

Research article

Published
2026-08-27

Cite as

Audric Berger, Margaux J.M. Lefebvre, Jacques Dainat, Davy Jiolle, Isabelle Moltini-Conclois, Loic Talignani, Emilio Mastriani, Sylvie Cornelié, Nicolas Berthet and Christophe Paupy
(2026) *ViroSeek: a viral detection pipeline for second-generation sequencing*,
Peer Community Journal,
6: e82.

Correspondence

audric.berger@ird.fr
margaux.jm.lefebvre@gmail.com
christophe.paupy@ird.fr

Peer-review

Peer reviewed and
recommended by
PCI Genomics,

<https://doi.org/10.24072/pci.genomics.100695>



This article is licensed
under the Creative Commons
Attribution 4.0 License.

ViroSeek: a viral detection pipeline for second-generation sequencing

Audric Berger^{,#1}, Margaux J.M. Lefebvre^{,#1,2},
Jacques Dainat^{,1}, Davy Jiolle^{,1}, Isabelle
Moltini-Conclois^{,1}, Loic Talignani^{,1}, Emilio
Mastriani^{,3}, Sylvie Cornelié^{,1}, Nicolas
Berthet^{,4,5}, and Christophe Paupy^{,1}

Volume 6 (2026), article e82

<https://doi.org/10.24072/pcjournal.767>

Abstract

Arbovirus emergences represent a rising public health issue and are exacerbated by climate change and globalization. Virome analysis has become a key approach for monitoring and managing infectious diseases, yet existing tools often remain technically complex and inaccessible to non-specialists. In this context, we present ViroSeek, a reproducible and accessible bioinformatics pipeline specifically designed for the taxonomic analysis of second-generation sequencing data from target-enriched libraries. ViroSeek performs a series of automated steps: quality control, trimming, host and bacterial sequence removal, assembly, taxonomic assignment, read remapping for quantification, and PCR duplicate removal. The whole process is designed to produce a clear, usable viral taxonomy table that is suitable for diversity studies. ViroSeek was empirically validated on enriched control samples containing a known panel of viruses. All the expected viruses were correctly detected. Bacterial and host contaminant sequences were effectively removed. The pipeline is freely available and fully documented, supporting its adoption and adaptation by the research community.

¹MIVEGEC, Univ. Montpellier, CNRS, IRD, Montpellier, France, ²Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, ³Harbin akeda data information technology co. LTD, Harbin Medical University, Selinus University, University of Catania, Italy, ⁴Institut Pasteur, Université Paris-cité, Unité Épidémiologie et Physiopathologie des Virus Oncogènes, Paris, France, ⁵Institut Pasteur de Bangui, Bangui, Central African Republic, #Equal contribution



Introduction

Viruses are a major global public health issue, particularly because of arboviral diseases and zoonotic viruses (de Almeida et al., 2023; Dong & Soong, 2021). Climate change, growing urbanization or population movements contribute to increasing the risks of viral emergence and transmission (Baker et al., 2022). In this context, virome analysis, which aims to study the viral diversity present in different environments or hosts, is now a major challenge in microbial ecology, human health and environmental monitoring (de Almeida et al., 2023; de Almeida et al., 2021; Quick et al., 2016). Traditional virus characterization primarily depends on laboratory-based methods such as cell culture (Grandien & Svedmyr, 1983). While effective, these approaches are time-intensive and generally restricted to the identification of individual viruses. In contrast, next-generation sequencing (NGS) has emerged as a high-throughput, cost-effective alternative for large-scale viral detection. Therefore, numerous bioinformatics tools have been developed to detect, classify and annotate viral sequences from metagenomic data (Kaiser et al., 2025).

However, despite their abundance, most of these pipelines have significant limitations. Some are too specialized (i.e. Metaphage or poreCov) (Brandt et al., 2021; Pandolfo et al., 2022), others are poorly maintained, difficult to install due to obsolete dependencies, or contain errors that prevent proper execution of the scripts. For example, a study comparing twenty virome pipelines for third-generation sequencing data revealed that only five of them were usable, the others being either unsuitable or impossible to operate (Kaiser et al., 2025).

Furthermore, several existing pipelines and tools are specifically designed for the detection of viruses infecting bacteria and archaea (Guo et al., 2021; Kieft et al., 2020; Ren et al., 2017), making them unsuitable for the identification of arboviruses. In this study, we focus on the analysis of viromes derived from second-generation (Illumina) sequencing data from target-enriched libraries. The pipeline is designed to handle both RNA and DNA viruses, with a particular emphasis on eukaryotic viruses, including arboviruses of public health concern.

Pipelines such as Taxprofiler (Stamouli et al., 2023), MetaDenovo (Lundin et al., 2025), PIMGAVir (Mastriani et al., 2022), MicroSeek (Pérot et al., 2022), or VirusTaxo (Raju et al., 2022) were initially considered. However, their integration proved challenging due to installation issues and output files were found to be unsuitable for some downstream analyses.

In the case of MetaDenovo, assembly is performed across all samples collectively. However, retaining assemblies on a per-sample basis facilitates downstream analyses such as phylogenetic placement and genetic diversity assessment. Furthermore, MetaDenovo performs taxonomic assignment using *BLASTp*, which requires prior ORF prediction and may introduce biases in viral sequences (Finkel et al., 2018). Although *BLASTn* avoids this step, protein-level searches provide higher sensitivity for divergent viruses (Pearson, 2014). Therefore, *BLASTx*, which evaluates all six reading frames without explicit ORF calling, offers a more robust alternative (Buchfink et al., 2021). This approach is implemented when *Diamond* is selected as the taxonomic classifier in Taxprofiler. However, raw reads are not assembled at any stage of the workflow, which does not enable the phylogenetic placement of sequences assigned to specific species. To address this limitation, we adopted an assembly-based strategy, enabling the reconstruction of longer viral sequences and facilitating downstream analyses of genetic diversity. While VirusTaxo provides taxonomic classification of viral sequences, it does not include an assembly step and requires pre-assembled contigs as input, limiting its standalone utility in a complete pipeline. Finally, other available pipelines also present accessibility limitations. For instance, PIMGAVir presented challenges including high RAM requirements and restrictive access rights, while MicroSeek lacked publicly available code or workflow documentation.

Although flexible workflows such as MetaDenovo or Viralmetagenome provide comprehensive solutions for the analysis of diverse viral communities, they are designed to accommodate a broad range of datasets and experimental designs. Consequently, they offer extensive flexibility and multiple configurable steps in this objective. ViroSeek was developed for a more specific

application: the analysis of short-read target-enriched libraries generated for the monitoring and characterization of eukaryotic viruses.

Rather than integrating multiple interchangeable tools, the design of ViroSeek relies on two main methodological choices: an assembly-based strategy to generate sample-specific viral contigs and taxonomic assignment using protein-level similarity searches. This optimized combination of methods provides consistency and interpretable results. This approach remains accessible to non-expert users and devoid of unnecessary options while maintaining sensitivity for divergent eukaryotic viruses. To validate the pipeline, we analyzed four paired-end metagenomic libraries with known viral compositions, and one simulated dataset containing viral sequences with bacterial contaminants. The pipeline is available in a public GitHub repository (<https://github.com/MargauxLefebvre/ViroSeek>) and is thoroughly documented to facilitate installation, usage and potential modification by the community.

Materials and methods

Description of the pipeline

ViroSeek is a pipeline designed for the analysis of target-enriched libraries (sequencing libraries generated after an experimental enrichment step for predefined targets), with specific optimization for managing high PCR duplicate rates and performing per-sample assembly (Figure 1). The primary inputs are paired or single-end FASTQ files generated by NGS. The pipeline can be executed using a simple command such as: `nextflow run MargauxLefebvre/ViroSeek -r v0.0.2 --input '/path/to/samples.csv'`, with all parameters set to their default values.

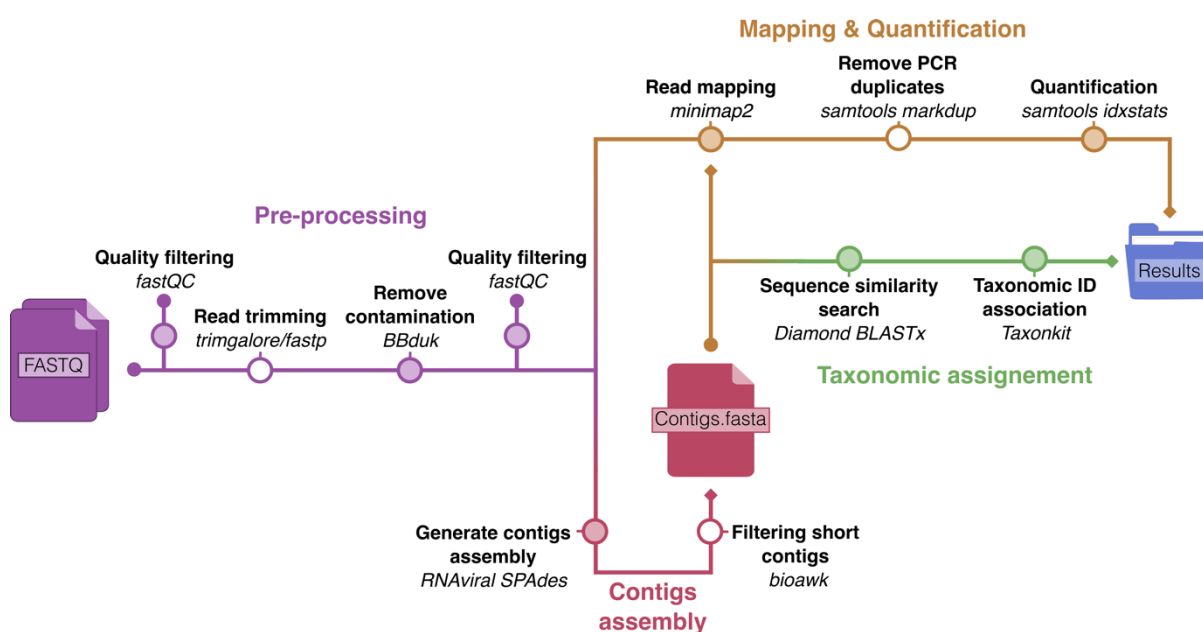


Figure 1 - Simplified workflow of the ViroSeek pipeline. The pipeline comprises four main stages for the processing and analysis of target-enriched libraries: pre-processing, assembly construction, taxonomic assignment, and read mapping for relative quantification. For each step (shown in bold), the corresponding software tool is indicated in italics. The steps marked with a white circle are optional and can be skipped. The version used for each tool of the pipeline is described in Supplementary Table 1.

