Bioinformatics, 2025, pp. 1–5

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

nf-core/viralmetagenome: A Novel Pipeline for Untargeted Viral Genome Reconstruction

Joon Klaps[1,*] Philippe Lemey[1] nf-core community[2] and Liana Eleni Kafetzopoulou[1]

[1]Rega Institute for Medical Research, Department of Microbiology, Immunology and Transplantation, KU Leuven, Herestraat 49, 3000, Leuven, Belgium and [2]A full list of contributors can be found at https://nf-co.re/community.
[*]Corresponding author. joon.klaps@kuleuven.be
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Motivation: Eukaryotic viruses present significant challenges for genome reconstruction and variant analysis due to their extensive diversity and potential genome segmentation. While de novo assembly followed by reference database matching and scaffolding is a commonly used approach, the manual execution of this workflow is extremely time-consuming, particularly due to the extensive reference curation required. Here, we address the critical need for an automated, scalable pipeline that can efficiently handle viral metagenomic analysis without manual intervention.
Results: We present nf-core/viralmetagenome, a comprehensive viral metagenomic pipeline for untargeted genome reconstruction and variant analysis of eukaryotic DNA and RNA viruses. Viralmetagenome is implemented as a Nextflow workflow that processes short-read metagenomic samples to automatically detect and assemble viral genomes, while also performing variant analysis. The pipeline features automated reference selection, consensus quality control metrics, comprehensive documentation, and seamless integration with containerization technologies, including Docker and Singularity. We demonstrate the utility and accuracy of our approach through validation on both simulated and real datasets, showing robust performance across diverse viral families in metagenomic samples.
Availability: nf-core/viralmetagenome is freely available at https://github.com/nf-core/viralmetagenome with comprehensive documentation at https://nf-co.re/viralmetagenome
Contact: joon.klaps@kuleuven.be
Supplementary information: Supplementary data are available at https://github.com/Joon-Klaps/nf-core-viralmetagenome-manuscript online.

Key words: viralmetagenome, bioinformatic pipeline, nextflow, viral metagenomics, viral assembly, viral variant analysis

Introduction

Reconstructing viral genomes from metagenomic sequencing data presents considerable computational challenges, particularly for viruses exhibiting extensive genetic diversity Baaijens et al., Deng et al., Meleshko et al.. This diversity is further compounded by segmented genomes in families like influenza, rotavirus, and bunyaviruses, where individual segments can evolve under distinct selective pressures and reassort, contributing to a complex landscape for genome reconstruction. While pipelines are often designed to target specific viruses and their subtypes Shepard et al., accurate and complete genome reconstruction of samples with unknown references typically requires manual curation of contigs and reference matching Tomkins-Tinch et al., Li et al. [b]. This manual curation process is time-consuming, making it impractical for large-scale metagenome studies or rapid response scenarios that involve emerging viral outbreaks of unknown origin.

To address these limitations, we developed nf-core/viralmetagenome, a comprehensive pipeline specifically designed for untargeted viral genome reconstruction. The pipeline is developed using Nextflow Di Tommaso et al. within the nf-core framework Ewels et al. [b], ensuring reproducibility through containerization with Docker Merkel and Singularity Kurtzer et al., and enabling portability across computational platforms such as local desktops, high-performance clusters and cloud environments.

Pipeline Description

nf-core/viralmetagenome implements an automated workflow performing *de novo* assembly, reference matching, and iterative consensus refinement for the reconstruction of viral genomes without prior target knowledge. The pipeline consists of

