus W*) that were filtered
411 with the host genome and *ML29* which was misclassified to *Mammarenavirus lassaense*. About half
412 (~1000) of *Orthopneumovirus hominis* reads were misclassified as *O. bovis* (FP). One *Vaccinia virus*
413 contig was misclassified as a *Orthopoxvirus cowpox* (FP). The execution time for the one-round version
414 (*Blastn.NT*$_{ABV}$) was ~32h and for two-round version (*Minimap.NT*$_{Vi}$*-Blastn.NT*$_{ABV}$) ~4h showing 8.4X
415 speedup. Annotating with just one round of Minimap2 against NT viruses (*Minimap.NT*$_{Vi}$) was ~3.4X
416 faster than with two rounds (*Minimap.NT*$_{Vi}$*-Blastn.NT*$_{ABV}$) with no decrease in accuracy.

417 Annotating against smaller RefSeq database with one and two rounds (*Minimap.RefSeq*$_{Vi}$-
418 *Blastn.RefSeq*$_{ABV}$ and *Minima.RefSeq*$_{Vi}$) had higher error rate compared to annotations against NT
419 resulting in three FP and six FN species.

420 Annotating with Lazypipe2 resulted in three FN and 16 FP predictions and execution time comparable
421 to that of the newer 2-round strategy (*Minimap.NT*$_{Vi}$*-Blastn.NT*$_{ABV}$). FNs included the same three viruses
422 as in *Minimap.NT*$_{Vi}$*-Blastn.NT*$_{ABV}$. Lazypipe2 also reported *Orthopneumovirus bovis* (FP) and additional
423 14 false positive bacteriophage species (Table 3).

424 Annotating with DIAMOND Blastp against complete NR (*Diamondp.NR*) had low recall and precision
425 compared to LazypipeX NT-based two-round annotations. *Diamondp.NR* reported five FPs and failed
426 to report 15 viral pathogens present in the benchmark (Table 3).

427 Annotating with online CZ ID pipeline resulted in four FNs and 20 FPs. FNs included two retroviruses
428 likely filtered with host genome, *Betapapillomavirus 2* and *Bocaparvovirus primate1,* which were likely
429 misclassified to *Human papillomavirus* (FP) and *Human bocavirus* (FP). The latter two of the reported
430 viruses refer to taxonomic entities very close to the correct predictions and may be attributed to
431 discrepancies between taxonomies applied in analysis and evaluation. In addition to these CZ ID

reported another four FP eukaryotic viruses and 14 unlabelled bacteriophages leading to an overall low precision (Table 3).

To summarize, switching from one-round annotations (*Blastn.NT* and *Blastn.NT$_{ABV}$*) to two-round annotations (*Minimap.NT$_{Vi}$-Blastn.NT* and *Minimap.NT$_{Vi}$-Blastn.NT$_{ABV}$)* resulted in an ~4-8X speedup with no decrease in accuracy. Background filtering and annotating in two rounds on the NT$_{ABV}$ segment versus annotating on complete NT resulted in ~7X speedup. Altogether, switching from BLASTN annotation on complete NT (*Blastn.NT*) to two-round annotation on NT$_{ABV}$ (*Minimap.NT$_{Vi}$-Blastn.NT$_{ABV}$)* resulted in a ~30X speedup with only slight decrease in precision and identical recall (Table 3). Also, the updated code performed better then Lazypipe2 and web CZID pipelines with fewer classification errors (Table 3).

Table 3 Accessing Lazypipe accuracy of virus taxon retrieval, execution time and memory consumption on simulated data from (Fosso et al., 2017). Compared pipelines are ordered by the descending F1-score. Column headers: TP, number of true positive species, FP, number of false positive species, FN, number of false negative species, F1, F1-score, Maxrss, maximum memory consumption.

[1]Other false negatives for Diamond.NR: Alphainfluenzavirus influenzae, Cytomegalovirus humanbeta5, Morbillivirus hominis, Orthoebolavirus zairense, Orthohepadnavirus hominoidei, Orthomarburgvirus marburgense, Phlebovirus napoliense, TTV-like mini virus and Uukuvirus uukuniemiense.
[2]Other false positives for Lazypipe2: Bacteriophage sp., Bifidobacterium phage Bbif-1, Caudoviricetes sp., Escherichia phage Lambda_ev058 and phi458, Herelleviridae sp., Inoviridae sp., Maltschvirus maltsch, Myoviridae sp. ctk251 and ctZhz2, RockefellervirusIPLA5, Siphoviridae sp. ctX581, Streptococcus phage IPP16, Streptococcus satellite phage Javan375 and uncultured human fecal virus.
[3]Other false positives for CZID: Biseptimavirus BU01, Caudoviricetes sp., Crassvirales sp., Myoviridae sp. ctk251/ctuJM17, Phietavirus B236, Podoviridae sp. ct7Ex2, Siphoviridae sp. ctCCX1/ctGz830/ctSqC25/ctX581/ctxvK3, uncultured human fecal virus and virus sp. ctJrn16.

									False Negative Predictions									False Positive Predictions													
Strategy	TP	FP	FN	Precision	Recall	F1	Execution Time	Maxrss (MB)	Betapapillomavirus 2	Betaretrovirus mastimon	Bocaparvovirus primate1	Human endogenous retrovirus K	Human endogenous retrovirus	Human endogenous retrovirus W	Mopeia Lassa virus reassortant 29	Orthobornavirus bornaense	Other False Negatives	AGM simian foamy virus	Betaretrovirus squmon	Borna disease virus	Human bocavirus	Human papillomavirus	Orthobunyavirus khatangaense	Orthopneumovirus bovis	Orthopoxvirus cowpox	Phlebovirus toscanaense	Simiispumavirus chlaet	Simiispumavirus pantrotro	Simplexvirus humanalpha1	Spleen focus-forming virus	unknown Hepeviridae	Other False Positives
Minimap.NT$_{Vi}$ - Blastn.NT	80	2	2	97.6 %	97.6 %	97.6 %	1:03:09:17	110,985							FN	FN		FP							FP							
Blastn.NT	80	2	2	97.6 %	97.6 %	97.6 %	4-21:17:37	25,857							FN	FN		FP							FP							
Minimap.NT$_{Vi}$ - Blastn.NT$_{ABV}$	79	2	3	97.5 %	96.3 %	96.9 %	3:53:04	20,827					FN	FN	FN									FP	FP							
Blastn.NT$_{ABV}$	79	2	3	97.5 %	96.3 %	96.9 %	1-08:28:43	25,713					FN	FN	FN									FP	FP							
Minimap.NT$_{Vi}$	79	2	3	97.5 %	96.3 %	96.9 %	1:08:28	12,166					FN	FN	FN									FP					FP			
Minimap.RefSeq$_{Vi}$ - Blastn.RefSeq$_{ABV}$	77	3	5	96.3 %	93.9 %	95.1 %	0:50:09	10,374		FN		FN	FN	FN	FN				FP								FP	FP				
Minimap.RefSeq$_{Vi}$	77	3	5	96.3 %	93.9 %	95.1 %	0:45:51	10,066		FN		FN	FN	FN	FN				FP								FP	FP				
Lazypipe2 --ann minimap	79	16	3	83.2 %	96.3 %	89.3 %	3:43:05	70,548					FN	FN	FN									FP								FP[2]
Diamondp.NR	66	5	15	93.0 %	81.5 %	86.8 %	1-07:50:35	33,550		FN	FN		FN	FN	FN	FN	FN[1]	FP	FP	FP								FP			FP	
CZID NT/NR	78	20	4	79.6 %	95.1 %	86.7 %	NA	NA	FN		FN		FN	FN								FP	FP	FP			FP		FP			FP[3]

3.3 Virus discovery from real data

3.3.1 Benchmarking against BLAST baseline

We compared identification of viral sequences from 90 mosquito samples (Truong Nguyen *et al.* 2022) by five chained strategies *Vi.chain1, Vi.chain2, Vi.chain3, Vi.chain3.env and Vi.chain4,* and Lazypipe2 --ann minimap using viral sequences identified by *Vi.chain0* as our "ground truth" baseline (for strategy definitions see section 2.2.3). This approach reflects limitations posed by real data, for which "true" viral sequences or taxa are unknown. The aim was not to measure absolute accuracy, but rather relative accuracy and relative time savings when comparing viral sequence identification to BLASTN/BLASTP annotations commonly applied in viral metagenomics. In this evaluation, true positives (TP) were

defined as viral sequences (>=500 bp) identified by both evaluated and the baseline strategy, False positives (FP) as viral sequences (>=500 bp) identified by the evaluated but not the baseline strategy, and False negatives (FN) as viral sequences (>=500 bp) identified by the baseline but not the evaluated strategy. Precision, recall and F1-metric were calculated from the corresponding TP, FP and FN basepair counts for the entire set of 90 samples with results presented in Fig 1 and Supp. Table 2.

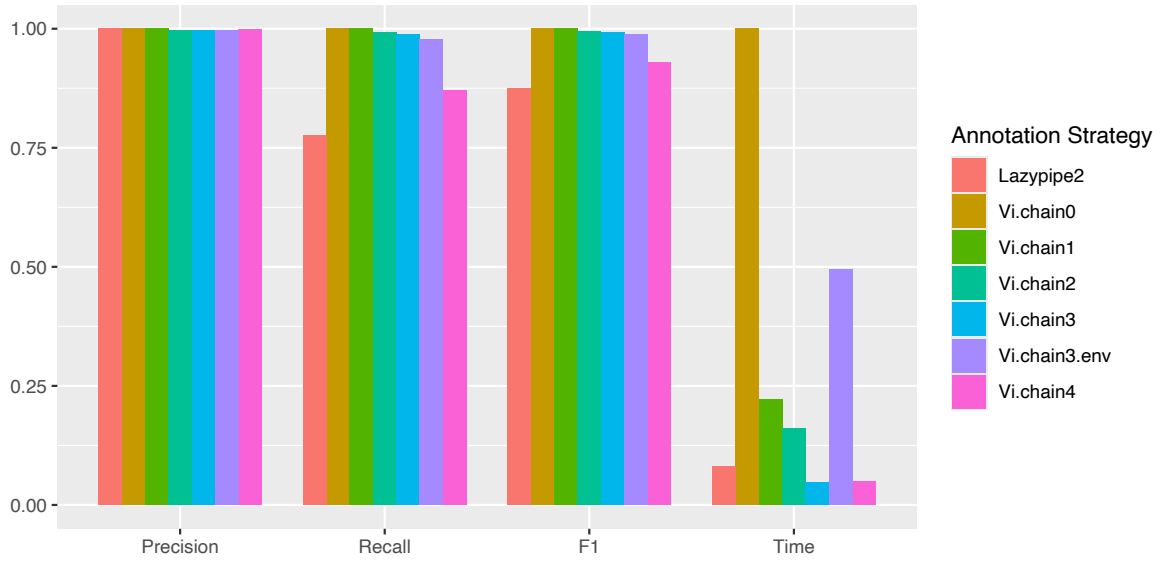


Figure 1 Overall performance of chained strategies (and Lazypipe2 –ann minimap) on the mosquito dataset. Precision, recall, F1-metric and Time were calculated relative to Vi.chain0 baseline strategy as described in the text. Precision, precision of identified viral base pairs, Recall, recall of viral base pairs, F1, F1-metric (harmonic mean of precision and recall), Time, execution time relative to Vi.chain0.

Supp. Table 2. Accessing Lazypipe accuracy of viral sequence identification with different annotation strategies on a dataset of 90 mosquito pools (Truong Nguyen et al., 2022). Ground truth baseline is defined here as viral sequences identified by Vi.chain0. TP, number of viral sequences (>=500 bp) identified by both evaluated and the baseline strategy, FP, number of viral sequences (>=500 bp) identified by the evaluated but not the baseline strategy, FN, number of viral sequences (>=500 bp) identified by the baseline but not the evaluated strategy, TP.bp/FP.bp/FN.bp, number of base pairs in TP/FP/FN sequences, F1, F1-metric, Time, execution time relative to Vi.chain0, Maxrss, maximum memory consumption relative to Vi.chain0.

Annotation strategy	Annotation settings	TP	FP	FN	TP.bp	FP.bp	FN.bp	Precision	Recall	F1	Time	Maxrss
Vi.chain0	—ann1 Blastn.NT_{ABV}, Blastp.$UniRef_{ABV}$	2,195	0	0	4,212,822	0	0	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %
Vi.chain1	—ann1 Minimap.NT_{ABV}, Blastn.NT_{ABV}, Blastp.$UniRef_{ABV}$	2,194	0	1	4,211,738	0	1,084	100.0 %	100.0 %	100.0 %	22.2 %	105.6 %
Vi.chain2	—ann1 Minimap.NT_{ABV}, Blastp.$UniRef_{ABV}$	2,154	3	41	4,181,975	11,386	30,847	99.7 %	99.3 %	99.5 %	16.0 %	98.9 %
Vi.chain3	—ann1 Minimap.NT_{ABV}, Diamondp.$UniRef_{ABV}$ very-sensitive	2,137	6	58	4,165,647	13,169	47,175	99.7 %	98.9 %	99.3 %	4.8 %	131.1 %
Vi.chain3.env	—ann1 Minimap.NT, Diamondp.UniRef —very-sensitive	2,102	5	93	4,121,336	12,659	91,486	99.7 %	97.8 %	98.8 %	49.5 %	417.4 %
Vi.chain4	—ann1 Minimap.NT_{ABV}, SANSparallel	1,832	13	363	3,663,913	8,629	548,909	99.8 %	87.0 %	92.9 %	5.0 %	63.1 %
Lazypipe2	—ann1 minimap	1626	0	569	3,268,076	0	944,746	1	77.6 %	87.4 %	8.0 %	63.9 %

Here we compare evaluated chains to the *Vi.chain0* baseline (BLASTN against NT_{ABV}, then contigs with no hits with BLASTP against $UniRef100_{ABV}$). In *Vi.chain1* we added Minimap2 before BLASTN, which resulted in ~1 kbp of FNs, no FPs and a ~5-fold speedup (Fig 1). In *Vi.chain2* we substituted BLASTN

488 to Minimap2, which resulted in ~11 kbp of FPs, ~31 kbp of FNs and ~6-fold speedup. In *Vi.chain3* we
489 substituted BLASTN with Minimap2 and BLASTP with DIAMOND blastp, which resulted in ~13 kbp of
490 FPs, ~47 kbp of FNs and ~20-fold speedup. In *Vi.chain3.env* we used the same searches as *Vi.chain3*
491 without prior background filtering and annotating against complete NT/UniRef100 databases. This had
492 no effect on FPs (~13 kbp) but, nearly doubled the number of false negatives (~91 kbp). We note that
493 it is debatable whether these are truly negative findings since here we compare against BLAST
494 searches on NT/UniRef100 that exclude eukaryotic sequences, thus FNs might represent contigs with
495 close homologs in eukaryotic sequences. Execution time for *Vi.chain3.env* was ~10X higher than for
496 *Vi.chain3* likely due annotating a larger assembly against larger databases. In *Vi.chain4* we substituted
497 BLASTN with Minimap2 and BLASTP with SANSparallel, which resulted in ~9 kbp of FPs, ~549 kbp of
498 FNs and ~20-fold speedup. Precision of all LazypipeX strategies was at least 99.7%. Recall was lowest
499 for strategy incorporating SANSparallel (87.0%), and above 97.8% for all other strategies. Recall by
500 Lazypipe2 was expectedly lower (77.6%) since this version only allowed annotations against either
501 nucleotide or protein database. For benchmarking we selected the nucleotide NT database leading to
502 poor recall of distant homologs.

503 The average runtime for mosquito samples (mean 74 Mbp, median 46 Mbp, range 0.2-338 Mbp) ranged
504 from ~5 hours (*Vi.chain1*) to ~1 hour (*Vi.chain3* and *Vi.chain4*) illustrating time savings of up to 19 out
505 of 20 hours compared to the BLAST baseline.

506 *3.3.2 Identified viral sequences*

507 Here we characterise viral sequences identified by LazypipeX *Vi.chain1* (Minimap2, BLASTN, BLASTP)
508 from the mosquito dataset and compare these to the previously published findings (Truong Nguyen *et*
509 *al.* 2022). We also examine a novel rhabdovirus discovered from the horse fly sample with *Vi.chain1*
510 augmented with a HMMscan search. Finally, we compare reported rhabdo-, nege-, toti- and flavivirus
511 sequences reported by *Vi.chain1* and CZ ID pipelines for the horse fly sample and two mosquito
512 samples (acc. SRR19859382-83). All viral contigs reported by *Vi.chain1* are listed in Supp. Table 3
513 and CZ ID reports in Supp. File 2.