Raw sequencing reads from the paired-end or single-end FASTQ files are pre-processed using *TrimGalore v0.6.10* (Krueger et al., 2023) or *fastp v0.23.4* (Chen, 2025; Chen et al., 2018) to remove adapter sequences and low-quality bases. Subsequently, potential contamination is filtered out using *BBduk v39.18* (Bushnell, 2026). By default, ViroSeek removes non-viral

ribosomal RNA (16S/18S and 23S/28S) employing the SILVA rRNA database as a reference (Quast et al., 2013, release 138). However, users may substitute this database with an alternative reference better suited to their specific application, such as a host genome. Quality control is performed after each filtering step using *FastQC v0.12.1* (Andrews et al., 2023).

For each sample, de novo assembly is carried out using *SPAdes v4.2.0* (Prjibelski et al., 2020) with the *--rnaviral* option, which is specifically designed for assembling viral RNA-seq data sets. The assembly can be filtered to retain only contigs exceeding a minimum length specified by the user, thereby reducing incorrect taxonomic assignments associated with shorter contigs.

Taxonomic assignment of assembled contigs is performed using *Diamond v2.1.13* (Buchfink et al., 2021) in *BLASTx* mode, which aligns nucleotide contigs against a protein database. By default, ViroSeek uses a database derived from the NCBI reference sequence non-redundant protein database (Pruitt et al., 2005), restricted to viral sequences. However, users may specify an alternative database generated with the *makedb* utility from *Diamond v2.1.13*. The *--range-culling* option is enabled to allow *Diamond* to merge multiple high-scoring segment pairs from the same subject sequence, resulting in a more accurate estimation of total query coverage. The alignment is run in frameshift-aware mode (*-F 15*) to account for potential indels or sequencing errors that might cause frameshifts in the RNA viral genome. Moreover, user-defined thresholds can be specified for the maximum e-value of alignment hits, the minimum percent identity, and the minimum query coverage. In addition, the user can also define the sensitivity mode of *Diamond* (*faster*, *fast*, *mid-sensitive*, *sensitive*, *more-sensitive*, *very-sensitive* or *ultra-sensitive*). Assignment is followed by classification with *Taxonkit v0.9.0* (Shen & Ren, 2021). To estimate the relative viral abundance (Kane et al., 2024; Roux et al., 2017), cleaned reads are mapped to the assembled contigs using *Minimap 2 v2.29* (Li, 2018) and PCR duplicates are removed using the *markdup* function from *Samtools v1.21* (Danecek et al., 2021). However, this *Samtools* function relies exclusively on read position and orientation, which may result in the erroneous classification of genuine signals as duplicates by chance (Danecek et al., 2021). To mitigate this potential bias, users may bypass the deduplication step by specifying the *--skip_dedup* flag.

All essential tools required by the pipeline are encapsulated within a *Singularity* (Kurtzer et al., 2017) or *Docker* (Merkel, 2014) container to ensure full reproducibility across computing environments. The workflow is automatically managed by *Nextflow v24.10.4* (di Tommaso et al., 2017) which orchestrates all processes and handles the retrieval and execution of the container, thereby ensuring seamless portability and reproducibility across platforms.

The pipeline generates a results directory for each sample, named accordingly and organized as follows:

- The *fastQC* directory contains the quality checks for each filtering stage.
- The *assembly_spades* directory contains the final assembly (*contigs.fasta*) not filtered as well as other internal files generated by *SPAdes*, useful mainly for debugging or further exploration of the assembly.
- *_contigs.filtered.fasta*: the final assembly file obtained after applying the length-based filtration step (according to the minimum length parameter).
- *_diamond_blastx.tsv*: Raw output file generated by *Diamond BLASTx*.
- *_contigs_read.tsv*: a tab-delimited summary listing all the contigs
- *_all_viral_taxonomy.txt*: raw tab-delimited file associating each assembled contig with its corresponding viral taxonomic assignment.
- *_taxonomy_viral.clean.txt*: a cleaned and aggregated tab-delimited table reporting read counts per viral taxon to estimate relative abundance.
- *_contigs_reads.tsv*: alignment summary file reporting the number of reads mapped to each contig.
- *_rmrdna.fastq.gz*: filtered read file containing sequences retained after quality filtering, trimming (if enabled), and removal of ribosomal DNA contamination.

Experimental infection and sample preparation

To validate the pipeline, we used four experimentally infected mosquito samples with known viral compositions. Females of *Aedes albopictus* (Strain Ia Lopé, Gabon) were experimentally infected with 1×10^7 PFU/mL for each of the six viruses tested (Table 1) and incubated for 14 days post blood feeding following a previous protocol (Jiolle et al., 2021). After incubation, females were killed and stored at -80°C . Pools of five females per virus were prepared and then used for RNA extraction. Whole mosquito tissues were homogenized in 100 μL of lysis buffer and 1.0 mm glass beads (Biospec products). RNA fragments were purified from each sample using the Nucleospin RNA extraction kit (Macherey-Nagel) according to manufacturer recommendations. Purified RNA was quantified using Qubit HS RNA (Thermo Fisher Scientific).

To simulate realistic metagenomic samples and introduce biological background noise, a separate RNA extract was prepared from a pool of 30 uninfected *Aedes aegypti* females (Paea, strain, Tahiti, French Polynesia). *Aedes aegypti* was chosen as a source of background RNA because large laboratory-reared colonies were readily available, ensuring a homogeneous and infection-free material. This strategy aimed at reproducing the biological noise usually encountered in field-collected samples, where viral RNA typically represents only a minor fraction of total RNA.

Three successive dilutions of the RNA extraction from infected mosquitoes were prepared. The first mixture (Mix A) was obtained by combining 4 μL from each pool corresponding to the six viral infections (Semliki Forest virus, Chikungunya virus, O’nyong-nyong virus, Bozo virus, Wesselsbron virus and West Nile virus). A 1:100 dilution of Mix A generated Mix B, and a subsequent 1:100 dilution of Mix B yielded Mix C. Each of these dilutions was then supplemented with *Aedes aegypti* RNA in order to mimic realistic biological samples (see Supplementary Information 1). This resulted in three final sample types: MixA sample, consisting of 18 μL of Mix A combined with 6 μL of mosquito RNA; MixB sample, containing 18 μL of Mix B and 6 μL of mosquito RNA; and MixC sample, prepared with 18 μL of Mix C and 6 μL of mosquito RNA.

Table 1 - Viruses used for experimental infections, strains and infection rate. The infection rate corresponds to the percentage of positive females by RTqPCR. NA: not available.

Genus	Virus name	Strain	Infection rate
<i>Alphavirus</i>	Semliki Forest virus	NA	80%
	Chikungunya virus	LR2006_OPY	90%
	O’Nyong Nyong virus	ONN229	40%
<i>Orthoflavivirus</i>	Wesselsbron virus	HB83P49	73%
	West Nile virus	ARB2015	80%
<i>Orthobunyavirus</i>	Bozo virus	ARB14898	90%

An additional sample (AltMix sample) consisting of a pool of eight *Ae. albopictus* specimens coinfectd with *Aedes albopictus* densovirus 2 (strain AalDV2) and chikungunya virus were prepared. The presence of both viruses in this pool was confirmed by RT-qPCR (Supplementary information 2). This sample enabled us to assess the pipeline’s capacity to detect this viral mix, including *Aedes albopictus* densovirus 2 (strain AalDV2). While the other viruses tested are RNA viruses, AalDV2 is a single-stranded DNA virus classified within the family *Parvoviridae* and *Brevihamaparvovirus dipteran* 2 species. The preparation of samples and mock viromes using experimental infections is illustrated in Figure 2.