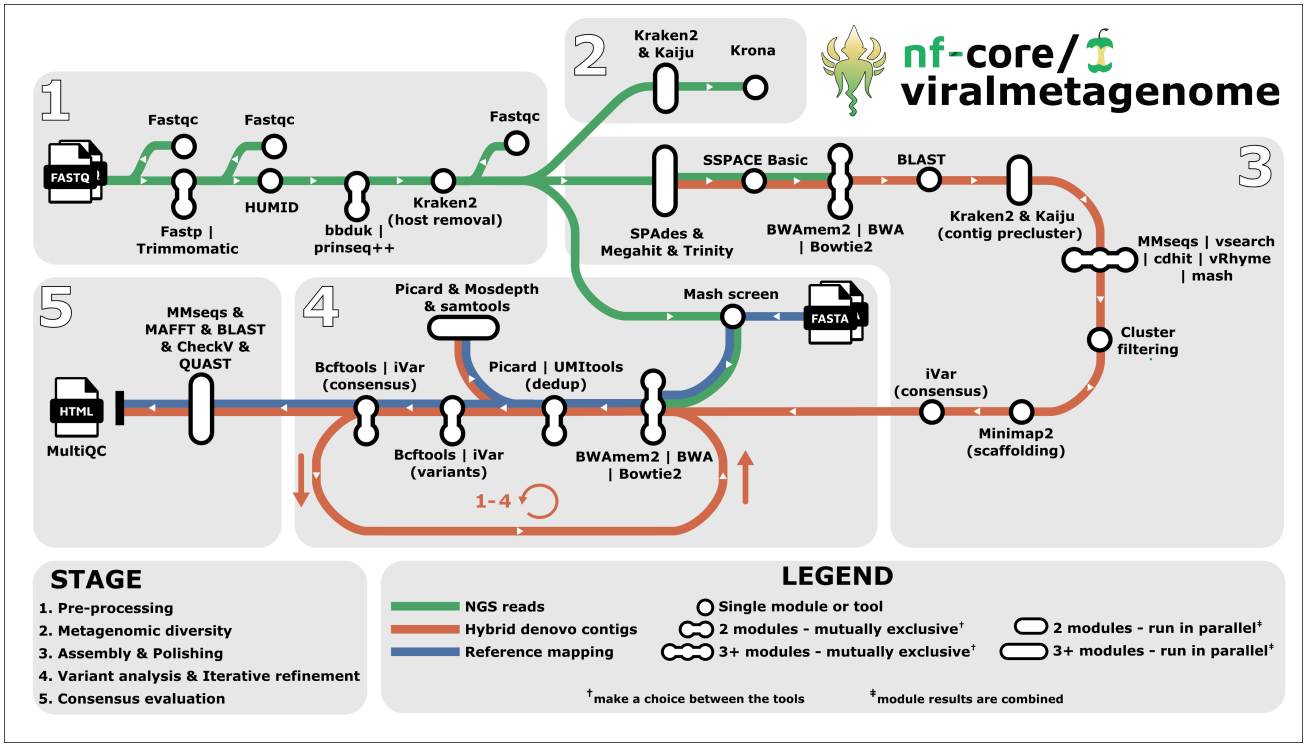

Fig. 1. Visual overview of the nf-core/viralmetagenome pipeline for untargeted viral genome reconstruction. nf-core/viralmetagenome processes short-read data through pre-processing, metagenomic diversity assessment, *de novo* assembly with multiple assemblers, scaffolding with automated reference identification and contig taxonomy-guided clustering, and iterative consensus refinement through read mapping and variant calling. Quality control metrics, assembly statistics, and coverage data are integrated into interactive MultiQC reports and standardised overview tables for downstream analysis.

five major stages: read preprocessing, metagenomic diversity assessment, contig assembly and scaffolding, iterative consensus refinement with variant analysis, and quality control (Figure 1). While this manuscript highlights key differences between particular tools, the pipeline offers multiple options to accommodate established user workflows and preferences. Tool details are in Supplementary Table 1.

Read preprocessing

Input reads provided via sample sheets containing sample names and short-read FASTQ paths are preprocessed using FastQC and have their adapter trimmed with Fastp Chen et al. (default) or Trimmomatic Bolger et al.. Fastp is overall faster and has automated adapter detection and trimming Chen et al.. UMI deduplication is implemented using HUMID Laros and UMI-tools Smith et al. once reads are mapped to a reference. Multiple sequencing runs are merged after trimming by specifying merge group identifiers in the sample sheet. Complexity filtering with BBduk Bushnell or prinseq++ Cantu et al. removes low-complexity sequences containing repetitive elements. Host removal uses Kraken2 Wood et al. against user-specified databases (default: human genome subset). However, users are encouraged to employ more comprehensive databases, including complete host genome and transcriptome (human and otherwise), common sequencer contaminants, and bacterial genomes, to ensure thorough decontamination Forbes et al..