514 For the mosquito dataset, *Vi.chain1* identified 1,242 viral contigs exceeding 1 kbp and covering a total
515 of 3,559 kbp. From these Minimap2 identified 78.1 %, BLASTN 1.6% and BLASTP 20.3%. Keeping in
516 mind application of chained annotations for virus discovery we next focus on >=1 kbp viral contigs
517 identified by Minimap2 and BLASTP. We expected that Minimap2 would identify viral sequences
518 already published for this dataset, while BLASTP would identify divergent and yet unpublished viruses.

519 *3.3.2.1 Identified rhabdoviruses*

520 Minimap2 in *Vi.chain1* identified 44 contigs (>=1kbp) covering *Enontekio merhavirus, E. ohlsrhavirus,*
521 *Hattula rhabdovirus, Inari rhabdovirus, Joutseno rhabdovirus 1 and 2, and Ohldorf virus* (see Supp.
522 Table 3). These include all rhabdoviruses previously published for this dataset (Truong Nguyen *et al.*
523 2022). BLASTP in *Vi.chain1* identified two contigs covering complete rhabdovirus genomes that were

provisionally named here according to municipalities they were discovered from as *Kuusamo primrhavirus* and *Inari merhavirus* (Fig 2). Comparing these against rhabdoviruses published in previous analysis (Truong Nguyen *et al.* 2022) confirmed very low nucleotide level homology (qcov <9%) for both sequences.

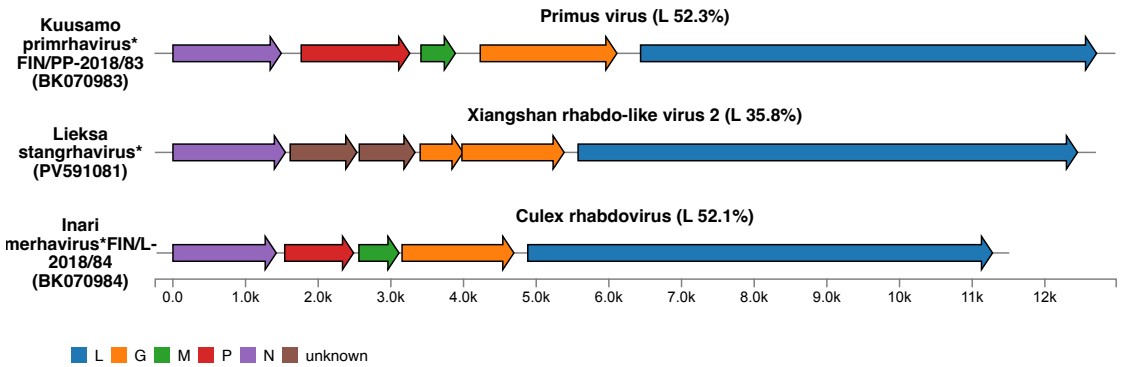


*Figure 2 ORF structure of novel Rhabdo-like sequences from mosquito samples and the horse-fly sample. Labels at the center represent closest homologs in UniRef with identity values in parenthesis. *Provisional virus names, isolates and accession numbers.*

The *Kuusamo primrhavirus* (identified in acc. SRR19859383) and *Inari merhavirus* (identified in acc. SRR19859382) both had five ORFs corresponding to N, P, M, G and L genes in a canonical rhabdovirus genome (Fig 2). In *Kuusamo primrhavirus* 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 acc. A0A6H0C864_9VIRU, A0A6H0C6Q8_9VIRU, A0A6H0C7L9_9VIRU and A0A6H0CA79_9VIRU), while ORF3 had no homologs in UniRef, although by genomic position this appears a putative M gene. In *Inari merhavirus* ORF1 had 35.8% identity to *Merida virus* nucleoprotein (A0A894KPN5_9RHAB), ORF2 had no homologs (a putative phosphoprotein), 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 merhavirus* had no significant nucleotide similarity to *Inari rhabdovirus* (acc. ON955143) previously identified from the same pooled sample.

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

553 Bootstrapped maximum likelihood (ML) trees were constructed from the L-gene for *Kuusamo*
554 *primrhavirus*, *Lieksa stangrhavirus* and *Inari merhavirus* as described in the methods section (Fig 3 and
555 4). In the phylogenetic tree (Fig 3), *Stangrhavirus lieksa* (acc. PV591081) was placed as a sister clade
556 to a clade with five ICTV *Stangrhavirus* isolates, while *Primrhavirus kuusamo* (acc. BK070983) was
557 placed within a monophyletic clade of ICTV *Primrhavirus* isolates, confirming the assignment of these
558 sequences to *Strangrhavirus* and *Primrhavirus* genera. *Inari merhavirus* (acc BK070984) was placed
559 as an outgroup to the clade of Merida viruses, and within the clade formed by other merhaviruses (Fig
560 4), confirming its assignment to the *Merhavirus* genus.

561 Comparing *K. primhavirus* (acc. BK070983) protein sequences to ICTV *Primrhavirus* isolates (acc.
562 MN661034, BK059423, MN567480, OL700129) showed above 66%, 46% and 63% aa divergence for
563 the N, L and G genes, respectively, suggesting this is a new *Primrhavirus*. Unlike ICTV primrhaviruses,
564 which were discovered in metagenomic studies of *Culex and Aedes* mosquitos (Walker *et al.* 2022), *K.*
565 *primrhavirus* was sequenced from *Ochlerotatus communis*, further supporting demarcation into a
566 separate species. Comparing *L. stangrhavirus* (acc. PV591081) protein sequences to ICTV
567 *Stangrhavirus* isolates (acc. MW434768, MN013386, MW434775, OL700091, OL700136) showed
568 above 65%, 46% and 63% aa divergence for the N, L and G genes, respectively, suggesting this is a
569 new *Stangrhavirus*. Additionally, the G gene in the new genome is split into two ORFs (Fig 2), and the
570 host species is a horse-fly, while ICTV stangrhaviruses are hosted by mosquitoes.

571 Comparing *Inari merhavirus* (acc. BK070984) protein sequences to ICTV *Merhavirus* isolates (acc.
572 OR479699, BK059424, ON955142, ON955143, KU194360, LC775065, AB604791) showed above
573 67%, 48% and 52% amino acid (aa) divergence for the N, L and G genes, suggesting this is a new
574 *Merhavirus*. The aa divergence was high (>70%, >60% and >70%, for N, L and G, respectively) also to
575 the *Merhavirus inari* (acc. ON955143) which was previously identified from the same sample. Thus, in
576 the pooled sample individual mosquitoes may have been infected or coinfected with the two different
577 merhaviruses.

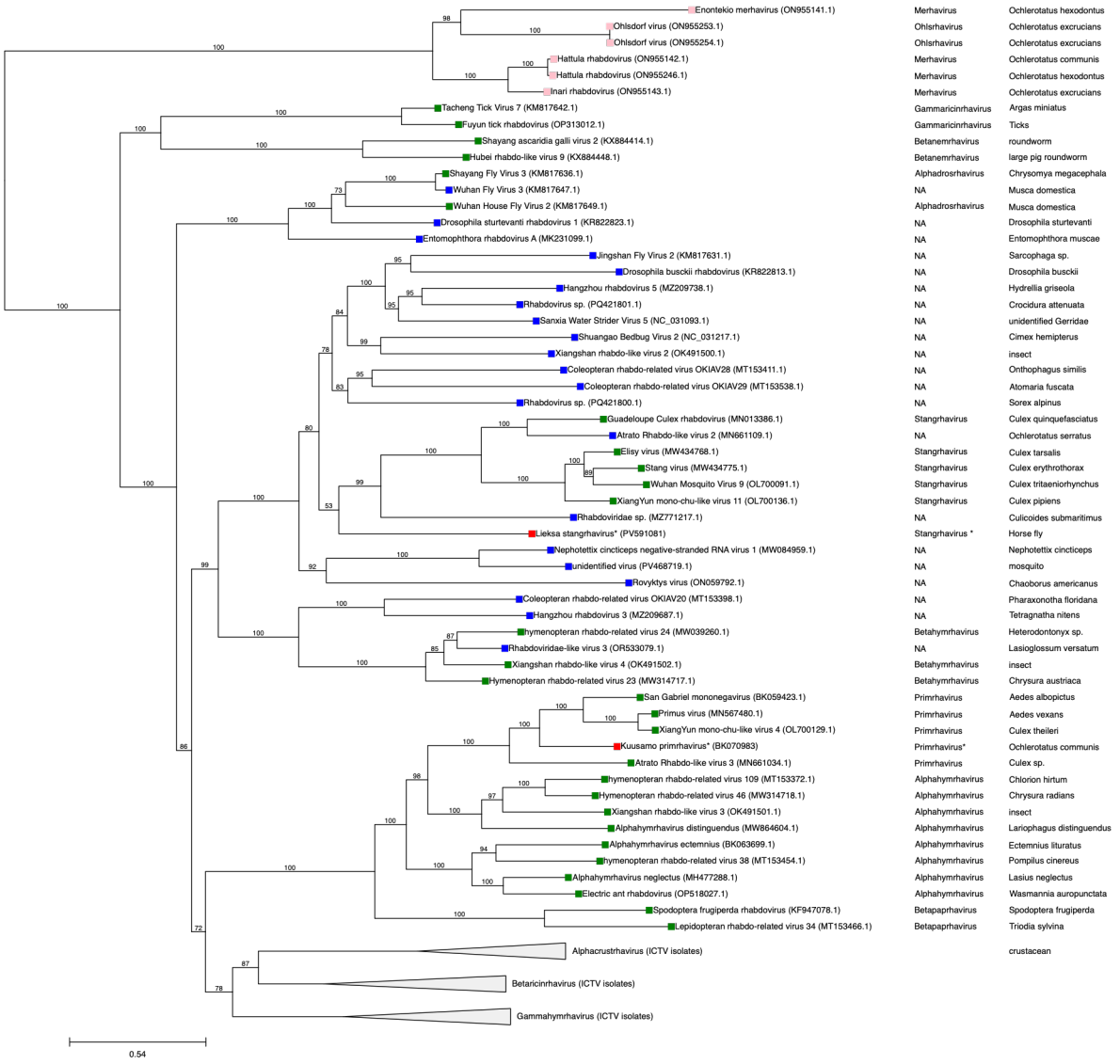


Figure 3 Phylogeny of Lieksa stangrhavirus and Kuusamo primrhavirus*. Tip colours: red, novel sequences, pink, Rhabdoviridae previously published for the mosquito dataset (Truong Nguyen* et al. *2022), green, ICTV Deltarhabdovirinae isolates, blue, homologs retrieved with BLASTX. Columns on the right denote virus genus and host species. Rooted on outgroup (infectious hematopoietic necrosis virus, acc. L40883, not shown), divergent clades collapsed for clarity. *Names assigned to new sequences are provisional.*

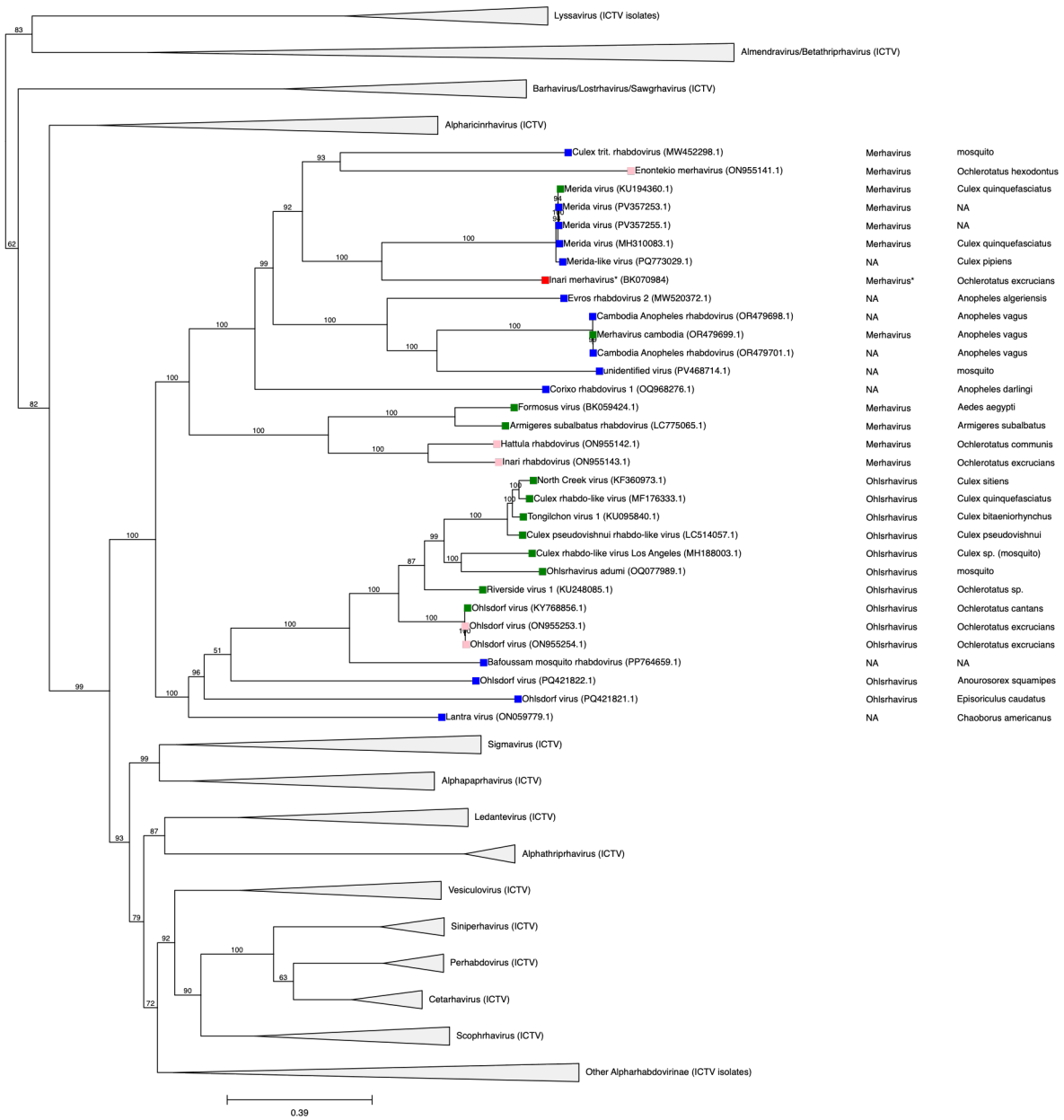
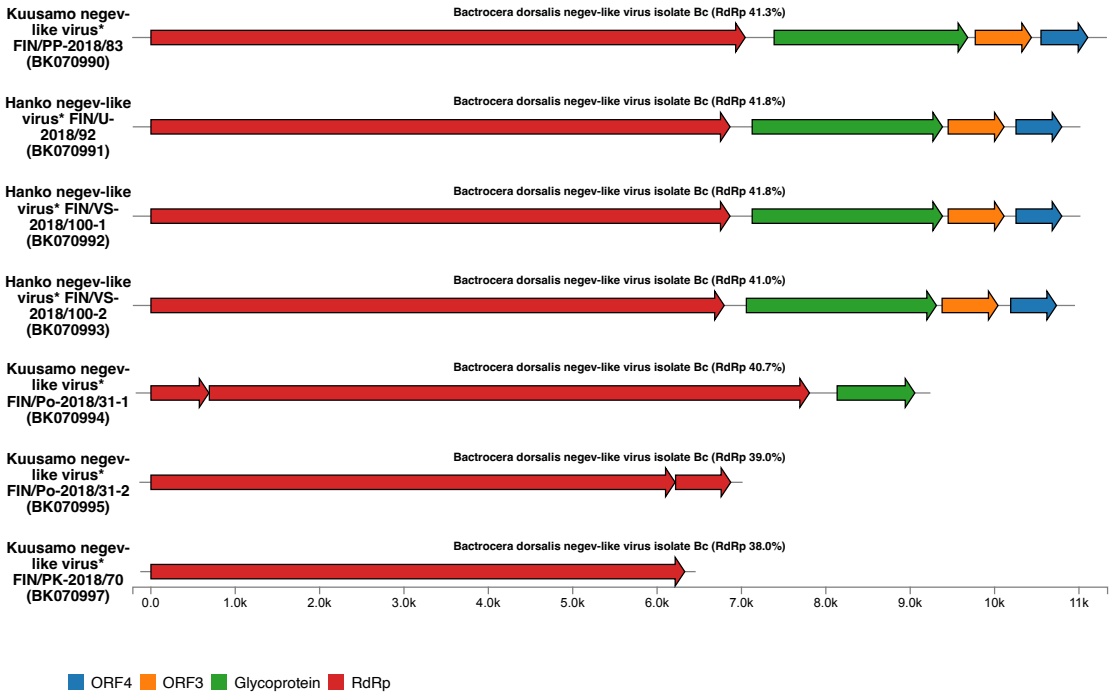


Figure 4 Phylogeny of Inari merhavirus. Tip colours: red, novel sequence, pink, rhabdoviruses previously identified from the mosquito dataset (Truong Nguyen* et al. *2022), green, ICTV Alpharhabdovirinae isolates, blue, homologs retrieved with BLASTX. Columns on the right denote virus genus and host species. Rooted on the outgroup (infectious hematopoietic necrosis virus, acc. L40883, not shown), divergent clades collapsed for clarity. *Names assigned to new sequences are provisional.*

3.3.2.2 Identified negeviruses

Minimap2 in *Vi.chain1* identified 56 contigs (<=1kbp) including *Kustavi negevirus, Mekrijärvi negevirus, Utsojoki negeviruses 1-3, Dezigoudou virus and Corboda virus* (Supp. Table 3), which covers all Negeviruses previously published for this dataset (Truong Nguyen *et al.* 2022). BLASTP in *Vi.chain1* identified four whole-genome negevirus-like sequences ranging between 11.1-11.4 kbp and 21 shorter sequences. Comparing these BLASTP findings against previously published Negeviruses (Truong Nguyen *et al.* 2022) confirmed very low nucleotide level homology (qcov <=3%). ORF structure for the

597 seven largest BLASTP findings is illustrated in Fig 5. These were provisionally named *Kuusamo* and
598 *Hanko negev-like viruses*, represented by four and three isolates, respectively (Fig 5).