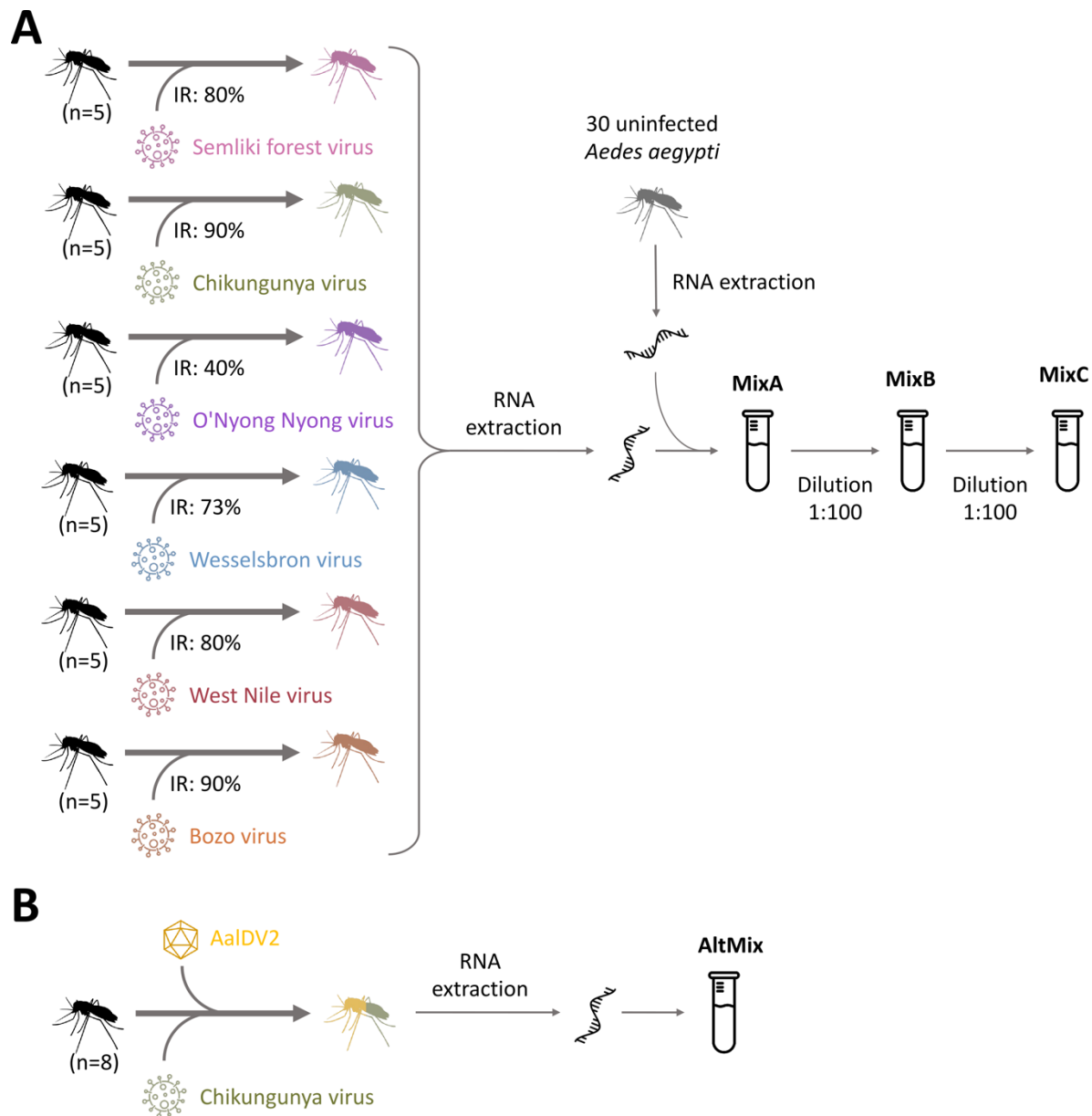


Figure 2 - Simplified preparation of the experimental samples. A. Preparation of samples MixA, MixB and MixC. The mosquitoes (shown in black) are experimentally infected female *Aedes albopictus*. For each experimental infection, the infection rate (IR) is specified. B. Preparation of sample AltMix. The mosquitoes (in black) are experimentally infected female *Aedes albopictus*. AalDV2: *Aedes albopictus* densovirus 2.

Libraries preparation and target enrichment

The RNA was converted to a first-strand cDNA using Superscript III and the second strand was synthesized using NEBNext® Ultra™ II Non-Directional RNA Second Strand Synthesis Module (E6111), followed by a purification using AMPure Xp beads (Beckman Coulter) and Qubit quantification (HS RNA).

For MixA to MixC samples, library preparation was performed following the SureSelect xt Target Enrichment System protocol (Supplementary information 1). Due to issues during library preparation, the following libraries and AltMix samples were performed using Twist protocol (Supplementary information 2). Twist viral probe panel (Twist Bioscience) was designed using CATCH (Metsky et al., 2019) to target 533 virus sequences from 15 families, as described in Supplementary Table 2. After verifying that the amount of DNA was sufficient for all pools, the

pools were clustered to generate 240 million raw reads on Novaseq6000 with 2×150 base pairs (bp) paired-end reads (IntegraGen).

Generation of the simulated data

Simulated paired-end metagenomic reads have been generated using *CAMISIM v1.3* (Fritz et al., 2019) as in Mastriani et al. (2022). This simulation was designed to provide a controlled and reproducible framework to evaluate specific aspects of the pipeline, including the detection of viral sequences at different abundance levels and the ability to filter out non-viral contaminant sequences (Antipov et al., 2020). Three viral sequences were selected: Hepatitis A virus (NC_001489.1) and Ippy virus segment S (NC_007905.1) and L (NC_007906.1). The bacterial genome of *Helicobacter hepaticus* (AE017125.1) was used as a contaminant. The inclusion of a bacterial genome as a dominant component was intended to simulate a challenging background and to assess the robustness of contamination filtering and taxonomic assignment steps. This design does not aim to fully replicate a specific biological system, but rather to test the pipeline under stringent and controlled conditions. We simulated 10 metagenomic samples as paired-end sequencing from Illumina HiSeq 2500 (the ART Illumina model) with a mean fragment length of 270 bp. The abundance distribution was created by sampling from a log-normal distribution with default parameters. Among the ten generated samples, we selected the one that provided the most challenging and informative test conditions (Simulation 9 in Supplementary Table 3). In this sample, bacterial contamination (intended to be filtered out by the pipeline) was predominant. Two viral sequences were present in balanced proportions (Hepatitis A virus and Ippy virus segment S), while a third was included at a lower abundance (Ippy virus segment L).

ViroSeek pipeline execution

Validation of ViroSeek v0.0.2 was performed using one simulated metagenome (SIM sample) and four samples from mosquitoes experimentally infected with known viruses (MixA, MixB, MixC and AltMix samples). The Nextflow pipeline (with *Nextflow v24.10.4*) was executed on all five control samples. Reads pre-processing was performed using *TrimGalore*. To minimize erroneous taxonomic assignments associated with short contigs, a minimum contig length of 140 bp was applied to the assembly. Additionally, stringent alignment parameters were used, including a maximum e-value of $1e-5$, a minimum percent identity of 96.0%, and a minimum query coverage of 51%. Following the *Diamond* documentation, the *--fast* option was used since the minimum percent identity exceeded 90%. These parameters were defined in order to identify all target species, while remaining as stringent as possible to avoid false positives. The database used for taxonomic classification was derived from the National Center for Biotechnology Information Reference Sequence non-redundant protein database (Pruitt et al., 2005) and was not restricted to viral sequences, unlike the default configuration implemented in ViroSeek. We selected a database not restricted to viral sequences in order to facilitate the detection of potential false positives, particularly when comparing results with other pipelines. It should also be noted that lower e-value thresholds are recommended when using smaller databases to reduce the occurrence of false-positive assignments.

Comparison between existing pipelines and ViroSeek

To assess the performance and suitability of ViroSeek, we compared it with several existing pipelines commonly used for viral metagenomic analysis: Taxprofiler v1.2.4 (Stamouli et al., 2023), MetaDenovo v1.3.0 (Lundin et al., 2025), Viralmetagenome v1.1.1 (Klaps et al., 2026), and VirusTaxo v2 (Github last accessed in October 2025; Raju et al., 2022). Comparison with MicroSeek could not be performed because the pipeline's code and workflow are not openly accessible. Furthermore, for PIMGAVir, we were unable to run the pipeline due to excessive RAM requirements.

For each test sample generated in this study, we compared the performance of different pipelines by evaluating computational resource usage (CPU time and maximum memory) and

sensitivity toward various virus species. Taxprofiler, MetaDenovo and Viralmetagenome were executed directly on the raw sequencing reads, whereas VirusTaxo was applied to the assemblies generated by ViroSeek, as described above. For Taxprofiler and MetaDenovo, the same NCBI database as ViroSeek was used, and Diamond served as the taxonomic classifier. Taxprofiler, MetaDenovo and Viralmetagenome are from the nf-core collection of workflows (Ewels et al., 2016), utilising reproducible software environments from the Bioconda (Grüning et al., 2018) and Biocontainers (da Veiga Leprevost et al., 2017) projects. They were executed with *Nextflow* v25.10.0 (di Tommaso et al., 2017).

For Taxprofiler, data was processed using *nf-core/taxprofiler* v1.2.4 (Stamouli et al., 2023), and was configured to perform quality control, including read merging, minimum length filtering at 30 bp, and duplicate removal. Short-read complexity assessment was also enabled. Sequencing quality control and short read preprocessing were respectively performed with *FastQC* (Andrews et al., 2023) and *fastp* (Chen et al. 2018; Chen 2025). Classification was carried out with Diamond (Buchfink et al., 2021). Pipeline results statistics were summarised with *MultiQC* (Ewels et al., 2016).

For MetaDenovo, data was processed using *nf-core/metadenovo* v1.3.0 (<https://doi.org/10.5281/zenodo.10666590>). More precisely in this workflow, raw reads were quality-checked using *FastQC* (Andrews et al., 2023) and trimmed with *TrimGalore* (Krueger et al., 2023). Reads were aligned using *BBMap* (<https://sourceforge.net/projects/bbmap/>), and alignment files were processed with *Samtools* (Danecek et al., 2021). Sequences were filtered to remove potential ribosomal DNA contamination using the same SILVA rRNA database as with ViroSeek (Quast et al., 2013, release 138). De novo assembly was performed using *SPAdes* (Prjibelski et al., 2020). Gene prediction was conducted using *Prodigal* (Hyatt et al., 2010), and taxonomic annotation was performed with *Diamond* (Buchfink et al., 2021), with lineage assignment using *Taxonkit* (Shen & Ren, 2021). Gene abundance was quantified using *featureCounts* from Subread (Yang et al., 2014). Pipeline results statistics were summarised with *MultiQC* (Ewels et al., 2016). For both pipelines, we kept only results with a minimum e-value of 1e-5.

For Viralmetagenome, data was processed using *nf-core/viralmetagenome* v1.1.1 (Klaps et al., 2026). Assembly of reads was conducted using *SPAdes* (Prjibelski et al., 2020), with a minimum contig size threshold of 140 bp. Sequence clustering was performed using the *MMseqs2* clustering method (Kallenborn et al., 2025). Variant calling and annotation steps were omitted from the workflow.