Metagenomic diversity assessment

Taxonomic classification of preprocessed reads is performed using two complementary approaches - Kaiju Menzel et al. and Kraken2 Wood et al. - to maximise detection sensitivity across diverse viral families. Results from both classifiers are visualised using Krona Ondov et al. [a].

De novo assembly and clustering

The assembly workflow implements multi-assembler approaches followed by clustering and scaffolding. *De novo* assembly is performed using SPAdes Meleshko et al. (RNAviral mode), MEGAHIT Li et al. [a], and Trinity Grabherr et al., capitalizing on distinct algorithmic strengths to maximise genome recovery across diverse viral families and variable read depths. Contigs can optionally be extended with SSPACE Basic Boetzer et al..

Reference identification uses BLASTn Altschul et al. against the Reference Viral Database (RVDB) Goodacre et al., retaining top five hits to facilitate identification of related genomic segments and appropriate reference sequences for contig scaffolding and clustering.

Clustering employs two stages: taxonomic pre-clustering using Kraken2 Wood et al. and Kaiju Menzel et al.. For more efficient targeted analyses, the user can opt to focus on specific taxonomic clades. Subsequent nucleotide similarity clustering is done with CD-HIT-EST Li and Godzik, VSEARCH Rognes et al., MMseqs2 Steinegger and Söding, vRhyme Kieft et al. [a], or Mash Ondov et al. [b]. All tools are valid options, though performance may vary depending on the dataset; for comprehensive benchmarking, we refer to Zielezinski et al.

Zielezinski et al. and Steinegger and Söding Steinegger and Söding.

As an optional filtering step of contig clusters, after assembly and extension, reads can be mapped to all contigs using BWAmem2 Vasimuddin et al. (default), BWA Li [a], or Bowtie2 Langmead et al.. Clusters are filtered based on the cumulated percentage of reads mapped to the contigs of a cluster. By filtering clusters, low-coverage assemblies can be identified that likely represent assembly artefacts.

For the final scaffolding step, all cluster members are mapped to the cluster representative or centroid using Minimap2 Li [b], followed by consensus calling with iVar Grubaugh et al. to generate reference-assisted assemblies. Regions with zero coverage depth can optionally be represented by the reference genome to produce a more complete scaffold genome for consensus calling.

Iterative consensus refinement and variant calling

The consensus module supports external reference-based analysis and scaffold refinement. Users can provide a separate reference genome or reference set for each sample with `--mapping_constraints`; when a reference set is provided, the most similar can be selected using Mash Ondov et al. [b].

Scaffold refinement performs up to 4 iterative cycles (default 2). Each iteration maps reads using BWAmem2 Vasimuddin et al., BWA Li [a], or Bowtie2 Langmead et al. to the consensus, followed by variant calling with BCFtools Danecek et al. or iVar Grubaugh et al.. Benchmarking by Bassano et al. Bassano et al. showed that BCFtools outperformed iVar in precision and recall. iVar detects more low-frequency variants, resulting in an increased false positive rate but decreases the number of identified false negatives Bassano et al.. Users are recommended to consider prioritising sensitivity or specificity when selecting the variant caller.

Optional deduplication can be performed using Picard or when UMI's are available with UMI-tools Smith et al.. Mapping statistics are generated using samtools (flagstat, idxstats, stats) Danecek et al., Picard CollectMultipleMetrics Broad Institute, and coverage statistics with mosdepth Pedersen and Quinlan.

Consensus Quality Control

Quality control employs CheckV Nayfach et al. for completeness estimates, BLASTn Altschul et al. for reference similarity, and MMseqs2 Steinegger and Söding against the annotated database Virosaurus Gleizes et al.. These analyses enable species identification, genomic segment classification, host association determination, and extraction of additional metadata embedded within the reference databases.

The refinement progression is evaluated through sequence alignment with MAFFT Katoh et al., which compares final consensus genomes against *de novo* contigs, intermediate consensus sequences from iterative cycles, and the scaffolding reference. All tool metrics are compiled into an interactive MultiQC report Ewels et al. [a]. Additionally, key metrics are extracted from the MultiQC report and compiled into standalone overview tables to facilitate downstream analysis across all processed samples.