599

600 *Figure 5 ORF structure of novel Negev-like virus sequences from mosquito samples. Labels at the center represent*
601 *closest homologs in UniRef with identity values in parenthesis. *Provisional virus names, isolates and accession*
602 *numbers.*

603 Bootstrapped ML tree was constructed with the RdRp-gene for the seven largest Negev-like virus
604 sequences as described in the methods section (Fig 6). In the phylogenetic tree novel sequences were
605 placed into two sister clades with isolates in the first claded hosted by *Ochlerotatus caspius*, and in the
606 second claded by *O. communis*. Both clades were placed at a distance of known members of the
607 proposed genus *Negevirus* and closest to negev-like viruses identified from *Bactrocera, Zeugodacus,*
608 *Ceratitis* and other *Diptera* (Fig 6). Comparing ORF1 protein sequence (concatenated
609 Methyltransferase, RNA-helicase and RdRp) between the *Kuusamo negev-like virus* (acc. BK070997),
610 *Hanko negev-like virus* (acc. BK070992) and exemplar isolates of the proposed genus *Negevirus* (as
611 listed in (Nunes *et al.* 2017)) showed below 7% aa divergence in within clade pairwise comparisons,
612 above 25% aa divergence in between clade comparisons and above 76% divergence to exemplar
613 *Negevirus* isolates (Supp. Table 4). These divergence levels (<7% within clades and >25% between
614 clades) suggest that the seven isolates identified here likely represent two novel negeviruses hosted by
615 *Ochlerotatus* mosquitoes.

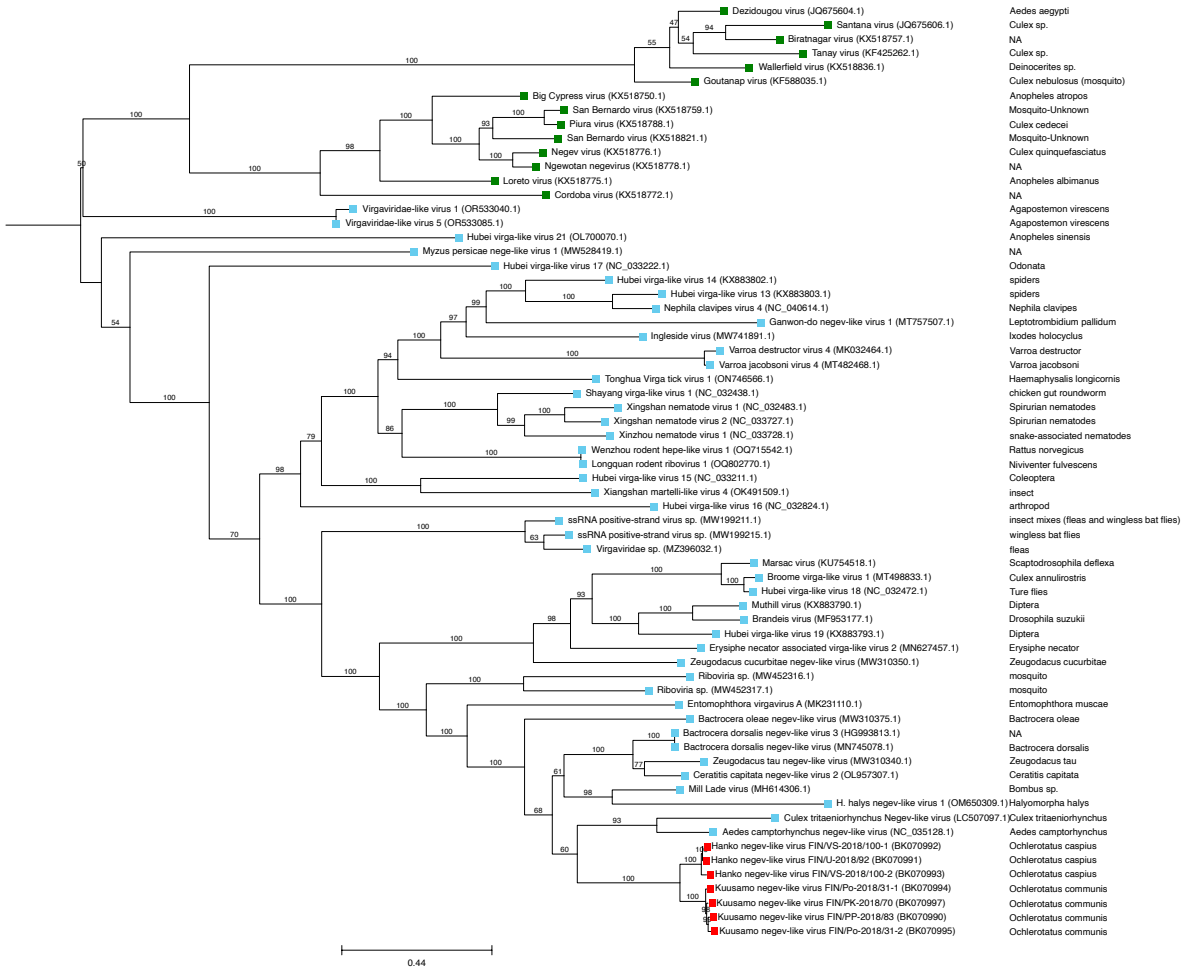


Figure 6 Phylogeny for novel negev-like virus sequences from mosquito samples. Tip colours: novel sequences in red, BLASTX homologs in blue, exemplar isolates of the proposed genus Negevirus (as listed in Nunes et al.*, 2017) in green. Host species in the right column. Scale bar is in Mya. Virus names assigned to new sequences are provisional.*

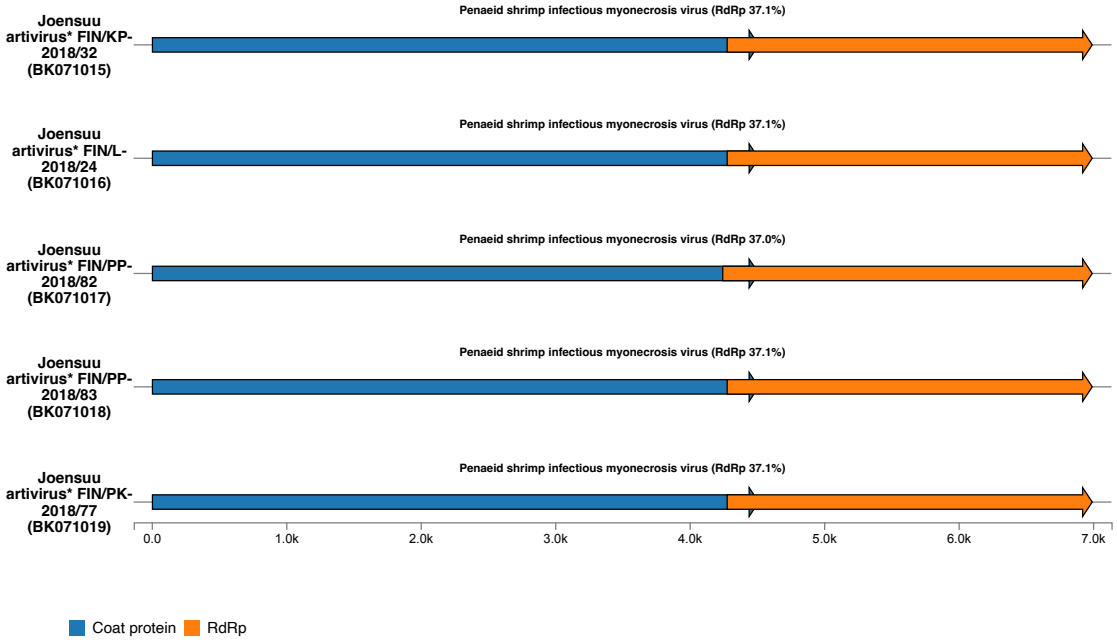
3.3.2.3 Identified totiviruses and artiviruses

The first search in *Vi.chain1*, Minimap2, identified 315 contigs (>=1kbp) covering 49 out of 52 *Totivirdae* species previously published for this dataset (Truong Nguyen *et al.* 2022) (Supp. Table 2). This search did not report *Hanko totivirus 10 (2.7 kbp), Ilomantsi totivirus 3* (1.2 kbp) and *Karstula totivirus (1.1 kbp)* published previously perhapse due to close similarities to other totivirus-like sequences.

BLASTP identified five whole-genome totivirus-like whole genome contigs ranging from 7.5 to 7.7 kbp and provisionally named as *Joensuu artivirus* (Fig 7). Comparison against previously published *Totivirus* sequences (Truong Nguyen *et al.* 2022) showed very low nucleotide level homology (qcov<14%) for all of these contigs.

Bootstrapped ML tree was constructed with concatenated ORF1 and ORF2 aa sequences. *Joensuu artivirus* formed a monophyletic sister clade to an unclassified *Grogonang virus 29* (OR270267) (Fig 8). The two closest classified viruses were *Artivirus sani* (acc. EU715328) and *Artivirus ni* (acc. GQ342961) confirming assignment to *Artivirus* (Fig 8). *Joensuu artivirus FIN/PK-2018/77* (BK071019)

634 was selected as the representative isolate and its RdRp aa sequence was compared to other members
635 of the clade and established member species of *Artivirus* (acc. KX882995, KX882989, AY570982,
636 MT293125, GQ342961, KX882972, EU715328, AB555544, KX882992). The representative isolate had
637 <40% aa identity to established *Artivirus* species and >99% aa identity within the *Joensuu artivirus*
638 sequences suggesting these represented isolates of a novel *Artivirus*.



639

640 *Figure 7 ORF structure of novel Artivirus sequences from mosquito samples. Labels at the center represent closest*
641 *homologs in UniRef with identity values in parenthesis. *Provisional virus names, isolates and accession numbers.*

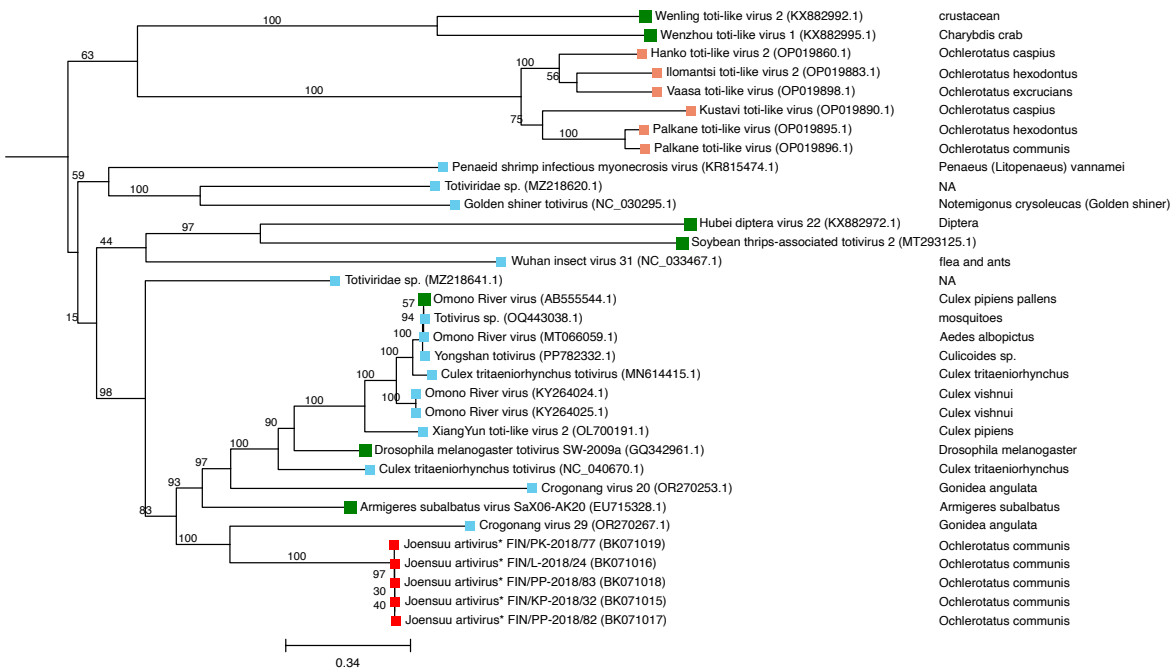


*Figure 8. Phylogeny of novel Artivirus sequences from mosquito samples. Tip colours: novel sequences in red, Totiviridae sequences previously published for this data in light red, BLASTX homologs in blue, ICTV Artivirus exemplars in green. Host species in denoted in the right column. Scale bar is in Mya. *Provisional virus names, isolates and accession numbers.*

3.3.2.4 Identified flaviviruses

The first search in *Vi.chain1*, Minimap2, identified 10 contigs (>=1kbp) covering all *Flaviviridae* species previously published for this dataset (Truong Nguyen *et al.* 2022) (Supp. Table 2). BLASTP in *Vi.chain1* identified six contigs mapped to *Flaviviridae*. Comparison against previously published *Flaviviridae* showed 72.0% nt identity between sequence provisionally named as *Hameenlinna flavivirus (acc. BK071037)* and *Hanko virus* (acc. ON949929.1), while other sequences were non-homologous. All six contigs had a long ORFs complete on the 5' end and partial on 3' end with flavivirus polyprotein homologs in UniRef (Figure 9).

Bootstrapped ML tree was constructed for the six BLASTP *Flaviviridae* sequences (>=1kbp) as described in the methods section (Fig 10). *Hameenlinna flavivirus FIN/KH-2018/38-2* was placed as a new isolate of previously published *Hameenlinna flavivirus FIN/KH-2018/38* (acc. ON955057). *Palkane flavi-like virus* (acc. BK071041) was placed as an outgroup to *Placeda* (acc. MW434111) and *Culex tritaeniorhynchus flavi-like* (acc. LC514290) viruses. Four novel sequences, including two isolates of *Lestijarvi flavivirus* (acc. BK071038 and BK071040) hosted by *Ochlerotatus diantaeus* and two isolates of *Palkane flavivirus* (acc. BK071036 and BK071039) hosted by *O. communis*, formed a monophyletic group very distant from previously published *Flaviviridae* and BLASTX homologs (Figure 10). Due to partial coverage of the polyprotein the *Lestijarvi* and *Palkane* sequences were not characterised in details, however the performed preliminary phylogenetic analysis indicates that these may represent isolates of novel flaviviruses in *O. diantaeus* and *O. communis* mosquitoes from Finland. Whether this

is indeed so remains to be shown by future studies. We note that *Palkane flavi-like virus*, *Lestijarvi flavivirus* and *Palkane flavivirus* are provisional names, which may be subject to change.