For each pipeline, the parameters were selected to closely align with those used in ViroSeek, thereby ensuring that any observed differences could be attributed primarily to pipeline performance rather than to discrepancies in software configuration.

Post-pipelines downstream analysis

After completion of the pipelines, a post-processing workflow was applied to harmonize taxonomic assignments, generate abundance tables and downstream analyses. All viral taxonomy output files generated (**_taxonomy_viral.clean.txt*) were first collected from individual sample directories and renamed to a standardized krona format. These files were concatenated into a single taxonomy file, allowing a unified representation of viral taxa across samples. Taxonomic tables were then reconstructed in *R* v4.2.3 (R core Team, 2023) and formatted to comply with the International Committee on Taxonomy of Viruses (ICTV) hierarchical structure. Only entries belonging to the Viruses realm were retained for downstream analyses. Read count data were aggregated per taxonomic assignment using a custom Python script modified from Hayer et al. (2023), producing an abundance table summarizing the total number of reads per taxon and per sample. This table was combined with the curated taxonomy and sample metadata to construct a phyloseq object, from package *phyloseq* v1.42.0 (McMurdie & Holmes, 2013), in *R* v4.2.3 (R core Team, 2023). Relative abundances were calculated by normalizing read counts to percentages within each sample. Taxa were agglomerated at the species level, and bar plots were generated to visualize viral community composition with *ggplot2* v3.5.1 (Wickham, 2016). Target viral genera contained in the probes panel were specifically extracted for graphical representation for clarity.

Results

Pipeline validation

The viral panel used in this study includes representative arboviruses of medical relevance as well as a DNA virus belonging to a distinct viral family, allowing evaluation across different genome types commonly encountered in surveillance datasets. Although this benchmarking provides a controlled and realistic evaluation of pipeline performance, it does not encompass the full diversity of known viral taxa, particularly highly divergent or under-represented viruses. Across five control datasets, ViroSeek detected 100% of expected viral species, including low-abundance targets, while maintaining low false positive rates. It outperformed Taxprofiler, MetaDenovo and VirusTaxo in both computational efficiency (approximately 2 and 20 times faster respectively) and sensitivity for virus detection. To validate the ViroSeek pipeline, we designed five samples containing known viral targets, belonging to alphaviruses, orthoflaviviruses, and DNA viruses. These samples covered different levels of complexity, from dilution series of viral mixes (MixA–MixC) to alternative compositions combining RNA and DNA viruses (AltMix, SIM). This setup provided controlled conditions to assess detection sensitivity and specificity.

The analysis found all the viruses expected in MixA (6/6) and MixB (6/6) samples, confirming ViroSeek's ability to detect a wide range of viruses in complex samples (Figure 3). In the most diluted condition (MixC), all expected viruses were still detected (6/6), however fewer species were detected overall and a ~20% reduction in relative abundance of enriched families was observed compared to MixA and MixB, indicating a sensitivity limit of the pipeline under low-abundance conditions. The target viruses were also correctly detected in SIM sample (2/2). In AltMix sample, chikungunya virus was the dominant virus detected and the presence of *Anopheles gambiae* densovirus (AgDV) was also reported (2/2). However, its low abundance prevented direct visualization in the graphs, but it is symbolized by arrows (Figure 3). The underrepresentation of AgDV might be attributed to the biology of densoviruses, whose single-stranded DNA genome is only captured during the replication phase, via the expression of viral transcripts (Kimmick et al., 1998). It is important to note that AgDV belongs to the family *Parvoviridae*, species *Brevihamaparvovirus dipteran 1*, and not to *Brevihamaparvovirus dipteran 2* as reported by ViroSeek. This taxonomic misassignment will be further discussed below. Furthermore, orthoflaviviruses not introduced during experimental infection were detected in MixA sample as Sepik virus, Rift Valley virus, cell fusing agent virus and *Aedes flavivirus*. The identification of Sepik virus and Rift Valley virus corresponds to a taxonomic misassignment, while Cell fusing agent virus and *Aedes flavivirus* are insect-specific flaviviruses commonly found in mosquito samples (Bolling et al., 2012; Fu et al., 2025). With the exception of cell fusing agent virus, which remained detectable in all samples, these other viruses were present at low abundance and were no longer detectable after successive dilution steps. In addition, the pipeline cleaning steps effectively removed contaminating sequences, such as bacterial sequences in SIM sample (Table 2) and host sequences in the other samples (Figure 3).

Two cases of taxonomic misassignment were observed. In AltMix sample, two contigs were assigned to AgDV, while we expected to detect AalDV2. Inspection of the sequences reveals that the strain has been correctly assigned to the capsid protein Q90187.2 (from AalDV2), but this protein has been linked to the wrong species (AgDV). This misassignment is likely due to an inconsistency in the NCBI database, where this sequence is referenced as AgDV rather than AalDV2.

In MixA sample, two contigs were falsely assigned to Sepik virus, a virus not included in the experimental mixture. After verification, the assignment was explained by a protein homology with Wesselsbron virus, preventing reliable discrimination between the two. It was not possible to increase the stringency of the filters without affecting the detection of other target viruses.

Finally, a 1948 bp contig assigned to the Usutu virus was detected in the AltMix sample, although no Usutu virus had been used in the preparation of this library. However, this sample had been prepared in parallel and pooled for sequencing with other Usutu-positive libraries as part of another project on *Culex* mosquitoes. As only a single contig was assigned to this virus, and

considering the library preparation context, this isolated detection suggests potential cross-contamination during preparation or pooling of libraries. This observation does not call into question the efficiency of the pipeline, but rather highlights the importance of rigor in the upstream stages of the experimental protocol, particularly when handling enriched libraries. It also emphasizes the need to interpret certain signals with caution and, when biological confirmation is required, to validate them using complementary approaches.

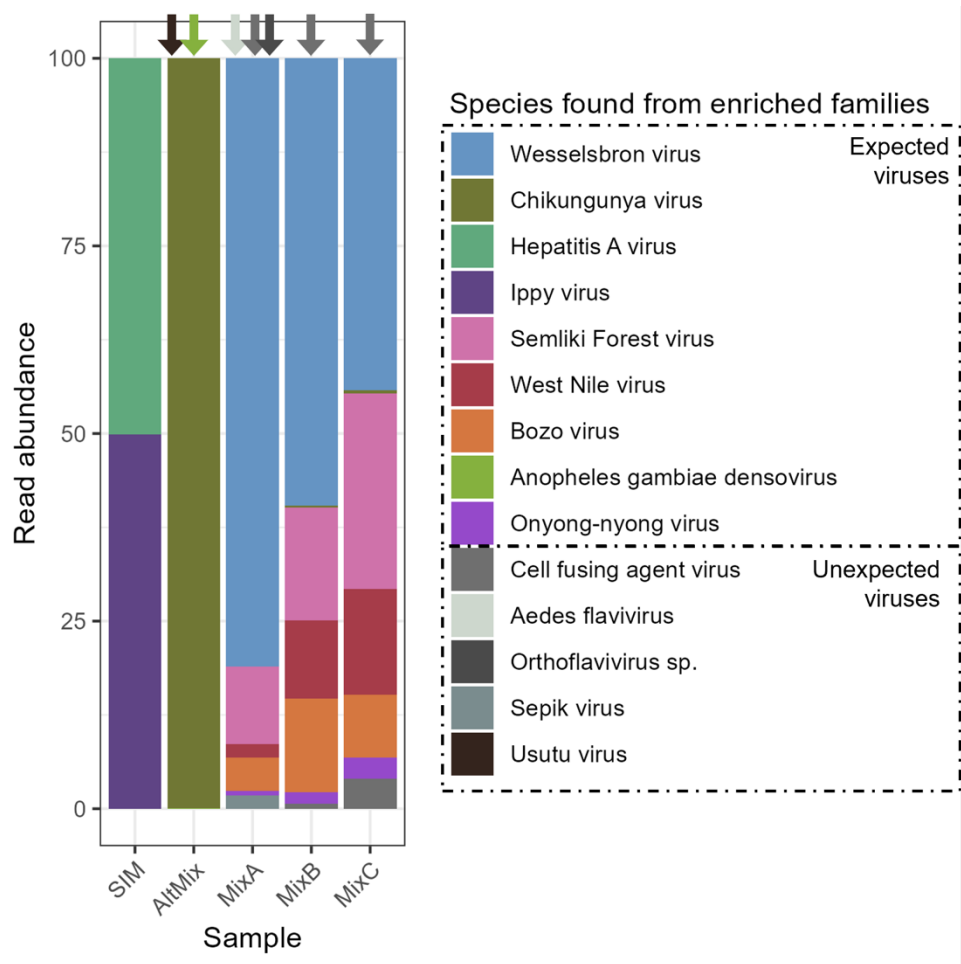


Figure 3 - Viral relative abundance from targeted families. Only viruses belonging to the targeted families are shown, all other viral taxa from non-targeted families are excluded for clarity. “Unexpected viruses” refer to additional viruses detected that were not the species used in the experimental infections but belonged to viral families targeted by the enrichment probe panel (Supplementary Table 2). Abundance corresponds to the percentage of sequencing reads mapped to each viral species. Coloured arrows correspond to viral species detected in the samples but not visible on the graph due to low abundance. *Anopheles gambiae* densovirus (AgDV) is classified as “expected” because sequence verification confirmed that the reported sequence is actually *Aedes albopictus* densovirus, the expected viral species. The discrepancy is due to a reference database misassignment (see Results).

Comparison of ViroSeek efficacy and sensitivity with existing pipelines

To evaluate the efficacy and sensitivity of ViroSeek, we compared its performance with that of existing metagenomic pipelines, focusing on its ability to detect viruses across diverse test samples.

Table 2 - Summary of viral genomes reconstruction for SIM sample. Included reference genome, expected genome size, contig length and corresponding read depth.

Reference genome	Reference length (bp)	Taxonomic assignation	Contig length (bp)	Number of reads mapped
Ippy virus segment S (NC_007905.1)	3366	Ippy virus segment S (YP_516231.1)	3314	370
Ippy virus segment L (NC_007906.1)	7316	Ippy virus segment L (YP_516233.1)	278	3
Hepatitis A virus (NC_001489.1)	7478	Hepatitis A virus (NP_740559.1)	586	25
		Hepatitis A virus (NP_041007.1)	6798	350
<i>Helicobacter hepaticus</i> (AE017125.1)	1799146	-	-	-

Firstly, in terms of computational performance, ViroSeek required approximately 184 CPU hours to process the five test samples, compared to 3,891 for Taxprofiler, 453 for MetaDenovo, and 295 for Viralmetagénome. Regarding memory usage, ViroSeek utilized a maximum of 42.3 GB RAM, whereas Taxprofiler, MetaDenovo, and Viralmetagénome required 40.3, 42, and 72 GB respectively (Figure 4). Thus, ViroSeek was the fastest pipeline, and its memory consumption was equivalent to Taxprofiler and MetaDenovo.

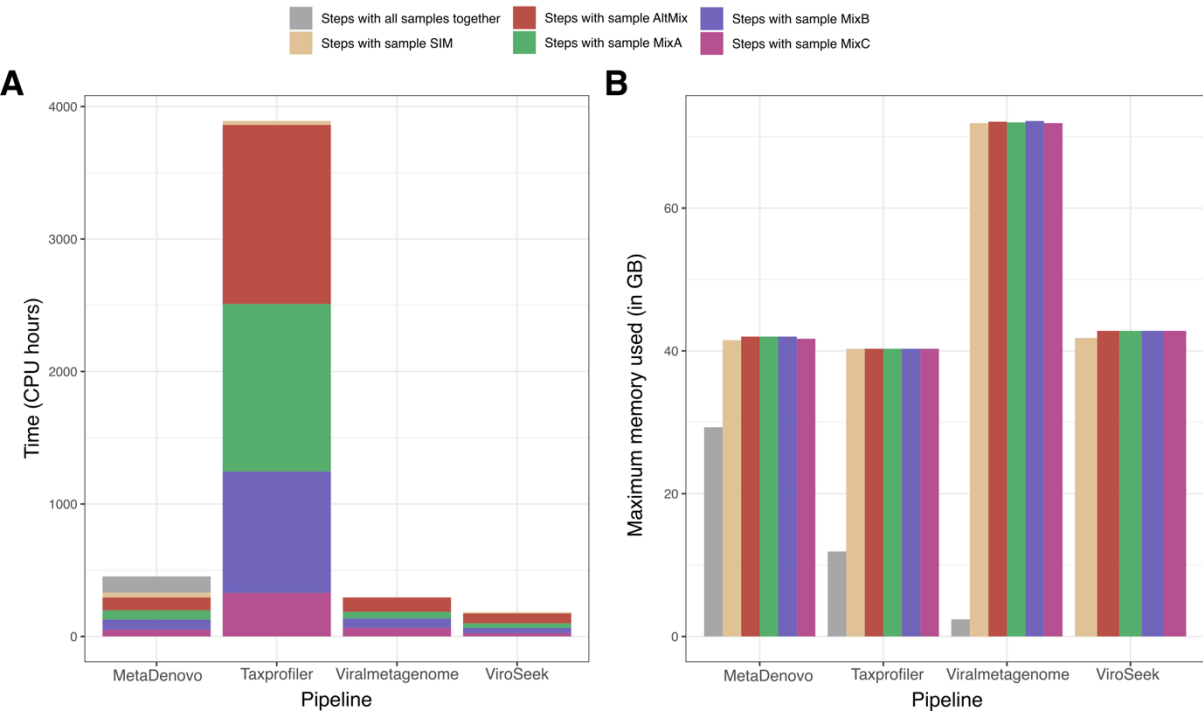


Figure 4 - Comparison of the performance of the different pipelines tested. A. Bar chart showing the computational time required by each pipeline, expressed in CPU hours. B. Bar chart presenting the maximum memory usage (in GB) for each pipeline. In both plots, colors correspond to the individual samples processed. For MetaDenovo, Taxprofiler and Viralmetagénome, some steps operate on all samples together: it is represented in gray.

ViroSeek successfully detected all expected viruses in every dataset, including the simulated sample (SIM), AltMix, and the three mixed samples (with all dilutions from 1 to 1:10000). Similarly, TaxProfiler and Viralmetagene recovered all expected viruses in all datasets (Table 3). Both pipelines correctly identified the expected dengue virus species (*Aedes albopictus* dengue virus 2). However, TaxProfiler additionally assigned a subset of reads to *Anopheles gambiae* dengue virus, reflecting ambiguity in species-level assignment. In contrast, MetaDeNovo showed reduced sensitivity, and detected all expected viruses only in the simulated sample. In AltMix, only Chikungunya virus was detected among the expected viruses, whereas four of the six expected viruses were recovered in MixA and MixB, with Semliki Forest virus and Chikungunya virus consistently missed. In MixC, only three expected viruses were detected, with Semliki Forest virus, Chikungunya virus, and Wesselsbron virus not recovered. VirusTaxo exhibited the lowest sensitivity, detecting only one expected virus in the simulated sample, none in AltMix, and three of the six expected viruses in MixA, B and C.

The number of unexpected viral detections differed substantially among pipelines (Table 3). ViroSeek produced no unexpected detection in the simulated dataset. The number of unexpected species increased in the three mixed samples, reaching 7, 16, 20 and 16 in AltMix, MixA, MixB and MixC, respectively. Aside from the Sepik virus and Rift Valley virus assignment, which has been confirmed to be a false positive, most unexpected viruses detected by ViroSeek corresponded to associated viruses of mosquitoes or viruses plausibly linked to the mosquito microbiome. These included insect-specific viruses such as *Aedes* binegeev-like virus 1, *Culex* picorna-like virus 1, *Aedes* flavivirus and cell-fusing agent virus (Cholleti et al., 2018; Guarido et al., 2021; Hoshino et al., 2009; Parry et al., 2021; Sadeghi et al., 2018). Additional detections included bacteriophages associated with bacterial taxa commonly found in the mosquito microbiome, including members of the *Herelleviridae*, *Ackermanviridae*, *Microviridae* and *Siphonoviridae* families, as well as phages infecting *Klebsiella* spp. or other bacterial groups known to colonize mosquitoes (Barylski et al., 2020; Kang et al., 2020; Mosquera et al., 2023; Nadeem et al., 2026). MetaDenovo generated substantially more unexpected viral detections, with 2, 28, 14, 14 and 37 unexpected species identified in the SIM, AltMix, MixA, MixB, MixC samples, respectively. Several of these assignments corresponded to arboviruses that were not present in the samples, including getah virus, Sindbis virus, Ross River virus, Mayaro virus, Venezuelan equine encephalitis virus and Middleburg virus. Similar patterns were observed with Taxprofiler, which detected 1 unexpected species in the simulated dataset, 40 in AltMix, 60 in MixA, 144 in MixB and 36 in MixC. In the simulated dataset, the *Mammarenavirus* sequence was misclassified as Kitale virus, a related member of the same viral genus. In the experimental datasets, numerous unexpected arboviruses were detected, including Mayaro virus, getah virus, Ross River virus, yellow fever virus or Zika virus.

Although VirusTaxo produced relatively few unexpected species assignments (0 in SIM, 2 in MixB and MixC and 3 in Mix A and C), these results should be interpreted cautiously because the majority of viral reads remained unclassified at species level. More than 90% of viral reads remained unclassified at species resolution. In the simulated dataset, hepatitis A virus accounted for approximately 50% of all reads but could not be assigned at the species level (Figure 5). Several contigs were correctly assigned at the family or genus level but lacked sufficient resolution for species identification, consistent with previous reports (Raju et al., 2022). Viralmetagene detected all expected viruses but generated more unexpected detections than ViroSeek, with 2, 21, 65, 65 and 42 unexpected species in the SIM, AltMix, MixA, MixB and MixC, respectively. Unexpected assignments included several arboviruses, such as West Nile virus, o'nyong-nyong virus, Semliki Forest virus in AltMix and dengue virus, Saint Louis encephalitis virus, yellow fever virus or Ross River virus for MixA, B and C.

Overall, while all pipelines, except MetaDenovo and VirusTaxo, achieved high sensitivity for the detection of expected viruses, important differences were observed in the number and nature of unexpected viral assignments (Figure 5). Unexpected detections generated by ViroSeek were predominantly composed of insect-specific viruses and bacteriophages plausibly associated with the mosquito microbiome, whereas competing pipelines frequently assigned reads to medically

relevant arboviruses that were not present in the samples. These results indicate that ViroSeek maintains high sensitivity while reducing biologically implausible taxonomic assignment relative to the other evaluated pipelines.

Table 3 - Summary of the sensitivity comparison between ViroSeek and the other pipelines evaluated. For each sample, the expected number of viral species was defined according to the experimental and bioinformatic protocols (see Materials and Methods). For each pipeline, we report both the number of expected viral species correctly detected and the total number of unexpected viral species identified. “Unexpected viruses” refer to additional viruses detected that were not the species used in the experimental infections but belonged to viral families targeted by the enrichment probe panel (Supplementary Table 2). “Confirmed false positives” are viral species that are “unexpected viruses” that are not reported to be naturally present in the mosquito virome and were therefore considered confirmed false-positive.