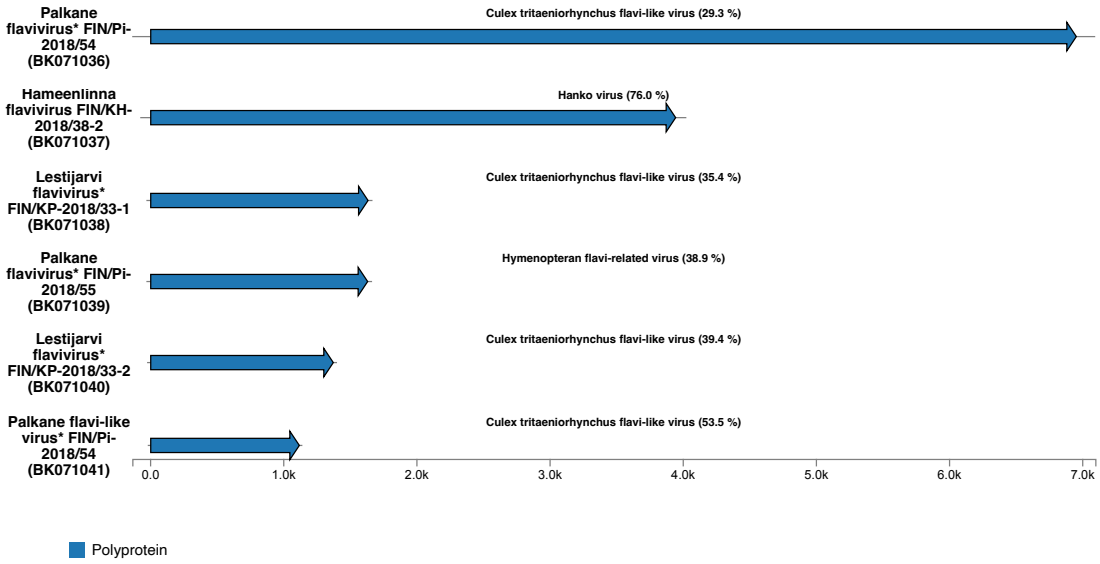


*Figure 9 ORF structure of novel Flavi-like sequences from mosquito samples. Labels at the center represent closest homologs in UniRef with identity values in parenthesis. *Provisional virus names, isolates and accession numbers.*

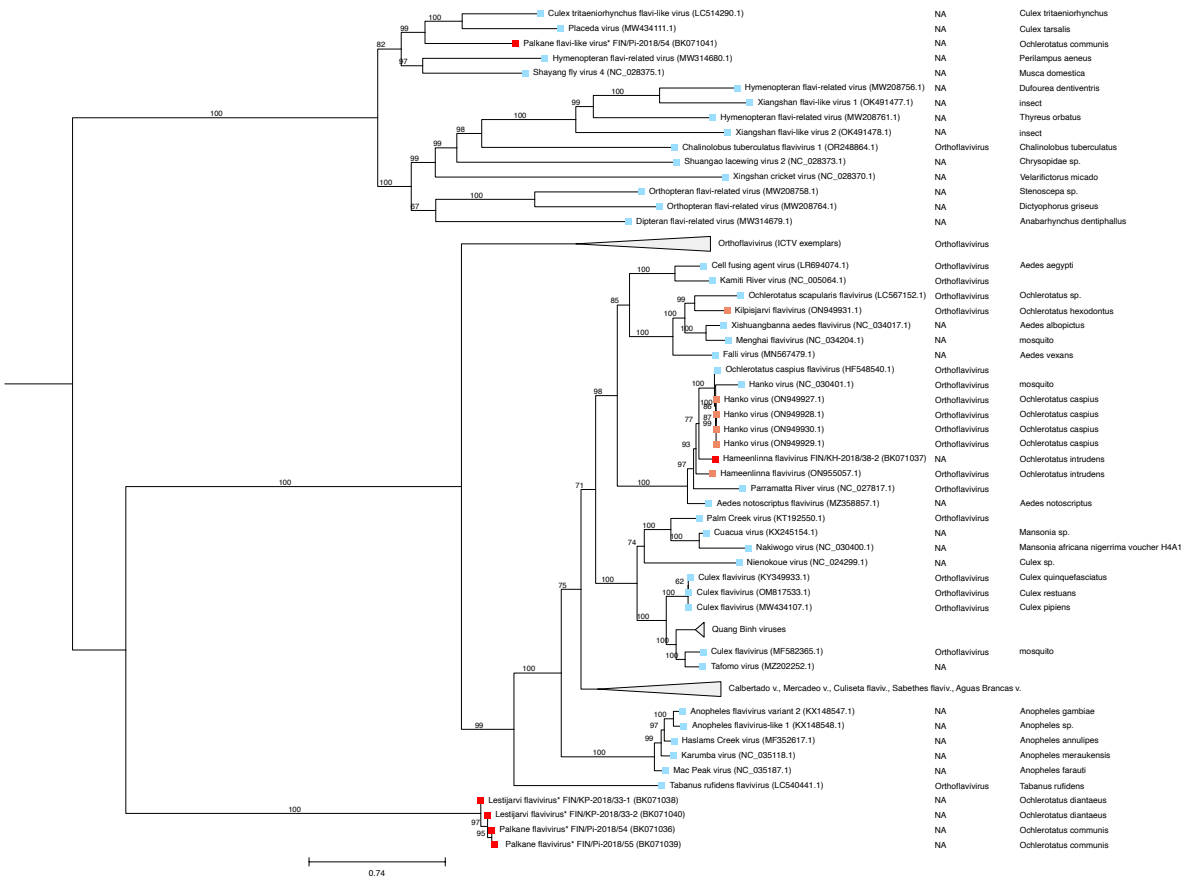


*Figure 10 Phylogeny of novel Flavi-like sequences from mosquito samples. Tip colours: novel sequences in red, homologs previously published for this data in light red, BLASTX homologs in blue, ICTV Orthoflavivirus exemplars in a collapsed clade in the upper part of the tree. Columns on the right denote genus and host species. Scale bar is in Mya. *Provisional virus names, isolates and accession numbers.*

3.3.3 Comparison of virus discovery by CZ ID and Lazypipe

Comparing CZ ID and LazypipeX findings for *Rhabdoviridae* from the horse fly and the two mosquito samples (acc. SRR19859382-83) showed that CZ ID was able to identify complete genome for *Inary merhavirus*, but only the L-gene for *Lieksa stangrhavirus* and 138-606 bp fragments of *Kuusamo primrhavirus* reported by LazypipeX (Supp. File 2). For negeviruses CZ ID was able to recover 0.1-1.1 kbp fragments, but not the full genome of *Kuusamo negev-like virus FIN/PP-2018/83* reported by LazypipeX (Supp. File 2). For totiviruses CZ ID did not report divergent totivirus contigs reported by LazypipeX for acc. SRR19859382 and also did not report the full genome *Joensuu artivirus FIN/PP-2018/83* reported by LazypipeX for acc. SRR19859383. For samples analysed with CZ ID LazypipeX and CZ ID did not report *Flaviviridae* contigs. There were also no large (>= 1kbp) novel rhabdo-, nege-, arti- or totivirus sequences reported by CZ ID that would not be reported by LazypipeX (Supp. File 2).

DISCUSSION

LazypipeX is a versatile and efficient pipeline for virus discovery and identification from metagenomic next-generation sequencing (mNGS) data, offering significant advancements in speed, accuracy, and flexibility.

692 LazypipeX's background filtering enables transparent removal of host and contamination sequences by
693 specifying database accessions and metadata in the configuration file. This approach achieved a 7-fold
694 speedup in a simulated benchmark with heavy host sequence backgrounds (Table 3), driven by reduced
695 data size and a smaller reference database for annotation. Benchmarking on real and synthetic datasets
696 demonstrated high filtering accuracy (Table 2 and Supp Table 1). The pipeline supports pre-built filters
697 for common use cases, custom filter creation, and simultaneous application of multiple filters. These
698 capabilities make LazypipeX adaptable to diverse scenarios, such as clinical samples from human or
699 animal patients, vector pools (e.g., mosquitoes and ticks), and wildlife specimens with unknown
700 species-level identification.

701 Similar to its predecessors LazypipeX employs *de novo* assembling to generate contigs, enhancing the
702 sensitivity and specificity of homology searches. This approach is shared by pipelines that classify
703 assemblies instead or in addition to classifying raw reads (Morfopoulou and Plagnol 2015; Scheuch,
704 Höper and Beer 2015; Alawi *et al.* 2019; Kalantar *et al.* 2020; Zwagemaker *et al.* 2021). LazypipeX
705 integrates both nucleotide and protein alignments, a feature rare among mNGS pipelines. A recent
706 comparison found that only two of 13 pipelines used both alignment types (Alawi et al., 2019; Vilsker et
707 al., 2019), and only DAMIAN incorporated Hidden Markov Models (de Vries et al., 2021). LazypipeX
708 users can tailor the annotation procedure by chaining homology searches (e.g., Minimap2, BLASTN,
709 BLASTP, BLASTX, Diamond, SANSparallel, HMMscan) against different databases, optimizing for
710 sensitivity or specificity depending on the dataset. This flexibility supports iterative adjustment of the
711 pipeline to match a variety of goals in viral metagenomics.

712 The two-round annotation strategy in LazypipeX significantly enhances virus identification efficiency.
713 An initial fast search against a virus-only database, followed by a second round using BLASTN against
714 a larger database, achieved an 8-fold speedup without compromising accuracy (Table 3). When
715 combined with eukaryotic background filtering, this approach yielded a 30-fold speedup compared to
716 single-round BLASTN against the complete NT database. Compared to Lazypipe2, DIAMOND blastp
717 on NR, and CZ ID, LazypipeX's two-round strategies (*Minimap.NT*$_{Vi}$*-Blastn.NT* and *Minimap.NT*$_{Vi}$-
718 *Blastn.NT*$_{ABV}$) demonstrated superior accuracy in identifying eukaryotic viral taxa (Table 3). For
719 example, DIAMOND blastp missed 15 viral pathogens, likely due to high protein-level similarity among
720 related viruses, while CZ ID reported multiple false positives. These results highlight the strength of
721 LazypipeX's nucleotide-based searches for detecting known viruses.

722 LazypipeX two-round strategy is not unique it's attempts to perform faster alignments prior to the more
723 accurate but slower alignments. The idea was inspired by VirusSeeker pipeline that starts by fast
724 identification of candidate viral sequences against virus-only nucleotide and protein databases and then
725 removes false positives with alignments against NCBI NT and NR databases (Zhao *et al.* 2017). Also,
726 in CZ ID reads are first mapped with fast nucleotide and protein aligners prior to aligning contigs with
727 BLAST against targeted databases (Kalantar *et al.* 2020). In GenomeDetective viral reads are identified
728 with fast DIAMOND search on UniRef90 prior to assembling and running BLASTN/BLASTX on RefSeq.

729 Lazypipe complements these approaches by offering flexibility to choose different search engines and
730 databases for both scanning and refinement stages.

731 For virus discovery, LazypipeX's chained annotation strategies proved highly effective. In a real dataset
732 of 90 mosquito pooled samples, the *Vi.chain1* strategy (Minimap2/BLASTN/BLASTP) reduced
733 execution time 5-fold compared to BLASTN/BLASTP baseline, with precision and recall above 99.7%
734 and 98.9%, respectively. Substituting BLASTP with faster tools like DIAMOND or SANSparallel
735 (*Vi.chain3, Vi.chain4*) achieved a 20-fold speedup, though *Vi.chain4* showed lower recall (87.0%).
736 *Vi.chain1* retrieved nearly all previously reported viral sequences (Truong Nguyen et al., 2022) via its
737 initial Minimap2 search, while subsequent BLASTP search identified novel, distantly related sequences.
738 This also explains why Lazypipe2 had low recall (78.6%): this version only allowed application of one
739 search in each analysis, thus, failing to report distant homologs. Incorporating HMMscan into *Vi.chain1*
740 enabled the discovery of a novel rhabdovirus whole-genome sequence in a horse fly sample, which CZ
741 ID only partially recovered (L gene only). Comparisons with CZ ID (Kalantar *et al.* 2020) for rhabdo-,
742 nege-, arti/toti-, and flaviviruses further demonstrated LazypipeX's superior recovery of novel genomes
743 (Supp. File 2). Findings by BLASTP in Vi.chain1 were characterized with ORF, homology and
744 phylogenetic analysis, which confirmed that most of these findings were unrelated to viral sequences
745 previously reported for this dataset and covering complete genomes for at least six novel viruses. These
746 results underscoring LazypipeX utility for detecting divergent viruses in complex mNGS datasets.

747 Based on our benchmarking we can suggest general guidelines for selecting Lazypipe strategies for
748 different use cases (Table 4). For virus identification we advocate using two-round strategies either
749 against NT_{ABV} for tissue samples or against complete NT for samples with potential retroviral sequences
750 and/or samples with complex eukaryotic backgrounds (Table 4). For virus discovery from tissue
751 samples, we advocate using strategies *Vi.chain1* or *Vi.chain3*. Vi.chain3 is particularly well suited for
752 large datasets, since this showed a 20-fold speedup relative to BLASTN/BLASTP with a very small
753 reduction in precision and recall (Supp. Table 2). *Vi.chain1* provided a 5-fold speedup with nearly no
754 reduction in precision/recall, thus this is a viable option for analysing smaller datasets or running
755 analysis of larger HPCs. For virus discovery from environment, faecal and other samples with complex
756 eukaryotic background we advocate using *Vi.chain3.env* as this will also bin eukaryotic sequences.

757 *Table 4. Recommended annotation strategies for different user cases*

Sequence Data	Analysis Goals	Recommended Annotation Strategy
Illumina libraries from tissue samples and other samples with clearly defined eukaryotic background	Virus identification excluding retroviruses. BLASTN accuracy with up to 30X speedup compared to BLASTN on complete NT.	*background filter + Minimap.NT$_{Vi}$ - Blastn.NT$_{ABV}$*
	Virus identification including retroviruses. BLASTN accuracy with up to 8X speedup compared to BLASTN.	*Minimap.NT$_{Vi}$ - Blastn.NT*
	Virus discovery excluding retroviruses. Precision/recall at ~100%/100% of combined BLASTN/BLASTP with up to 5X speedup.	*background filter + Vi.chain1*
	Virus discovery excluding retroviruses. Precision/recall at ~100%/99% of combined BLASTN and BLASTP with up to 20X speedup.	*background filter + Vi.chain3*
Illumina libraries from environmental, fecal and other samples with complex eukaryotic backgrounds	Virus identification including retroviruses. BLASTN accuracy with up to 8X speedup compared to BLASTN.	*Minimap.NT$_{Vi}$ - Blastn.NT*
	Virus discovery including retroviruses. Precision/recall at ~100%/98% of combined BLASTN and BLASTP with up to 20X speedup.	*Vi.chain3.env*

758

759

760 **Future perspectives**

761 We plan to extend Lazypipe for use in field settings, particularly for integration with the Oxford Nanopore
762 MinION platform. This version will be optimized for both offline and online modes, enabling real-time
763 analysis of mNGS data in remote locations.

764 ## 4 Author Contributions

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

770 ## 5 Acknowledgements

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

772 ## 6 Conflicts of Interest

773 The authors have no conflict of interest to declare.

7 Data Availability Statement

LazypipeX 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 5).

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 Tables 1, 2, 3, 4 and 5

Supplementary Files 1 and 2.

10 References

Alawi M, Burkhardt L, Indenbirken D *et al.* DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples. *Sci Rep* 2019;**9**, DOI: 10.1038/s41598-019-52881-4.

Alburkat H, Smura T, Bouilloud M *et al.* 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* 2024;**18**:e0012142.

Alfaro-Alarcón A, Hetzel U, Smura T *et al.* Boid Inclusion Body Disease Is Also a Disease of Wild Boa Constrictors. *Microbiol Spectr* 2022;**10**:e01705-22.

Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.

Altschul SF, Madden TL, Schäffer AA *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.

Bassi C, Guerriero P, Pierantoni M *et al.* Novel Virus Identification through Metagenomics: A Systematic Review. *Life* 2022;**12**:2048.

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

Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021;**18**:366–8.

Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60.

Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;**25**:1972–3.

Eddy SR. Accelerated Profile HMM Searches. *PLOS Comput Biol* 2011;**7**:e1002195.

Erdin M, Altan E, Suomalainen N *et al.* Genetic characterization of rat hepatitis E virus (Rocahepevirus ratti) in urban brown rats (Rattus norvegicus) in Helsinki, Finland. *Arch Virol* 2025;**170**:215.

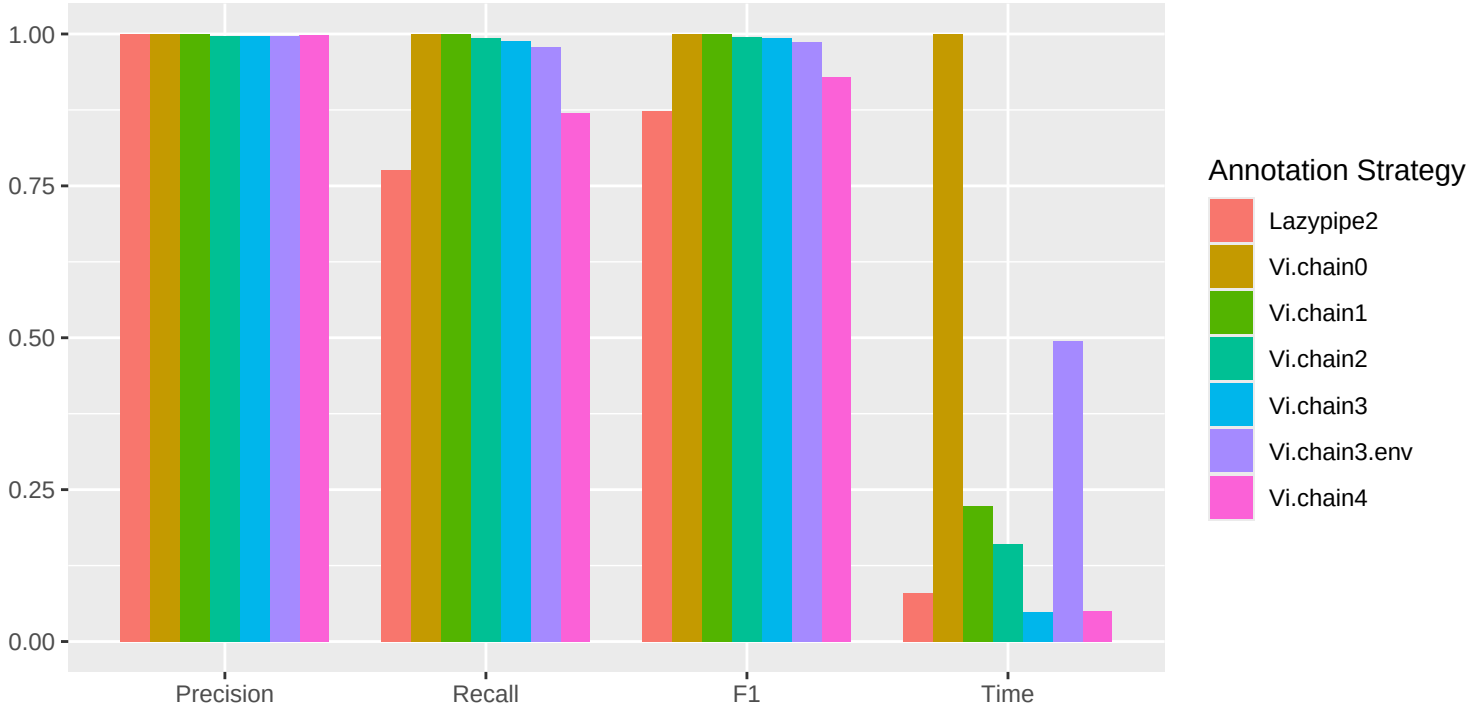
Erdin M, Smura T, Kalkan KK *et al.* Detection of divergent Orthohantavirus tulaense provides insight into wide host range and viral evolutionary patterns. *Npj Viruses* 2024;**2**:1–13.

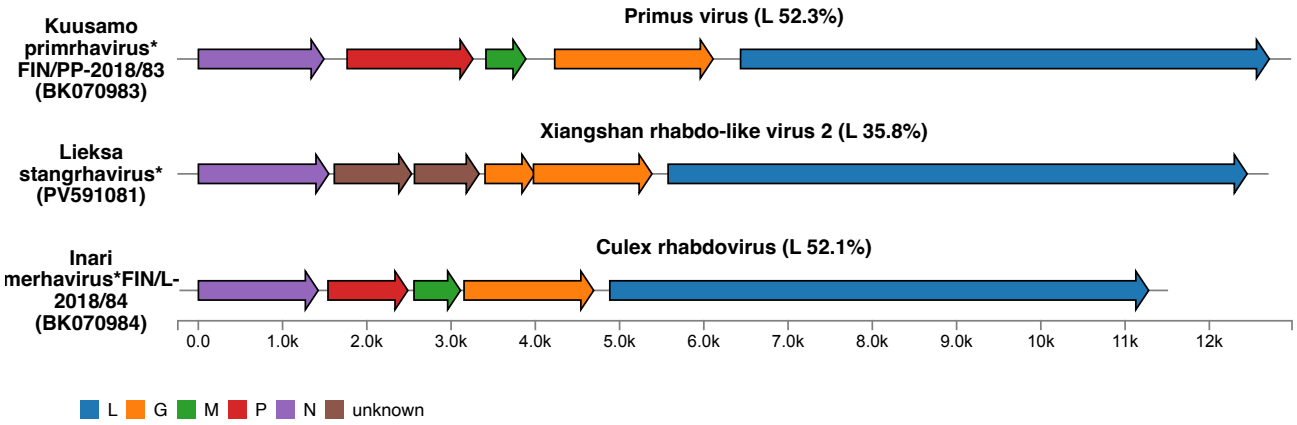
Forbes KM, Webala PW, Jääskeläinen AJ *et al.* Bombali Virus in Mops condylurus Bat, Kenya. *Emerg Infect Dis* 2019;**25**:955–7.

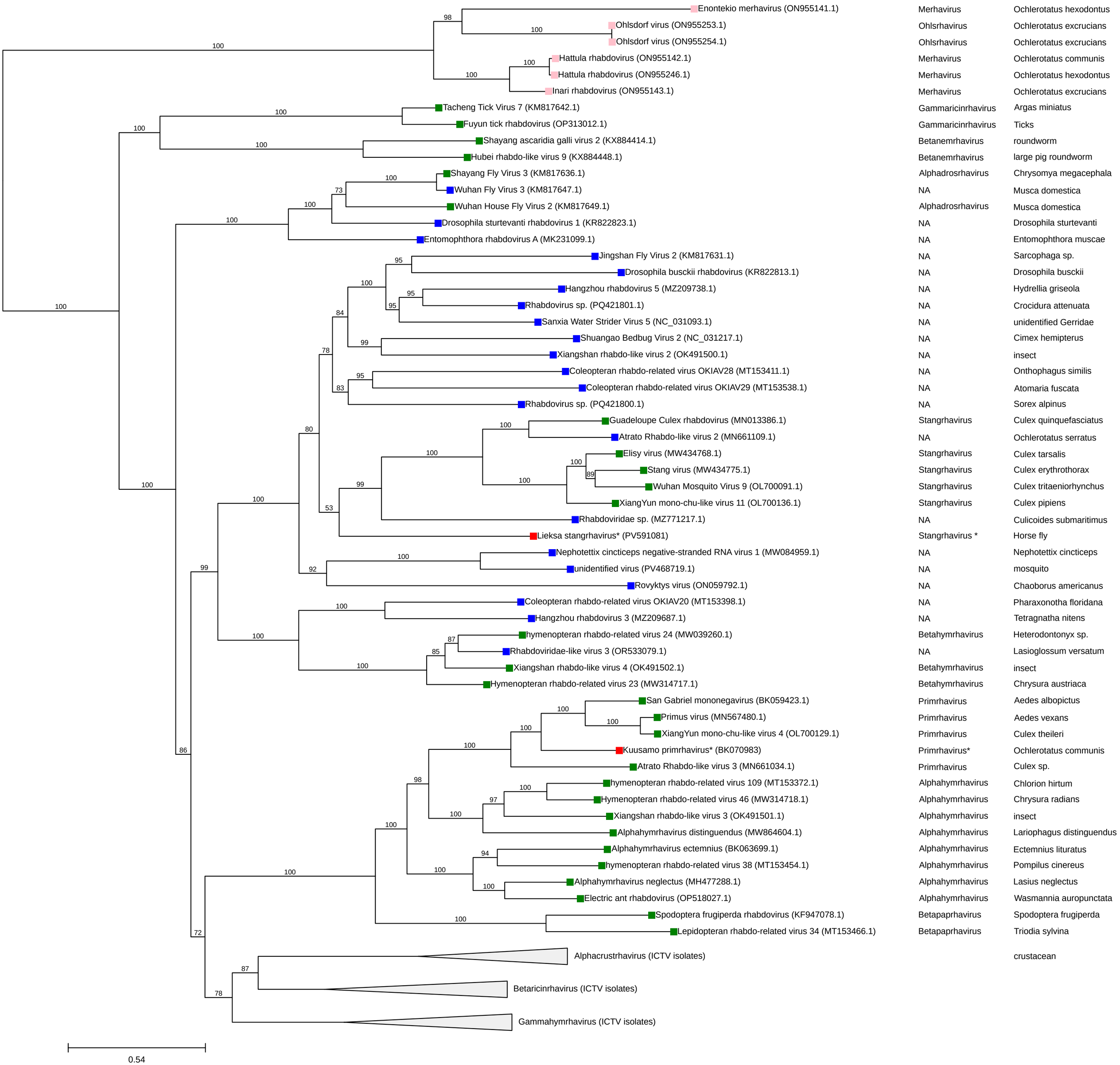
Fosso B, Santamaria M, D'Antonio M *et al.* MetaShot: an accurate workflow for taxon classification of host-associated microbiome from shotgun metagenomic data. *Bioinformatics* 2017;**33**:1730–2.

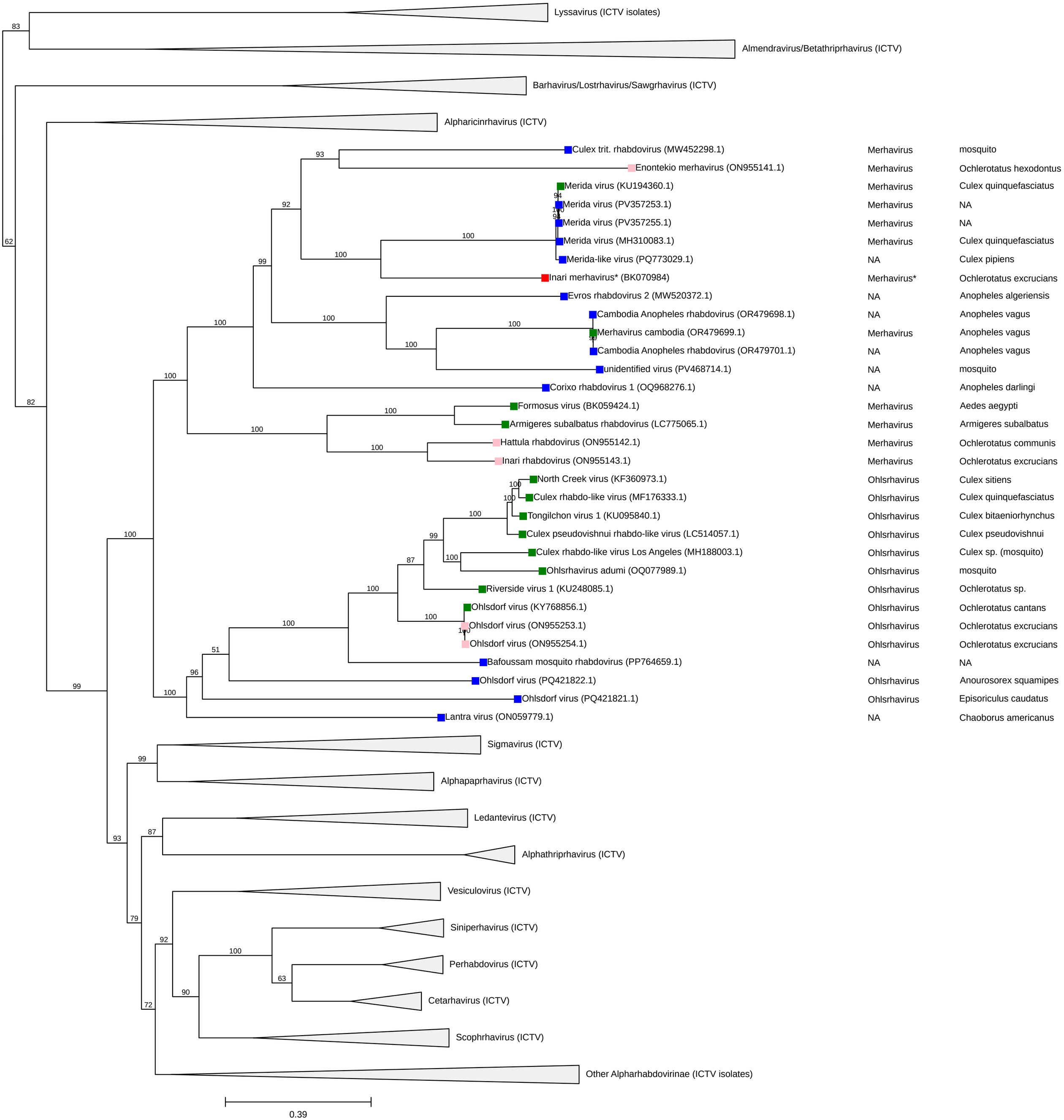
819 Ibañez-Lligoña M, Colomer-Castell S, González-Sánchez A *et al.* Bioinformatic Tools for NGS-Based
820 Metagenomics to Improve the Clinical Diagnosis of Emerging, Re-Emerging and New Viruses.
821 *Viruses* 2023;**15**:587.
822 Kalantar KL, Carvalho T, de Bourcy CFA *et al.* IDseq—An open source cloud-based pipeline and
823 analysis service for metagenomic pathogen detection and monitoring. *GigaScience*
824 2020;**9**:giaa111.
825 Kalyaanamoorthy S, Minh BQ, Wong TKF *et al.* ModelFinder: fast model selection for accurate
826 phylogenetic estimates. *Nat Methods* 2017;**14**:587–9.
827 Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in
828 Performance and Usability. *Mol Biol Evol* 2013;**30**:772–80.
829 Korhonen EM, Truong Nguyen PT, Faolotto G *et al.* Complete coding sequence of an Aedes flavivirus
830 strain isolated from Aedes albopictus collected in Northern Italy. *Microbiol Resour Announc*
831 2024;**13**:e00146-24.
832 Kuivanen S, Levanov L, Kareinen L *et al.* Detection of novel tick-borne pathogen, Alongshan virus, in
833 Ixodes ricinus ticks, south-eastern Finland, 2019. *Eurosurveillance* 2019;**24**, DOI: 10.2807/1560-
834 7917.ES.2019.24.27.1900394.
835 Lehikoinen J, Nurmi K, Ainola M *et al.* Epstein-Barr Virus in the Cerebrospinal Fluid and Blood
836 Compartments of Patients With Multiple Sclerosis and Controls. *Neurol Neuroimmunol*
837 *Neuroinflammation* 2024;**11**:e200226.
838 Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Prepr*
839 *ArXiv13033997* 2013.
840 Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100.
841 Li Y, Wang H, Nie K *et al.* VIP: an integrated pipeline for metagenomics of virus identification and
842 discovery. *Sci Rep* 2016;**6**:23774.
843 Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju.
844 *Nat Commun* 2016;**7**:11257.
845 Minh BQ, Schmidt HA, Chernomor O *et al.* IQ-TREE 2: New Models and Efficient Methods for
846 Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 2020;**37**:1530–4.
847 Mistry J, Chuguransky S, Williams L *et al.* Pfam: The protein families database in 2021. *Nucleic Acids*
848 *Res* 2021;**49**:D412–9.
849 Morfopoulou S, Plagnol V. Bayesian mixture analysis for metagenomic community profiling.
850 *Bioinformatics* 2015;**31**:2930–8.
851 Mull N, Erdin M, Smura T *et al.* Novel Ozark Orthohantavirus in Hispid Cotton Rats (Sigmodon hispidus),
852 Arkansas, USA. *Emerg Infect Dis* 2023;**29**:2576–8.
853 Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal
854 binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res*
855 *Int J Rapid Publ Rep Genes Genomes* 2008;**15**:387–96.
856 Nunes MRT, Contreras-Gutierrez MA, Guzman H *et al.* Genetic characterization, molecular
857 epidemiology, and phylogenetic relationships of insect-specific viruses in the taxon Negevirus.
858 *Virology* 2017;**504**:152–67.
859 Ogola JG, Alburkat H, Smura T *et al.* Detection and genetic characterization of alphacoronaviruses in
860 co-roosting bat species, southeastern Kenya. 2024:2024.12.23.24319537.
861 Pérot P, Bigot T, Temmam S *et al.* Microseek: A Protein-Based Metagenomic Pipeline for Virus
862 Diagnostic and Discovery. *Viruses* 2022;**14**:1990.
863 Plyusnin I, Vapalahti O, Sironen T *et al.* Enhanced Viral Metagenomics with Lazypipe 2. *Viruses*
864 2023;**15**:431.
865 Robinson JT, Thorvaldsdottir H, Turner D *et al.* igv.js: an embeddable JavaScript implementation of the
866 Integrative Genomics Viewer (IGV). *Bioinformatics* 2023;**39**:btac830.
867 Sakaguchi S, Nakano T, Nakagawa S. NeoRdRp2 with improved seed data, annotations, and scoring.
868 *Front Virol* 2024;**4**, DOI: 10.3389/fviro.2024.1378695.
869 Scheuch M, Höper D, Beer M. RIEMS: A software pipeline for sensitive and comprehensive taxonomic
870 classification of reads from metagenomics datasets. *BMC Bioinformatics* 2015;**16**, DOI:
871 10.1186/s12859-015-0503-6.
872 Shen W, Le S, Li Y *et al.* SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation.
873 *PLOS ONE* 2016;**11**:e0163962.
874 Shen W, Ren H. TaxonKit: A practical and efficient NCBI taxonomy toolkit. *J Genet Genomics Yi Chuan*
875 *Xue Bao* 2021;**48**:844–50.
876 Singh U, Wurtele ES. orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics* 2021;**37**:3019–
877 20.