		SIM	AltMix	MixA	MixB	MixC
	Number of expected viruses	2	2	6	6	6
ViroSeek	Number of expected viruses found	2	2	6	6	6
	Number of unexpected viruses	0	7	16	20	16
	Confirmed false positive	0	0	2	0	1
MetaDenovo	Number of expected viruses found	2	1	4	4	3
	Number of unexpected viruses	0	28	14	14	37
	Confirmed false positive	0	13	1	0	2
Taxprofiler	Number of expected viruses found	2	2	6	6	6
	Number of unexpected viruses	1	40	160	144	36
	Confirmed false positive	1	18	79	68	8
VirusTaxo	Number of expected viruses found	1	0	3	3	3
	Number of unexpected viruses	0	2	3	2	2
	Confirmed false positive	0	0	0	0	0
Viralmetagenome	Number of expected viruses found	2	2	6	6	6
	Number of unexpected viruses	2	21	65	65	42
	Confirmed false positive	2	12	23	26	29

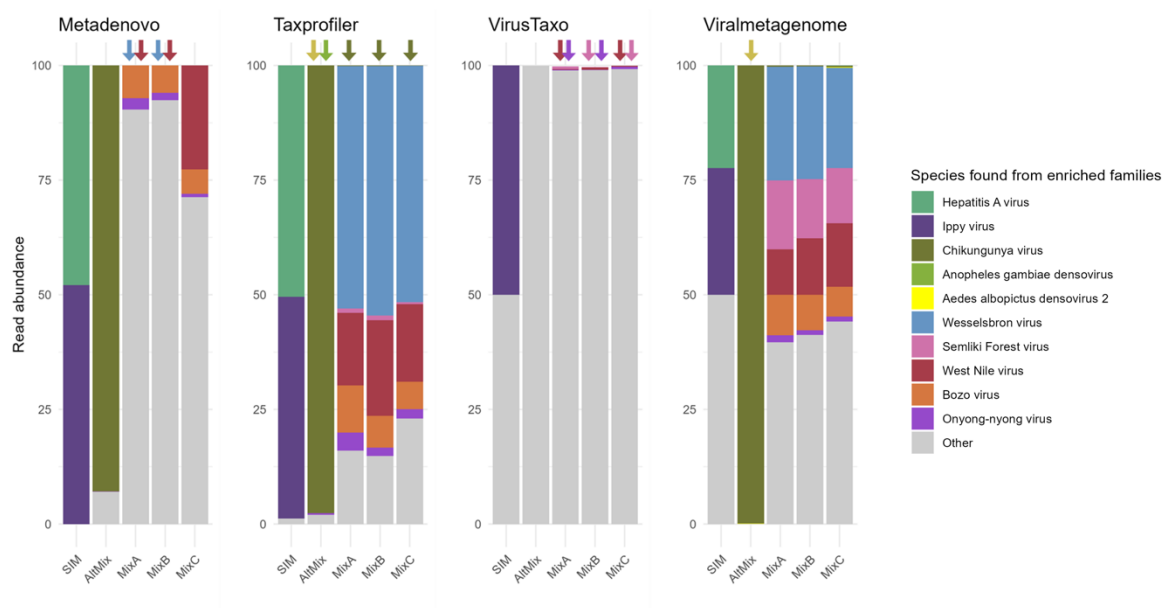


Figure 5 - Viral relative abundance from targeted families across alternative pipelines. Only viruses belonging to the targeted families are shown, all other viral taxa are excluded for clarity. “Other” refers to species not used in the experimental infections but that belonged to viral families targeted by the enrichment probe panel (Supplementary Table 2). Due to the high number of false positives, these signals were grouped into a single category to simplify the interpretation of the figure. Coloured arrows correspond to viral species detected in the samples but not visible on the graph due to low abundance.

Discussion

Several virome analysis pipelines already exist, as detailed in the introduction, but many of them are either highly specialized, complex to install or use, or poorly maintained. ViroSeek addresses this gap by providing a user-friendly workflow optimized for virus-enriched libraries, making it accessible to non-expert users while maintaining robust performance. Beyond producing reliable taxonomic assignments, the pipeline also generates assembled contigs that can be reused for phylogenetic analyses, supporting comprehensive evolutionary and comparative studies (Kane et al., 2024). During validation, ViroSeek consistently identified all expected viruses, even with high viral complexity or low-abundance targets, including single-stranded DNA densoviruses, which are only efficiently detected during the replicative phase when viral transcripts are produced. ViroSeek was not only faster but also provided more accurate taxonomic assignments than the other tools evaluated in this study (Figures 4 and 5). These characteristics make it a particularly relevant and effective tool for the analysis of viral metagenomes, especially in the context of arbovirus research.

In the present study, the interpretation of unexpected viral detections requires caution because the complete virome of whole mosquitoes is not fully characterized. Consequently, unexpected viral species cannot systematically be considered as false positives. Many of the other taxa detected by ViroSeek corresponded to insect-specific viruses previously reported in mosquitoes or to bacteriophages potentially associated with members of the mosquito microbiome. The biological relevance of these detections therefore cannot be excluded. However, we detect only two false positives (Rift Valley virus and Sepik virus) in all samples and one assignment of contigs to incorrect taxa (e.g. AgDV instead of AaIDV2) in AltMix highlights the limitation of current taxonomic databases. Such misassignments are not necessarily attributable to ViroSeek itself, but rather to the lack of sufficiently discriminant markers; when viral sequences are highly conserved between closely related taxa, taxonomic classification becomes uncertain. This shows the importance of manual validation for critical cases, as well as the need for carefully curated databases and the development of more specific probes to improve discrimination between related

viral species. In contrast, several competing pipelines frequently assigned reads to medically important arboviruses that were absent from the experimental design, including yellow fever virus or Zika virus. Such assignments cannot possibly originate from the virome of the mosquitoes and are therefore considered taxonomic classification errors. These results suggest that ViroSeek not only achieves high sensitivity for the detection of expected viruses but also reduces false assignment compared with other evaluated pipelines. This characteristic is particularly important for arbovirus surveillance, where false identification of medically relevant viruses might mislead epidemiological interpretations.

Another critical point highlighted concerns the quality of upstream experimental steps. The detection of a contig assigned to the Usutu virus in a sample that did not experimentally contain it, highlights the risk of cross contamination during library pooling, especially with high viral concentrations. Even if this result does not invalidate the performance of the pipeline, it does illustrate the importance of rigorous control of laboratory practices. ViroSeek's effectiveness has been empirically validated under predefined conditions, in particular with regard to taxonomic assignment parameters such as range culling, minimum query cover and identity thresholds. These parameters have been optimized to guarantee a compromise between sensitivity and accuracy. Reducing these thresholds to increase detection could expose users to an increased risk of false positives or erroneous assignments, particularly in the case of viruses that are genetically close or poorly referenced in databases. We therefore evaluated the impact of the sequence identity used for taxonomic assignment. Lowering the identity threshold from 96% to 90% increased the number of unexpected viral species detected across samples (Supplementary Figure 1). However, this lower threshold did not result in an increase in false positives. Instead, most additional detections corresponded to insect-associated viruses. Because the complete virome of mosquitoes remains incompletely characterized, it is difficult to determine whether these additional assignments reflect real viral diversity that would be missed under more stringent criteria or an increased rate of false positives. Consequently, no definitive conclusion can be drawn regarding the optimal identity threshold for all applications. To accommodate different research objectives, the identity threshold remains configurable by user. However, we recommend testing ViroSeek on control samples with known viral composition to calibrate thresholds and ensure result reliability. In the absence of such controls, the parameters experimentally validated in this study should be used. Furthermore, as demonstrated by our results with Orthoflaviviruses, sample dilution is advisable when the objective is to characterize the overall virome composition, but should be avoided when targeting low-abundance viral families.

ViroSeek has only been evaluated on eukaryotic viruses, although the workflow is compatible with both DNA and RNA viruses. Consequently, its application to bacteriophage or broader bacterial virome studies may be limited, and specialised tools such as VIBRANT (Kieft et al., 2020), VirSorter2 (Guo et al., 2021) or VirFinder (da Veiga Leprevost et al., 2017) may be more appropriate in these contexts. In addition, the pipeline was specifically developed for enriched sequencing libraries, and its performance on shotgun metagenomic datasets has not been extensively assessed. Its sensitivity in untargeted metagenomic applications may be lower than workflows specifically designed for shotgun metagenomics, such as EasyMetagenome (Bai et al., 2025).

Moreover, to maintain an accessible workflow for users with limited bioinformatics expertise, the pipeline implements a single assembler (*rnaviral SPAdes*) and a single taxonomic assignment approach (*BLASTx*). Although alternative assemblers, including other *SPAdes* modes and *PenguiN*, were considered, the *rnaviral SPAdes* was selected because it provides a balanced compromise between assembly quality and robustness for viral datasets. Jochheim et al. (2024) have shown that the *rnaviral* option can recover more complete viral genomes with fewer mismatches compared with other *SPAdes* assembly modes. While *PenguiN* may generate longer contigs, *rnaviral SPAdes* produces more conservative and precise reconstructions (Jochheim et al., 2024). For taxonomic assignment, ViroSeek relies on *BLASTx*, which was selected because it combines the sensitivity of protein-level searches while avoiding the biases introduced by explicit ORF prediction (Buchfink et al., 2021; Pearson, 2014).

One current limitation of ViroSeek concerns the handling of PCR duplicates. By default, duplicate removal is performed using the coordinate-based tool *Samtools markdup*. In the context of viral metagenomic enriched libraries, independent reads sharing identical mapping positions may be incorrectly classified as PCR duplicates and removed. The use of unique molecular identifiers (UMI) during library preparation can help overcome this issue by enabling the identification of true amplification-derived duplicates in the downstream analysis. Incorporating support for UMI-based deduplication, for example with UMI-tools (Smith et al., 2017), represents a potential direction for future development of ViroSeek. In the meantime, users may avoid biases associated with *Samtools markdup* by running the pipeline with the `--skip_dedup` option.