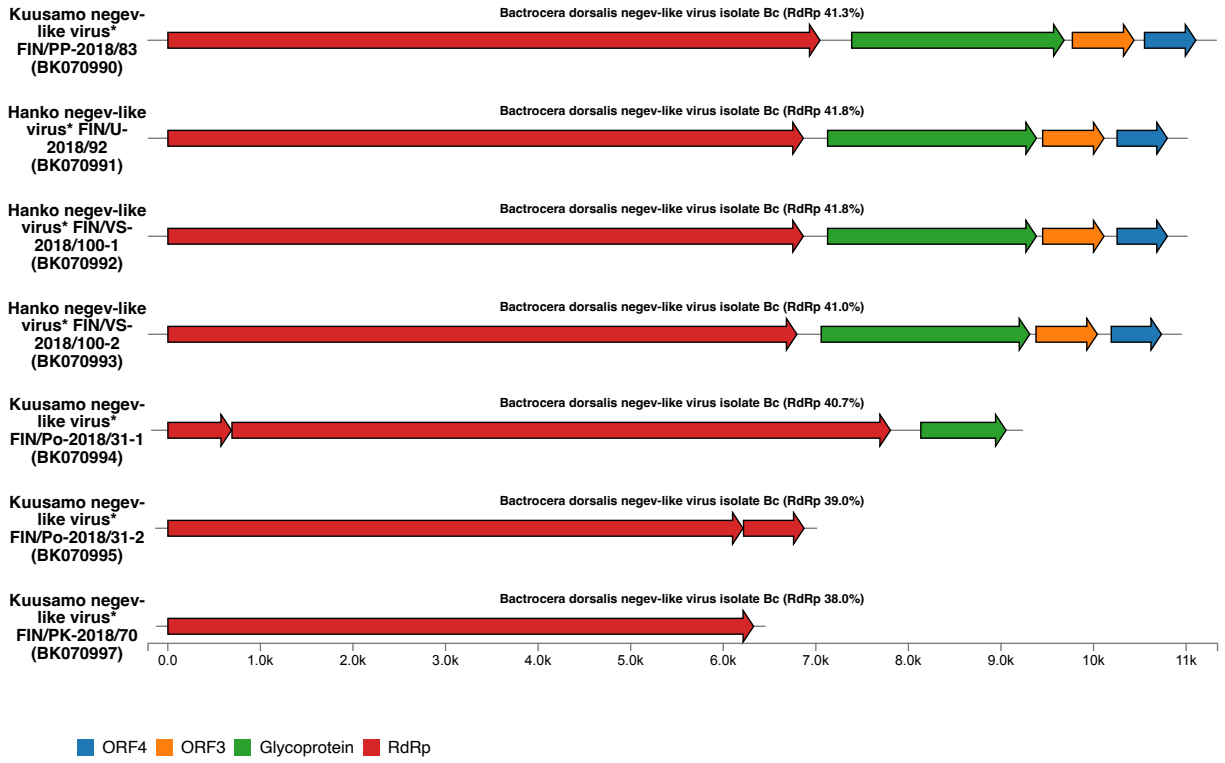
Somervuo P, Holm L. SANSparallel: interactive homology search against Uniprot. *Nucleic Acids Res* 2015;**43**:W24–9.
Tangherlini M, Dell'Anno A, Zeigler Allen L *et al.* Assessing viral taxonomic composition in benthic marine ecosystems: reliability and efficiency of different bioinformatic tools for viral metagenomic analyses. *Sci Rep* 2016;**6**:28428.
Tarasov A, Vilella AJ, Cuppen E *et al.* Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 2015;**31**:2032–4.
Törönen P, Holm L. PANNZER-A practical tool for protein function prediction. *Protein Sci Publ Protein Soc* 2022;**31**:118–28.
Truong Nguyen PT, Culverwell CL, Suvanto MT *et al.* Characterisation of the RNA Virome of Nine Ochlerotatus Species in Finland. *Viruses* 2022;**14**:1489.
Velden N van der. *Geneviewer: Gene Cluster Visualizations*., 2025.
Vilsker M, Moosa Y, Nooij S *et al.* Genome Detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics* 2019;**35**:871–3.
Virtanen J, Hautala K, Utriainen M *et al.* Equine dermatitis outbreak associated with parapoxvirus. 2023:2023.09.01.555671.
Virtanen J, Zalewski A, Kołodziej-Sobocińska M *et al.* Diversity and transmission of Aleutian mink disease virus in feral and farmed American mink and native mustelids. *Virus Evol* 2021;**7**:veab075.
de Vries JJC, Brown JR, Fischer N *et al.* Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. *J Clin Virol* 2021;**141**:104908.
Walker PJ, Freitas-Astúa J, Bejerman N *et al.* ICTV Virus Taxonomy Profile: Rhabdoviridae 2022. *J Gen Virol* 2022;**103**:001689.
Zakham F, Albalawi AE, Alanazi AD *et al.* Viral RNA Metagenomics of Hyalomma Ticks Collected from Dromedary Camels in Makkah Province, Saudi Arabia. *Viruses* 2021;**13**:1396.
Zhao G, Wu G, Lim ES *et al.* VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology* 2017;**503**:21–30.
Zwagemaker F, Schmitz D, Nooij S *et al.* DennisSchmitz/Jovian: 1.2.07. 2021, DOI: 10.5281/zenodo.4899547.

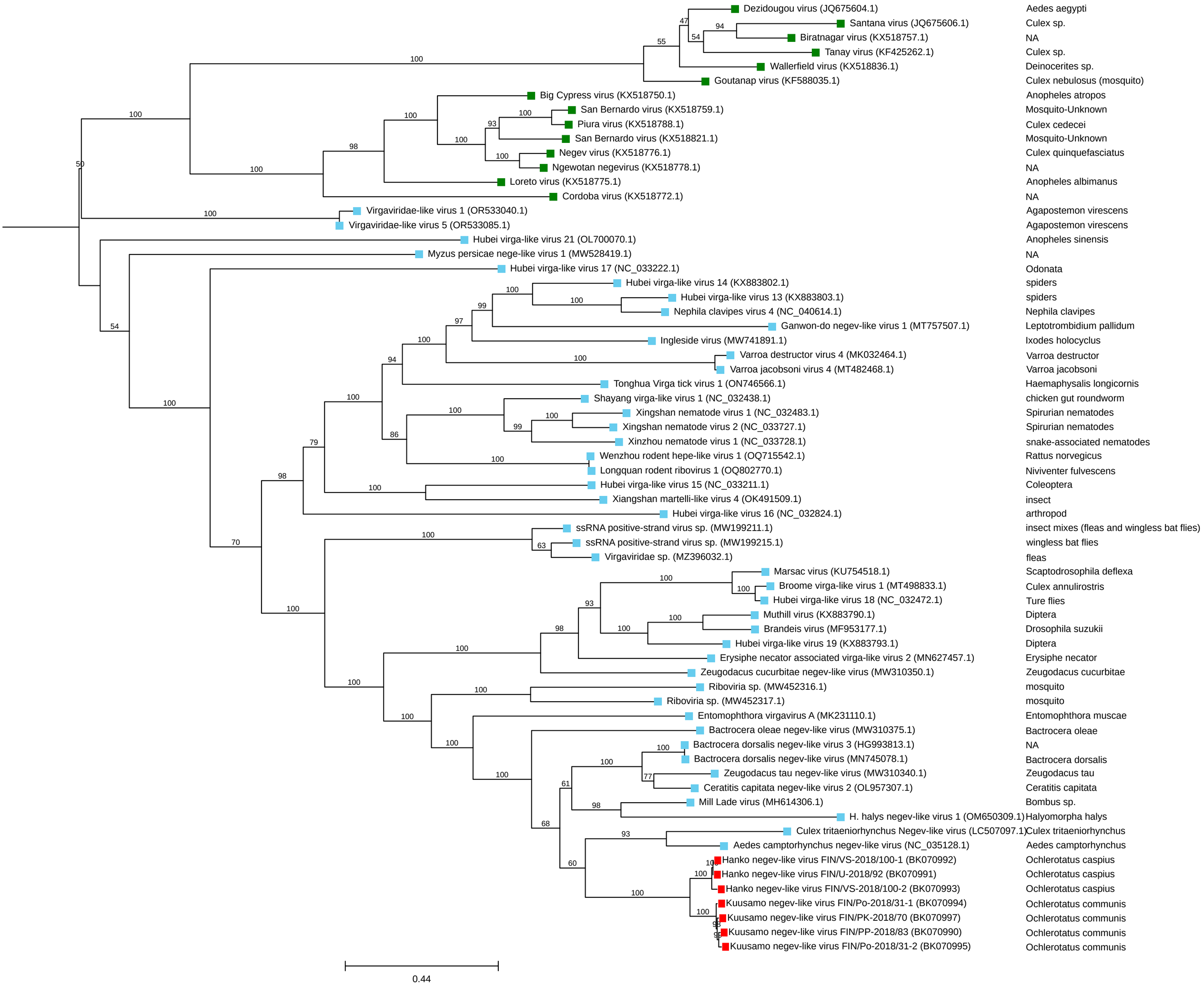

1.00
0.75
0.50
0.25
0.00
Precision
Recall
F1
Time
Annotation Strategy
Lazypipe2
Vi.chain0
Vi.chain1
Vi.chain2
Vi.chain3
Vi.chain3.env
Vi.chain4

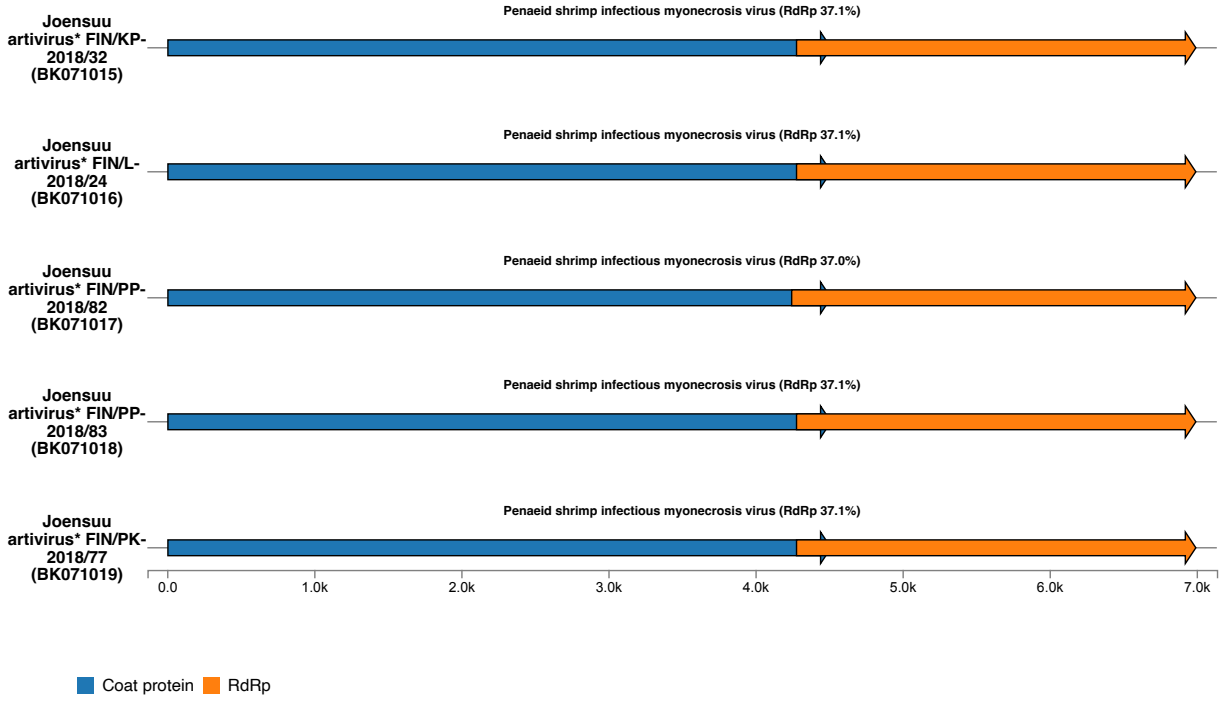

Kuusamo primrhavirus* FIN/PP-2018/83 (BK070983)
Primus virus (L 52.3%)
Lieksa stangrhavirus* (PV591081)
Xiangshan rhabdo-like virus 2 (L 35.8%)
Inari merhavirus*FIN/L-2018/84 (BK070984)
Culex rhabdovirus (L 52.1%)
0.0
1.0k
2.0k
3.0k
4.0k
5.0k
6.0k
7.0k
8.0k
9.0k
10k
11k
12k
L
G
M
P
N
unknown

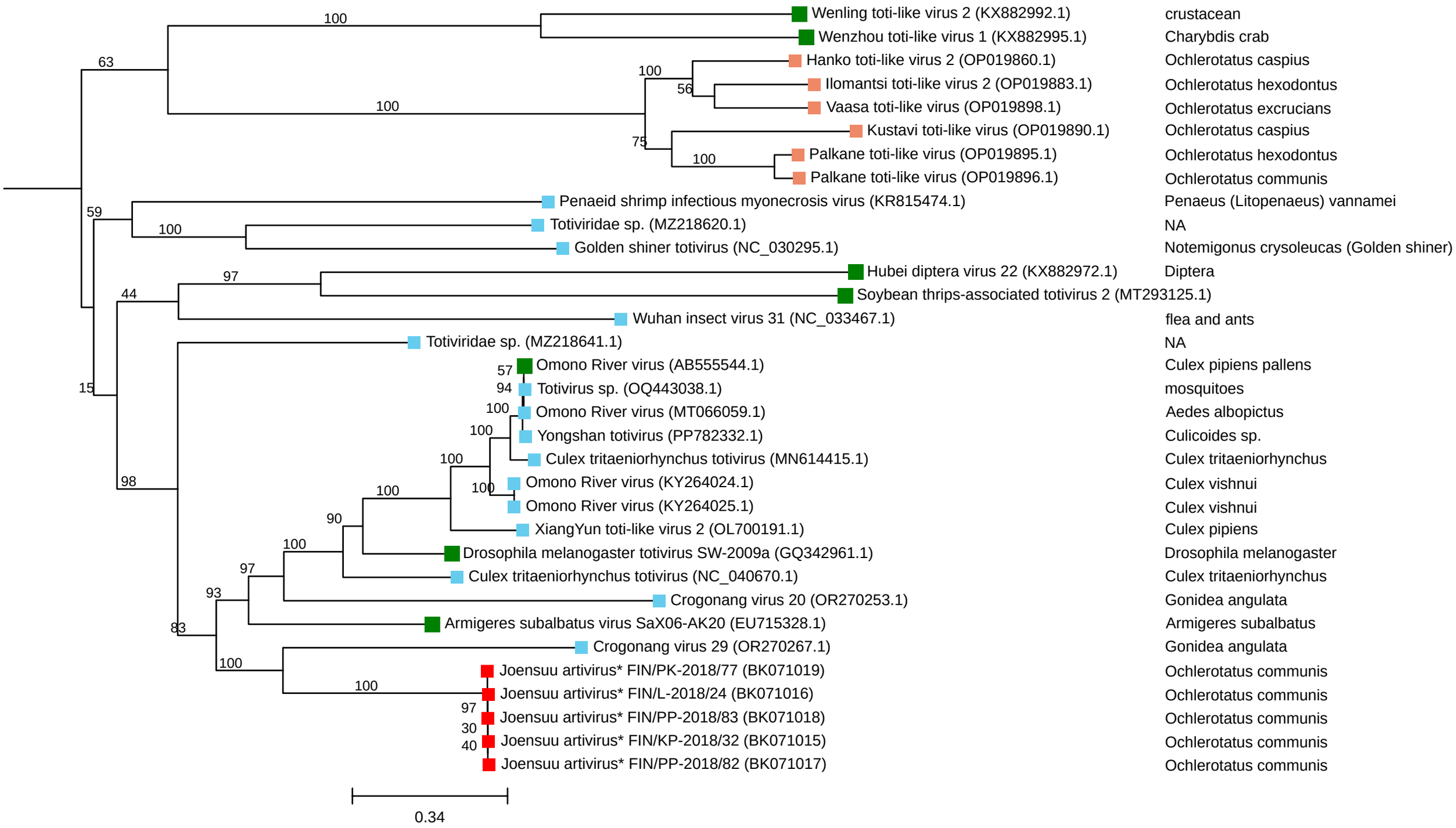

Enontekio merhavirus (ON955141.1) Merhavirus Ochlerotatus hexodontus
Ohlsdorf virus (ON955253.1) Ohlsrhavirus Ochlerotatus excrucians
Ohlsdorf virus (ON955254.1) Ohlsrhavirus Ochlerotatus excrucians
Hattula rhabdovirus (ON955142.1) Merhavirus Ochlerotatus communis
Hattula rhabdovirus (ON955246.1) Merhavirus Ochlerotatus hexodontus
Inari rhabdovirus (ON955143.1) Merhavirus Ochlerotatus excrucians
Tacheng Tick Virus 7 (KM817642.1) Gammaricinrhavirus Argas miniatus
Fuyun tick rhabdovirus (OP313012.1) Gammaricinrhavirus Ticks
Shayang ascaridia galli virus 2 (KX884414.1) Betanemrhavirus roundworm
Hubei rhabdo-like virus 9 (KX884448.1) Betanemrhavirus large pig roundworm
Shayang Fly Virus 3 (KM817636.1) Alphadrosrhavirus Chrysomya megacephala
Wuhan Fly Virus 3 (KM817647.1) NA Musca domestica
Wuhan House Fly Virus 2 (KM817649.1) Alphadrosrhavirus Musca domestica
Drosophila sturtevanti rhabdovirus 1 (KR822823.1) NA Drosophila sturtevanti
Entomophthora rhabdovirus A (MK231099.1) NA Entomophthora muscae
Jingshan Fly Virus 2 (KM817631.1) NA Sarcophaga sp.
Drosophila busckii rhabdovirus (KR822813.1) NA Drosophila busckii
Hangzhou rhabdovirus 5 (MZ209738.1) NA Hydrellia griseola
Rhabdovirus sp. (PQ421801.1) NA Crocidura attenuata
Sanxia Water Strider Virus 5 (NC_031093.1) NA unidentified Gerridae
Shuangao Bedbug Virus 2 (NC_031217.1) NA Cimex hemipterus
Xiangshan rhabdo-like virus 2 (OK491500.1) NA insect
Coleopteran rhabdo-related virus OKIAV28 (MT153411.1) NA Onthophagus similis
Coleopteran rhabdo-related virus OKIAV29 (MT153538.1) NA Atomaria fuscata
Rhabdovirus sp. (PQ421800.1) NA Sorex alpinus
Guadeloupe Culex rhabdovirus (MN013386.1) Stangrhavirus Culex quinquefasciatus
Atrato Rhabdo-like virus 2 (MN661109.1) NA Ochlerotatus serratus
Elisy virus (MW434768.1) Stangrhavirus Culex tarsalis
Stang virus (MW434775.1) Stangrhavirus Culex erythrothorax
Wuhan Mosquito Virus 9 (OL700091.1) Stangrhavirus Culex tritaeniorhynchus
XiangYun mono-chu-like virus 11 (OL700136.1) Stangrhavirus Culex pipiens
Rhabdoviridae sp. (MZ771217.1) NA Culicoides submaritimus
Lieksa stangrhavirus* (PV591081) Stangrhavirus * Horse fly
Nephotettix cincticeps negative-stranded RNA virus 1 (MW084959.1) NA Nephotettix cincticeps
unidentified virus (PV468719.1) NA mosquito
Rovyktys virus (ON059792.1) NA Chaoborus americanus
Coleopteran rhabdo-related virus OKIAV20 (MT153398.1) NA Pharaxonotha floridana
Hangzhou rhabdovirus 3 (MZ209687.1) NA Tetragnatha nitens
hymenopteran rhabdo-related virus 24 (MW039260.1) Betahymrhavirus Heterodontonyx sp.
Rhabdoviridae-like virus 3 (OR533079.1) NA Lasioglossum versatum
Xiangshan rhabdo-like virus 4 (OK491502.1) Betahymrhavirus insect
Hymenopteran rhabdo-related virus 23 (MW314717.1) Betahymrhavirus Chrysura austriaca
San Gabriel mononegavirus (BK059423.1) Primrhavirus Aedes albopictus
Primus virus (MN567480.1) Primrhavirus Aedes vexans
XiangYun mono-chu-like virus 4 (OL700129.1) Primrhavirus Culex theileri
Kuusamo primrhavirus* (BK070983) Primrhavirus* Ochlerotatus communis
Atrato Rhabdo-like virus 3 (MN661034.1) Primrhavirus Culex sp.
hymenopteran rhabdo-related virus 109 (MT153372.1) Alphahymrhavirus Chlorion hirtum
Hymenopteran rhabdo-related virus 46 (MW314718.1) Alphahymrhavirus Chrysura radians
Xiangshan rhabdo-like virus 3 (OK491501.1) Alphahymrhavirus insect
Alphahymrhavirus distinguendus (MW864604.1) Alphahymrhavirus Lariophagus distinguendus
Alphahymrhavirus ectemnius (BK063699.1) Alphahymrhavirus Ectemnius lituratus
hymenopteran rhabdo-related virus 38 (MT153454.1) Alphahymrhavirus Pompilus cinereus
Alphahymrhavirus neglectus (MH477288.1) Alphahymrhavirus Lasius neglectus
Electric ant rhabdovirus (OP518027.1) Alphahymrhavirus Wasmannia auropunctata
Spodoptera frugiperda rhabdovirus (KF947078.1) Betapaprhavirus Spodoptera frugiperda
Lepidopteran rhabdo-related virus 34 (MT153466.1) Betapaprhavirus Triodia sylvina
Alphacrustrhavirus (ICTV isolates) crustacean
Betaricinrhavirus (ICTV isolates)
Gammahymrhavirus (ICTV isolates)
0.54

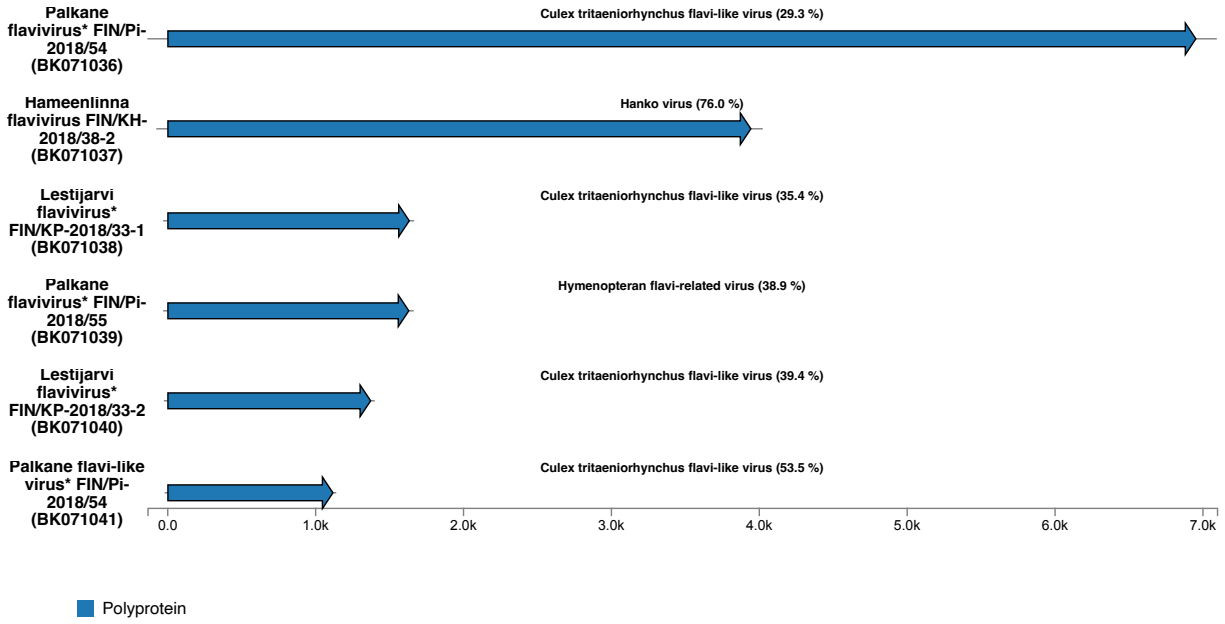

Lyssavirus (ICTV isolates)
Almendravirus/Betathriprhavirus (ICTV)
Barhavirus/Lostrhavirus/Sawgrhavirus (ICTV)
Alphaericinrhavirus (ICTV)
Culex trit. rhabdovirus (MW452298.1) Merhavirus mosquito
Enontekio merhavirus (ON955141.1) Merhavirus Ochlerotatus hexodontus
Merida virus (KU194360.1) Merhavirus Culex quinquefasciatus
Merida virus (PV357253.1) Merhavirus NA
Merida virus (PV357255.1) Merhavirus NA
Merida virus (MH310083.1) Merhavirus Culex quinquefasciatus
Merida-like virus (PQ773029.1) NA Culex pipiens
Inari merhavirus* (BK070984) Merhavirus* Ochlerotatus excrucians
Evros rhabdovirus 2 (MW520372.1) NA Anopheles algeriensis
Cambodia Anopheles rhabdovirus (OR479698.1) NA Anopheles vagus
Merhavirus cambodia (OR479699.1) Merhavirus Anopheles vagus
Cambodia Anopheles rhabdovirus (OR479701.1) NA Anopheles vagus
unidentified virus (PV468714.1) NA mosquito
Corixo rhabdovirus 1 (OQ968276.1) NA Anopheles darlingi
Formosus virus (BK059424.1) Merhavirus Aedes aegypti
Armigeres subalbatus rhabdovirus (LC775065.1) Merhavirus Armigeres subalbatus
Hattula rhabdovirus (ON955142.1) Merhavirus Ochlerotatus communis
Inari rhabdovirus (ON955143.1) Merhavirus Ochlerotatus excrucians
North Creek virus (KF360973.1) Ohlsrhavirus Culex sitiens
Culex rhabdo-like virus (MF176333.1) Ohlsrhavirus Culex quinquefasciatus
Tongilchon virus 1 (KU095840.1) Ohlsrhavirus Culex bitaeniorhynchus
Culex pseudovishnui rhabdo-like virus (LC514057.1) Ohlsrhavirus Culex pseudovishnui
Culex rhabdo-like virus Los Angeles (MH188003.1) Ohlsrhavirus Culex sp. (mosquito)
Ohlsrhavirus adumi (OQ077989.1) Ohlsrhavirus mosquito
Riverside virus 1 (KU248085.1) Ohlsrhavirus Ochlerotatus sp.
Ohlsdorf virus (KY768856.1) Ohlsrhavirus Ochlerotatus cantans
Ohlsdorf virus (ON955253.1) Ohlsrhavirus Ochlerotatus excrucians
Ohlsdorf virus (ON955254.1) Ohlsrhavirus Ochlerotatus excrucians
Bafoussam mosquito rhabdovirus (PP764659.1) NA NA
Ohlsdorf virus (PQ421822.1) Ohlsrhavirus Anourosorex squamipes
Ohlsdorf virus (PQ421821.1) Ohlsrhavirus Episoriculus caudatus
Lantra virus (ON059779.1) NA Chaoborus americanus
Sigmavirus (ICTV)
Alphapaprhavirus (ICTV)
Ledantevirus (ICTV)
Alphathriprhavirus (ICTV)
Vesiculovirus (ICTV)
Siniperhavirus (ICTV)
Perhabdovirus (ICTV)
Cetarhavirus (ICTV)
Scophrhavirus (ICTV)
Other Alpharhabdovirinae (ICTV isolates)
0.39