In conclusion, ViroSeek provides a user-friendly and reproducible workflow for the analysis of target-enriched metagenomic sequencing data. Across all datasets, it accurately detected the expected viruses, including low-abundance and phylogenetically diverse strains, with fewer erroneous taxonomic assignments compared with other commonly used pipelines. In addition to taxonomic identification, the generation of assembled contigs enables downstream phylogenetic and comparative analyses. By facilitating standardized and reproducible virome analyses, ViroSeek has the potential to support arbovirus surveillance and studies of viral diversity across a wide range of ecological and epidemiological contexts.

Acknowledgment

The authors acknowledge the ISO 9001-certified IRD i-Trop HPC (member of the South Green Platform) at IRD Montpellier for providing HPC resources that contributed to the research results reported in this paper. URL: <http://www.southgreen.fr>. The bioinformatics analyses were also performed on the Core Cluster of the Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013). We would also like to thank James Fellows Yates for his assistance in addressing the resource limitations encountered with Taxprofiler. We would like to thank Mathilde Jaquet for helping us with the phyloseq package. Mosquitoes were reared and tested thanks to the Vectopole platform dedicated to vectors at IRD.

A preprint version 2 of this article has been peer-reviewed and recommended by PCI Genomics (<https://doi.org/10.24072/pci.genomics.100695>; Molik, 2026).

Funding

This study received financial support from the Agence Nationale de la Recherche (ANR PRC TIGERBRIDGE, grant n° 16-CE35-0010-01 to CP). AB has been granted an IRD PhD Fellowship.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Data and code availability

The data for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB93907.

All the scripts, the related metadata and documentation are available in the DataSuds repository (IRD, France) at <https://doi.org/10.23708/LTXIKA>. Related documentation is openly available and granted under CC-BY license.

The pipeline is also available in a GitHub repository: <https://github.com/MargauxLefebvre/ViroSeek>.

References

- de Almeida, J. P., Aguiar, E. R., Armache, J. N., Olmo, R. P., & Marques, J. T. (2021). The virome of vector mosquitoes. *Current Opinion in Virology*, 49, 7–12. <https://doi.org/10.1016/j.coviro.2021.04.002>
- de Almeida, I. F. de, Codeço, C. T., Lana, R. M., Bastos, L. S., Oliveira, S. de S., Ferreira, D. A. da C., Godinho, V. B., Riback, T. I. S., Cruz, O. G., & Coelho, F. C. (2023). The expansion of chikungunya in Brazil. *The Lancet Regional Health – Americas*, 25. <https://doi.org/10.1016/j.lana.2023.100571>
- Andrews, S., Biggins, L., Inglesfield, S., Carr, H., & Montgomery, J. (2023, March 1). Babraham Bioinformatics --FastQC A Quality Control tool for High Throughput Sequence Data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Antipov, D., Raiko, M., Lapidus, A., & Pevzner, P. A. (2020). MetaviralSPAdes: Assembly of viruses from metagenomic data. *Bioinformatics*, 36(14), 4126–4129. <https://doi.org/10.1093/bioinformatics/btaa490>
- Bai, D., Chen, T., Xun, J., Ma, C., Luo, H., Yang, H., Cao, C., Cao, X., Cui, J., Deng, Y.-P., Deng, Z., Dong, W., Dong, W., Du, J., Fang, Q., Fang, W., Fang, Y., Fu, F., Fu, M., ... Liu, Y.-X. (2025). EasyMetagenome: A user-friendly and flexible pipeline for shotgun metagenomic analysis in microbiome research. *iMeta*, 4(1), e70001. <https://doi.org/10.1002/imt2.70001>
- Baker, R. E., Mahmud, A. S., Miller, I. F., Rajeev, M., Rasambainarivo, F., Rice, B. L., Takahashi, S., Tatem, A. J., Wagner, C. E., Wang, L.-F., Wesolowski, A., & Metcalf, C. J. E. (2022). Infectious disease in an era of global change. *Nature Reviews Microbiology*, 20(4), 193–205. <https://doi.org/10.1038/s41579-021-00639-z>
- Barylski, J., Kropinski, A. M., Alikhan, N.-F., Adriaenssens, E. M., & ICTV Report Consortium. (2020). ICTV Virus Taxonomy Profile: Herelleviridae. *Journal of General Virology*, 101(4), 362–363. <https://doi.org/10.1099/jgv.0.001392>
- Bolling, B. G., Olea-Popelka, F. J., Eisen, L., Moore, C. G., & Blair, C. D. (2012). Transmission dynamics of an insect-specific flavivirus in a naturally infected *Culex pipiens* laboratory colony and effects of co-infection on vector competence for West Nile virus. *Virology*, 427(2), 90–97. <https://doi.org/10.1016/j.virol.2012.02.016>
- Brandt, C., Krautwurst, S., Spott, R., Lohde, M., Jundzill, M., Marquet, M., & Hölzer, M. (2021). poreCov-An Easy to Use, Fast, and Robust Workflow for SARS-CoV-2 Genome Reconstruction via Nanopore Sequencing. *Frontiers in Genetics*, 12, 711437. <https://doi.org/10.3389/fgene.2021.711437>
- Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4), 366–368. <https://doi.org/10.1038/s41592-021-01101-x>
- Bushnell, B. (2026). BBMap [Computer software]. <https://sourceforge.net/projects/bbmap/>
- Chen, S. (2025). fastp 1.0: An ultra-fast all-round tool for FASTQ data quality control and preprocessing. *iMeta*, 4(5), e70078. <https://doi.org/10.1002/imt2.70078>
- Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
- Cholleti, H., Hayer, J., Fafetine, J., Berg, M., & Blomström, A.-L. (2018). Genetic characterization of a novel picorna-like virus in *Culex* spp. Mosquitoes from Mozambique. *Virology Journal*, 15(1), 71. <https://doi.org/10.1186/s12985-018-0981-z>
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), giab008. <https://doi.org/10.1093/gigascience/giab008>
- Dong, X., & Soong, L. (2021). Emerging and Re-emerging Zoonoses are Major and Global Challenges for Public Health. *Zoonoses*, 1, 999. <https://doi.org/10.15212/ZOONOSES-2021-0001>

- Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>
- Finkel, Y., Stern-Ginossar, N., & Schwartz, M. (2018). Viral Short ORFs and Their Possible Functions. *PROTEOMICS*, 18(10), 1700255. <https://doi.org/10.1002/pmic.201700255>
- Fritz, A., Hofmann, P., Majda, S., Dahms, E., Dröge, J., Fiedler, J., Lesker, T. R., Belmann, P., DeMaere, M. Z., Darling, A. E., Sczyrba, A., Bremges, A., & McHardy, A. C. (2019). CAMISIM: Simulating metagenomes and microbial communities. *Microbiome*, 7(1), 17. <https://doi.org/10.1186/s40168-019-0633-6>
- Fu, Y., Zhao, W., Wu, S., Li, J., Liu, Q., Jiang, F., Lu, H., Kang, L., Xia, Q., & Cui, F. (2025). Risk assessment of an Aedes flavivirus and its effect on pathogenic flavivirus replication in mosquitoes. *Parasites & Vectors*, 18, 88. <https://doi.org/10.1186/s13071-025-06711-4>
- Grandien, M., & Svedmyr, A. (1983). 20 --Laboratory techniques for diagnosis of virus infections. In E. Lycke & E. Norrby (Eds), *Textbook of Medical Virology* (pp. 189–201). Butterworth-Heinemann. <https://doi.org/10.1016/B978-0-407-00253-1.50025-6>
- Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7), 475–476. <https://doi.org/10.1038/s41592-018-0046-7>
- Guarido, M. M., Govender, K., Riddin, M. A., Schrama, M., Gorsich, E. E., Brooke, B. D., Almeida, A. P. G., & Venter, M. (2021). Detection of Insect-Specific Flaviviruses in Mosquitoes (Diptera: Culicidae) in Northeastern Regions of South Africa. *Viruses*, 13(11), 2148. <https://doi.org/10.3390/v13112148>
- Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., Pratama, A. A., Gazitúa, M. C., Vik, D., Sullivan, M. B., & Roux, S. (2021). VirSorter2: A multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*, 9(1), 37. <https://doi.org/10.1186/s40168-020-00990-y>
- Hayer, J., Dainat, J., Marcy, E., & Bañuls, A.-L. (2023). Baargin: A Nextflow workflow for the automatic analysis of bacterial genomics data with a focus on Antimicrobial Resistance. *Journal of Open Source Software*, 8(90), 5397. <https://doi.org/10.21105/joss.05397>
- Hoshino, K., Isawa, H., Tsuda, Y., Sawabe, K., & Kobayashi, M. (2009). Isolation and characterization of a new insect flavivirus from Aedes albopictus and Aedes flavopictus mosquitoes in Japan. *Virology*, 391(1), 119–129. <https://doi.org/10.1016/j.virol.2009.06.025>
- Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1), 119. <https://doi.org/10.1186/1471-2105-11-119>
- Jiollé, D., Moltini-Conclois, I., Obame-Nkoghe, J., Yangari, P., Porciani, A., Scheid, B., Kengne, P., Ayala, D., Failloux, A.-B., & Paupy, C. (2021). Experimental infections with Zika virus strains reveal high vector competence of Aedes albopictus and Aedes aegypti populations from Gabon (Central Africa) for the African virus lineage. *Emerging Microbes & Infections*, 10(1), 1244–1253. <https://doi.org/10.1080/22221751.2021.1939167>
- Jochheim, A., Jochheim, F. A., Kolodyazhnaya, A., Morice, É., Steinegger, M., & Söding, J. (2024). Strain-resolved de-novo metagenomic assembly of viral genomes and microbial 16S rRNAs. *Microbiome*, 12(1), 187. <https://doi.org/10.1186/s40168-024-01904-y>
- Kaiser, N. L., Groschup, M. H., & Sadeghi, B. (2025). Identification of bioinformatic pipelines for virus monitoring using nanopore sequence data: A systematic assessment. *Journal of Virological Methods*, 336, 115153. <https://doi.org/10.1016/j.jviromet.2025.115153>
- Kallenborn, F., Chacon, A., Hundt, C., Sirelkhatim, H., Didi, K., Cha, S., Dallago, C., Mirdita, M., Schmidt, B., & Steinegger, M. (2025). GPU-accelerated homology search with MMseqs2. *Nature Methods*, 22(10), 2024–2027. <https://doi.org/10.1038/s41592-025-02819-8>
- Kane, Y., Tendu, Alexander, Li, Ruiya, Chen, Yanhua, Mastriani, Emilio, Lan, Jiaming, Catherine Hughes, Alice, Berthet, Nicolas, & Wong, G. (2024). Viral diversity in wild and urban rodents of Yunnan Province, China. *Emerging Microbes & Infections*, 13(1), 2290842. <https://doi.org/10.1080/22221751.2023.2290842>