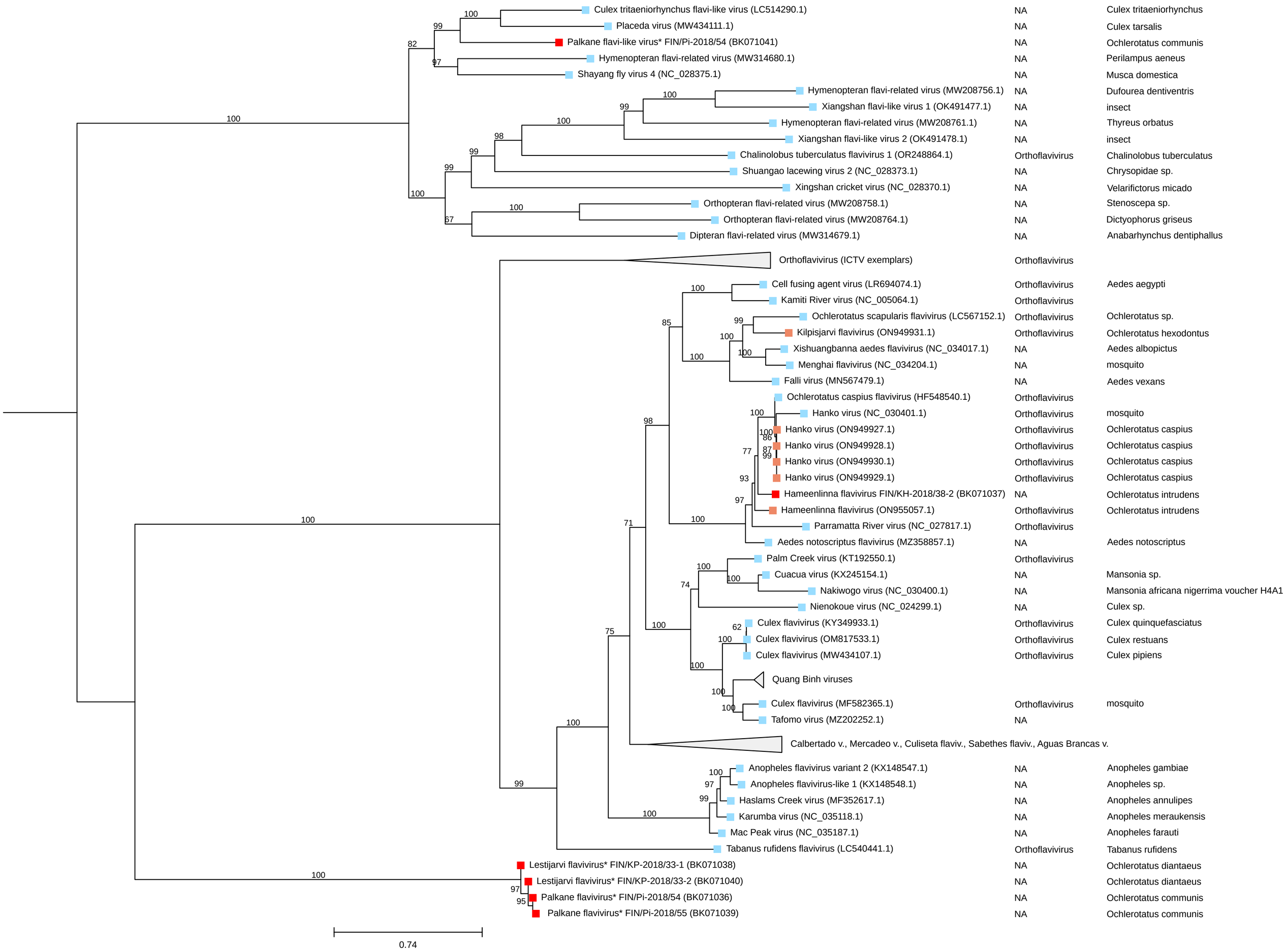

Kuusamo negev-like virus* FIN/PP-2018/83 (BK070990)
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 41.3%)
Hanko negev-like virus* FIN/U-2018/92 (BK070991)
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 41.8%)
Hanko negev-like virus* FIN/VS-2018/100-1 (BK070992)
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 41.8%)
Hanko negev-like virus* FIN/VS-2018/100-2 (BK070993)
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 41.0%)
Kuusamo negev-like virus* FIN/Po-2018/31-1 (BK070994)
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 40.7%)
Kuusamo negev-like virus* FIN/Po-2018/31-2 (BK070995)
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 39.0%)
Kuusamo negev-like virus* FIN/PK-2018/70 (BK070997)
Bactrocera dorsalis negev-like virus isolate Bc (RdRp 38.0%)
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


Dezidougou virus (JQ675604.1) Aedes aegypti
Santana virus (JQ675606.1) Culex sp.
Biratnagar virus (KX518757.1) NA
Tanay virus (KF425262.1) Culex sp.
Wallerfield virus (KX518836.1) Deinocerites sp.
Goutanap virus (KF588035.1) Culex nebulosus (mosquito)
Big Cypress virus (KX518750.1) Anopheles atropos
San Bernardo virus (KX518759.1) Mosquito-Unknown
Piura virus (KX518788.1) Culex cedecei
San Bernardo virus (KX518821.1) Mosquito-Unknown
Negev virus (KX518776.1) Culex quinquefasciatus
Ngewotan negevirus (KX518778.1) NA
Loreto virus (KX518775.1) Anopheles albimanus
Cordoba virus (KX518772.1) NA
Virgaviridae-like virus 1 (OR533040.1) Agapostemon virescens
Virgaviridae-like virus 5 (OR533085.1) Agapostemon virescens
Hubei virga-like virus 21 (OL700070.1) Anopheles sinensis
Myzus persicae nege-like virus 1 (MW528419.1) NA
Hubei virga-like virus 17 (NC_033222.1) Odonata
Hubei virga-like virus 14 (KX883802.1) spiders
Hubei virga-like virus 13 (KX883803.1) spiders
Nephila clavipes virus 4 (NC_040614.1) Nephila clavipes
Ganwon-do negev-like virus 1 (MT757507.1) Leptotrombidium pallidum
Ingleside virus (MW741891.1) Ixodes holocyclus
Varroa destructor virus 4 (MK032464.1) Varroa destructor
Varroa jacobsoni virus 4 (MT482468.1) Varroa jacobsoni
Tonghua Virga tick virus 1 (ON746566.1) Haemaphysalis longicornis
Shayang virga-like virus 1 (NC_032438.1) chicken gut roundworm
Xingshan nematode virus 1 (NC_032483.1) Spirurian nematodes
Xingshan nematode virus 2 (NC_033727.1) Spirurian nematodes
Xinzhou nematode virus 1 (NC_033728.1) snake-associated nematodes
Wenzhou rodent hepe-like virus 1 (OQ715542.1) Rattus norvegicus
Longquan rodent ribovirus 1 (OQ802770.1) Niviventer fulvescens
Hubei virga-like virus 15 (NC_033211.1) Coleoptera
Xiangshan martelli-like virus 4 (OK491509.1) insect
Hubei virga-like virus 16 (NC_032824.1) arthropod
ssRNA positive-strand virus sp. (MW199211.1) insect mixes (fleas and wingless bat flies)
ssRNA positive-strand virus sp. (MW199215.1) wingless bat flies
Virgaviridae sp. (MZ396032.1) fleas
Marsac virus (KU754518.1) Scaptodrosophila deflexa
Broome virga-like virus 1 (MT498833.1) Culex annulirostris
Hubei virga-like virus 18 (NC_032472.1) Ture flies
Muthill virus (KX883790.1) Diptera
Brandeis virus (MF953177.1) Drosophila suzukii
Hubei virga-like virus 19 (KX883793.1) Diptera
Erysiphe necator associated virga-like virus 2 (MN627457.1) Erysiphe necator
Zeugodacus cucurbitae negev-like virus (MW310350.1) Zeugodacus cucurbitae
Riboviria sp. (MW452316.1) mosquito
Riboviria sp. (MW452317.1) mosquito
Entomophthora virgavirus A (MK231110.1) Entomophthora muscae
Bactrocera oleae negev-like virus (MW310375.1) Bactrocera oleae
Bactrocera dorsalis negev-like virus 3 (HG993813.1) NA
Bactrocera dorsalis negev-like virus (MN745078.1) Bactrocera dorsalis
Zeugodacus tau negev-like virus (MW310340.1) Zeugodacus tau
Ceratitis capitata negev-like virus 2 (OL957307.1) Ceratitis capitata
Mill Lade virus (MH614306.1) Bombus sp.
H. halys negev-like virus 1 (OM650309.1) Halyomorpha halys
Culex tritaeniorhynchus Negev-like virus (LC507097.1) Culex tritaeniorhynchus
Aedes camptorhynchus negev-like virus (NC_035128.1) Aedes camptorhynchus
Hanko negev-like virus FIN/VS-2018/100-1 (BK070992) Ochlerotatus caspius
Hanko negev-like virus FIN/U-2018/92 (BK070991) Ochlerotatus caspius
Hanko negev-like virus FIN/VS-2018/100-2 (BK070993) Ochlerotatus caspius
Kuusamo negev-like virus FIN/Po-2018/31-1 (BK070994) Ochlerotatus communis
Kuusamo negev-like virus FIN/PK-2018/70 (BK070997) Ochlerotatus communis
Kuusamo negev-like virus FIN/PP-2018/83 (BK070990) Ochlerotatus communis
Kuusamo negev-like virus FIN/Po-2018/31-2 (BK070995) Ochlerotatus communis
0.44


Penaeid shrimp infectious myonecrosis virus (RdRp 37.1%)
Joensuu artivirus* FIN/KP-2018/32 (BK071015)
Penaeid shrimp infectious myonecrosis virus (RdRp 37.1%)
Joensuu artivirus* FIN/L-2018/24 (BK071016)
Penaeid shrimp infectious myonecrosis virus (RdRp 37.0%)
Joensuu artivirus* FIN/PP-2018/82 (BK071017)
Penaeid shrimp infectious myonecrosis virus (RdRp 37.1%)
Joensuu artivirus* FIN/PP-2018/83 (BK071018)
Penaeid shrimp infectious myonecrosis virus (RdRp 37.1%)
Joensuu artivirus* FIN/PK-2018/77 (BK071019)
0.0
1.0k
2.0k
3.0k
4.0k
5.0k
6.0k
7.0k
Coat protein
RdRp


Wenling toti-like virus 2 (KX882992.1) crustacean
Wenzhou toti-like virus 1 (KX882995.1) Charybdis crab
Hanko toti-like virus 2 (OP019860.1) Ochlerotatus caspius
Ilomantsi toti-like virus 2 (OP019883.1) Ochlerotatus hexodontus
Vaasa toti-like virus (OP019898.1) Ochlerotatus excrucians
Kustavi toti-like virus (OP019890.1) Ochlerotatus caspius
Palkane toti-like virus (OP019895.1) Ochlerotatus hexodontus
Palkane toti-like virus (OP019896.1) Ochlerotatus communis
Penaeid shrimp infectious myonecrosis virus (KR815474.1) Penaeus (Litopenaeus) vannamei
Totiviridae sp. (MZ218620.1) NA
Golden shiner totivirus (NC_030295.1) Notemigonus crysoleucas (Golden shiner)
Hubei diptera virus 22 (KX882972.1) Diptera
Soybean thrips-associated totivirus 2 (MT293125.1)
Wuhan insect virus 31 (NC_033467.1) flea and ants
Totiviridae sp. (MZ218641.1) NA
Omono River virus (AB555544.1) Culex pipiens pallens
Totivirus sp. (OQ443038.1) mosquitoes
Omono River virus (MT066059.1) Aedes albopictus
Yongshan totivirus (PP782332.1) Culicoides sp.
Culex tritaeniorhynchus totivirus (MN614415.1) Culex tritaeniorhynchus
Omono River virus (KY264024.1) Culex vishnui
Omono River virus (KY264025.1) Culex vishnui
XiangYun toti-like virus 2 (OL700191.1) Culex pipiens
Drosophila melanogaster totivirus SW-2009a (GQ342961.1) Drosophila melanogaster
Culex tritaeniorhynchus totivirus (NC_040670.1) Culex tritaeniorhynchus
Crogonang virus 20 (OR270253.1) Gonidea angulata
Armigeres subalbatus virus SaX06-AK20 (EU715328.1) Armigeres subalbatus
Crogonang virus 29 (OR270267.1) Gonidea angulata
Joensuu artivirus* FIN/PK-2018/77 (BK071019) Ochlerotatus communis
Joensuu artivirus* FIN/L-2018/24 (BK071016) Ochlerotatus communis
Joensuu artivirus* FIN/PP-2018/83 (BK071018) Ochlerotatus communis
Joensuu artivirus* FIN/KP-2018/32 (BK071015) Ochlerotatus communis
Joensuu artivirus* FIN/PP-2018/82 (BK071017) Ochlerotatus communis
0.34


Palkane flavivirus* FIN/Pi-2018/54 (BK071036)
Culex tritaeniorhynchus flavi-like virus (29.3 %)
Hameenlinna flavivirus FIN/KH-2018/38-2 (BK071037)
Hanko virus (76.0 %)
Lestijarvi flavivirus* FIN/KP-2018/33-1 (BK071038)
Culex tritaeniorhynchus flavi-like virus (35.4 %)
Palkane flavivirus* FIN/Pi-2018/55 (BK071039)
Hymenopteran flavi-related virus (38.9 %)
Lestijarvi flavivirus* FIN/KP-2018/33-2 (BK071040)
Culex tritaeniorhynchus flavi-like virus (39.4 %)
Palkane flavi-like virus* FIN/Pi-2018/54 (BK071041)
Culex tritaeniorhynchus flavi-like virus (53.5 %)
0.0
1.0k
2.0k
3.0k
4.0k
5.0k
6.0k
7.0k
Polyprotein


Culex tritaeniorhynchus flavi-like virus (LC514290.1) NA Culex tritaeniorhynchus
Placeda virus (MW434111.1) NA Culex tarsalis
Palkane flavi-like virus* FIN/Pi-2018/54 (BK071041) NA Ochlerotatus communis
Hymenopteran flavi-related virus (MW314680.1) NA Perilampus aeneus
Shayang fly virus 4 (NC_028375.1) NA Musca domestica
Hymenopteran flavi-related virus (MW208756.1) NA Dufourea dentiventris
Xiangshan flavi-like virus 1 (OK491477.1) NA insect
Hymenopteran flavi-related virus (MW208761.1) NA Thyreus orbatus
Xiangshan flavi-like virus 2 (OK491478.1) NA insect
Chalinolobus tuberculatus flavivirus 1 (OR248864.1) Orthoflavivirus Chalinolobus tuberculatus
Shuangao lacewing virus 2 (NC_028373.1) NA Chrysopidae sp.
Xingshan cricket virus (NC_028370.1) NA Velarifictorus micado
Orthopteran flavi-related virus (MW208758.1) NA Stenoscepa sp.
Orthopteran flavi-related virus (MW208764.1) NA Dictyophorus griseus
Dipteran flavi-related virus (MW314679.1) NA Anabarhynchus dentiphallus
Orthoflavivirus (ICTV exemplars) Orthoflavivirus
Cell fusing agent virus (LR694074.1) Orthoflavivirus Aedes aegypti
Kamiti River virus (NC_005064.1) Orthoflavivirus
Ochlerotatus scapularis flavivirus (LC567152.1) Orthoflavivirus Ochlerotatus sp.
Kilpisjarvi flavivirus (ON949931.1) Orthoflavivirus Ochlerotatus hexodontus
Xishuangbanna aedes flavivirus (NC_034017.1) NA Aedes albopictus
Menghai flavivirus (NC_034204.1) NA mosquito
Falli virus (MN567479.1) NA Aedes vexans
Ochlerotatus caspius flavivirus (HF548540.1) Orthoflavivirus
Hanko virus (NC_030401.1) Orthoflavivirus mosquito
Hanko virus (ON949927.1) Orthoflavivirus Ochlerotatus caspius
Hanko virus (ON949928.1) Orthoflavivirus Ochlerotatus caspius
Hanko virus (ON949930.1) Orthoflavivirus Ochlerotatus caspius
Hanko virus (ON949929.1) Orthoflavivirus Ochlerotatus caspius
Hameenlinna flavivirus FIN/KH-2018/38-2 (BK071037) NA Ochlerotatus intrudens
Hameenlinna flavivirus (ON955057.1) Orthoflavivirus Ochlerotatus intrudens
Parramatta River virus (NC_027817.1) Orthoflavivirus
Aedes notoscriptus flavivirus (MZ358857.1) NA Aedes notoscriptus
Palm Creek virus (KT192550.1) Orthoflavivirus
Cuacua virus (KX245154.1) NA Mansonia sp.
Nakiwogo virus (NC_030400.1) NA Mansonia africana nigerrima voucher H4A1
Nienokoue virus (NC_024299.1) NA Culex sp.
Culex flavivirus (KY349933.1) Orthoflavivirus Culex quinquefasciatus
Culex flavivirus (OM817533.1) Orthoflavivirus Culex restuans
Culex flavivirus (MW434107.1) Orthoflavivirus Culex pipiens
Quang Binh viruses
Culex flavivirus (MF582365.1) Orthoflavivirus mosquito
Tafomo virus (MZ202252.1) NA
Calbertado v., Mercadeo v., Culiseta flaviv., Sabethes flaviv., Aguas Brancas v.
Anopheles flavivirus variant 2 (KX148547.1) NA Anopheles gambiae
Anopheles flavivirus-like 1 (KX148548.1) NA Anopheles sp.
Haslams Creek virus (MF352617.1) NA Anopheles annulipes
Karumba virus (NC_035118.1) NA Anopheles meraukensis
Mac Peak virus (NC_035187.1) NA Anopheles farauti
Tabanus rufidens flavivirus (LC540441.1) Orthoflavivirus Tabanus rufidens
Lestijarvi flavivirus* FIN/KP-2018/33-1 (BK071038) NA Ochlerotatus diantaeus
Lestijarvi flavivirus* FIN/KP-2018/33-2 (BK071040) NA Ochlerotatus diantaeus
Palkane flavivirus* FIN/Pi-2018/54 (BK071036) NA Ochlerotatus communis
Palkane flavivirus* FIN/Pi-2018/55 (BK071039) NA Ochlerotatus communis
0.74