- Kang, X., Wang, Y., Li, S., Sun, X., Lu, X., Rajaofera, M. J. N., Lu, Y., Kang, L., Zheng, A., Zou, Z., & Xia, Q. (2020). Comparative Analysis of the Gut Microbiota of Adult Mosquitoes From Eight Locations in Hainan, China. *Frontiers in Cellular and Infection Microbiology*, 10. <https://doi.org/10.3389/fcimb.2020.596750>
- Kieft, K., Zhou, Z., & Anantharaman, K. (2020). VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, 8(1), 90. <https://doi.org/10.1186/s40168-020-00867-0>
- Kimmick, M. W., Afanasiev, B. N., Beaty, B. J., & Carlson, J. O. (1998). Gene Expression and Regulation from the p7 Promoter of Aedes Densonucleosis Virus. *Journal of Virology*, 72(5), 4364–4370. <https://doi.org/10.1128/jvi.72.5.4364-4370.1998>
- Klaps, J., Lemey, P., Bletsa, M., nf-core community, & Kafetzopoulou, L. E. (2026). nf-core/viralmetagenome: A novel pipeline for untargeted viral genome reconstruction. *Bioinformatics*, 42(5), btag187. <https://doi.org/10.1093/bioinformatics/btag187>
- Krueger, F., James, F., Ewels, P., Afyounian, E., Weinstein, M., Schuster-Boeckler, B., Hulselmans, G., & sclamons. (2023). FelixKrueger/TrimGalore: V0.6.10 - add default decompression path (Version 0.6.10) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.7598955>
- Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12(5): e0177459. <https://doi.org/https://doi.org/10.1371/journal.pone.0177459>
- Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Lundin, D., Leo, D. D., emnilsson, tfalkarkea, bot, nf-core, Jaber, M., Lecomber, D., Campos, H. G. de, Miller, E., Yates, J. A. F., & Garcia, M. U. (2025). nf-core/metatdenovo: Nf-core/metatdenovo 1.2.0 [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.15687194>
- Mastriani, E., Bienes, K. M., Wong, G., & Berthet, N. (2022). PIMGAVir and Vir-MinION: Two Viral Metagenomic Pipelines for Complete Baseline Analysis of 2nd and 3rd Generation Data. *Viruses*, 14(6), 1260. <https://doi.org/10.3390/v14061260>
- McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE*, 8(4), e61217. <https://doi.org/10.1371/journal.pone.0061217>
- Merkel, D. (2014). Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux Journal*, 2014(239), Article 2.
- Metsky, H. C., Siddle, K. J., Gladden-Young, A., Qu, J., Yang, D. K., Brehio, P., Goldfarb, A., Piantadosi, A., Wohl, S., Carter, A., Lin, A. E., Barnes, K. G., Tully, D. C., Corleis, B., Hennigan, S., Barbosa-Lima, G., Vieira, Y. R., Paul, L. M., Tan, A. L., ... Matranga, C. B. (2019). Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. *Nature Biotechnology*, 37(2), 160–168. <https://doi.org/10.1038/s41587-018-0006-x>
- Molik, D. (Ed.). (2026). Reproducible viral detection in difficult sequencing data. *Peer Community in Genomics*, 100695. <https://doi.org/10.24072/pci.genomics.100695>
- Mosquera, K. D., Khan, Z., Wondwosen, B., Alsanius, B., Hill, S. R., Ignell, R., & Lorenzo, M. G. (2023). Odor-mediated response of gravid Aedes aegypti to mosquito-associated symbiotic bacteria. *Acta Tropica*, 237, 106730. <https://doi.org/10.1016/j.actatropica.2022.106730>
- Nadeem, S. A., Ali, I., Hussain, H., Ullah, I., Ali, W., Alzahrani, K. J., Ali, H., Khan, Z. I., Abass, K. S., & Rahman, R. ur. (2026). Metagenomic analysis of bacterial and viral communities of Aedes aegypti and Aedes albopictus. *Journal of Genetic Engineering and Biotechnology*, 24(1), 100643. <https://doi.org/10.1016/j.jgeb.2025.100643>
- Pandolfo, M., Telatin, A., Lazzari, G., Adriaenssens, E. M., & Vitulo, N. (2022). MetaPhage: An Automated Pipeline for Analyzing, Annotating, and Classifying Bacteriophages in Metagenomics Sequencing Data. *mSystems*, 7(5), e00741-22. <https://doi.org/10.1128/msystems.00741-22>

- Parry, R., James, M. E., & Asgari, S. (2021). Uncovering the Worldwide Diversity and Evolution of the Virome of the Mosquitoes *Aedes aegypti* and *Aedes albopictus*. *Microorganisms*, 9(8), 1653. <https://doi.org/10.3390/microorganisms9081653>
- Pearson, W. R. (2014). BLAST and FASTA Similarity Searching for Multiple Sequence Alignment. In *Methods in Molecular Biology* (pp. 75–101). Humana Press. https://doi.org/10.1007/978-1-62703-646-7_5
- Pérot, P., Bigot, T., Temmam, S., Regnault, B., & Eloit, M. (2022). Microseek: A Protein-Based Metagenomic Pipeline for Virus Diagnostic and Discovery. 14(9), 1990. <https://doi.org/10.3390/v14091990>
- Prijbelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics*, 70(1), e102. <https://doi.org/10.1002/cpbi.102>
- Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(suppl_1), D501–D504. <https://doi.org/10.1093/nar/gki025>
- Quast, C., Priesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1), D590–D596. <https://doi.org/10.1093/nar/gks1219>
- Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Bore, J. A., Koundouno, R., Dudas, G., Mikhail, A., Ouédraogo, N., Afrough, B., Bah, A., Baum, J. H. J., Becker-Ziaja, B., Boettcher, J. P., Cabeza-Cabrerizo, M., Camino-Sánchez, Á., Carter, L. L., ... Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589), 228–232. <https://doi.org/10.1038/nature16996>
- R core Team. (2023). R: The R Project for Statistical Computing [Computer software]. R foundation for statistical computing. <https://www.r-project.org/>
- Raju, R. S., Al Nahid, A., Chondrow Dev, P., & Islam, R. (2022). VirusTaxo: Taxonomic classification of viruses from the genome sequence using k-mer enrichment. *Genomics*, 114(4), 110414. <https://doi.org/10.1016/j.ygeno.2022.110414>
- Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). VirFinder: A novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, 5(1), 69. <https://doi.org/10.1186/s40168-017-0283-5>
- Roux, S., Emerson, J. B., Elie-Fadrosh, E. A., & Sullivan, M. B. (2017). Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ*, 5, e3817. <https://doi.org/10.7717/peerj.3817>
- Sadeghi, M., Altan, E., Deng, X., Barker, C. M., Fang, Y., Coffey, L. L., & Delwart, E. (2018). Virome of > 12 thousand *Culex* mosquitoes from throughout California. *Virology*, 523, 74–88. <https://doi.org/10.1016/j.virol.2018.07.029>
- Shen, W., & Ren, H. (2021). TaxonKit: A practical and efficient NCBI taxonomy toolkit. *Journal of Genetics and Genomics, Special Issue on Microbiome*, 48(9), 844–850. <https://doi.org/10.1016/j.jgg.2021.03.006>
- Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, 27(3), 491–499. <https://doi.org/10.1101/gr.209601.116>
- Stamouli, S., Beber, M. E., Normark, T., Christensen, T. A., Andersson-Li, L., Borry, M., Jamy, M., Community, N.-C., & Yates, J. A. F. (2023). nf-core/taxprofiler: Highly parallelised and flexible pipeline for metagenomic taxonomic classification and profiling (p. 2023.10.20.563221). *bioRxiv*. <https://doi.org/10.1101/2023.10.20.563221>
- di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4), 316–319. <https://doi.org/10.1038/nbt.3820>
- da Veiga Leprevost, F., Grüning, B. A., Alves Aflitos, S., Röst, H. L., Uszkoreit, J., Barsnes, H., Vaudel, M., Moreno, P., Gatto, L., Weber, J., Bai, M., Jimenez, R. C., Sachsenberg, T., Pfeuffer,

- J., Vera Alvarez, R., Griss, J., Nesvizhskii, A. I., & Perez-Riverol, Y. (2017). BioContainers: An open-source and community-driven framework for software standardization. *Bioinformatics*, 33(16), 2580–2582. <https://doi.org/10.1093/bioinformatics/btx192>
- Wickham, H. (2016). Data Analysis. In H. Wickham (Ed.), *Ggplot2: Elegant Graphics for Data Analysis* (pp. 189–201). Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4_9
- Yang, L., Smyth, G., & Shi, wei. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7), 923–930. <https://doi.org/10.1093/bioinformatics/btt656>