Applications

To assess the performance of nf-core/viralmetagenome under challenging scenarios, we simulated coinfection scenarios by mixing paired-end reads from public HIV-1 genomes with varying diversity (80-99% similarity), resulting in 13 samples (See supplementary table 2). Nf-core/viralmetagenome successfully identified coinfections in all mixed samples when genetic similarity was low to moderate ($\leq$ 96.7% ANI).

We investigated how reference genome selection affects consensus accuracy. While similar references ($\geq$ 96%) minimally impact consensus quality, divergent references introduce up to 187 nucleotide mismatches (Figure 2). These findings underscore the critical importance of appropriate reference selection for scaffolding and sequence alignment procedures. Nf-core/viralmetagenome addresses this challenge by automatically selecting the most appropriate reference genome based on sequence similarity; however, when no suitable genome is found in the supplied database, the contigs are scaffolded to a *de novo* contig (specifically, the cluster representative), ensuring that the final consensus genomes are as accurate as possible.

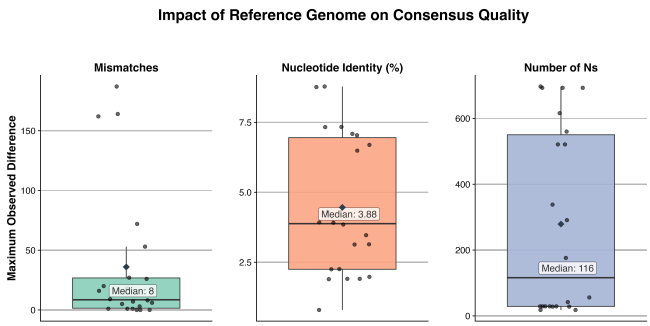

Fig. 2. Boxplot of maximum observed differences between consensus sequences generated with different reference genomes during scaffolding. Mean highlighted by a diamond.

We validated nf-core/viralmetagenome performance on real metagenomic samples spanning human and plant pathogens. Here, the pipeline successfully generated high-quality or near-complete genomes across viral families including segmented (Lassa virus, Orthonairovirus, Tomato spotted wilt tospovirus) and non-segmented viruses (SARS-CoV-2, West Nile virus, Potato virus Y,Youcai mosaic virus, and Monkeypox virus).

Processing 28 samples (supplementary methods) required 412 CPU hours and a maximum of 79GB RAM on an HPC, excluding taxonomic classification steps. The automated reference selection during scaffolding offers substantial improvements over manual curation by reducing processing time while preserving reconstruction accuracy. Performance correlates strongly with reference database comprehensiveness as consensus genomes tended to be more complete and similar to the true consensus sequence when scaffolding reference was closer to the true viral genome. This emphasizes the need to keep databases like RVDB Goodacre et al. and Virosaurus Gleizes et al. up-to-date. Since nf-core/viralmetagenome is primarily designed for eukaryotic viruses, bacteriophage analysis requires different approaches and users are encouraged to explore pipelines targeting phages such as VIRify Rangel-Pineros et al., VIBRANT Kieft et al. [b], VirSorter2 Guo et al..

Conclusion

nf-core/viralmetagenome addresses a critical need in viral genomics by providing an automated, scalable solution for untargeted viral genome reconstruction. The pipeline successfully automates the traditionally time-consuming and manual execution process of viral genome assembly from short-read metagenomic data through its integrated workflow of *de novo* contig assembly, automated reference selection, clustering algorithms, and iterative refinement strategies.

Our validation demonstrates the pipeline's broad applicability across diverse eukaryotic viral families, achieving high-quality genome reconstruction while ensuring reproducibility and ease of deployment across different computational environments.

As viral surveillance and outbreak response increasingly rely on metagenomic sequencing, automated pipelines like nf-core/viralmetagenome will be essential for the timely identification of pathogen strains. The pipeline represents a significant step forward in making viral genome reconstruction accessible to researchers without requiring extensive bioinformatics expertise, facilitating broader adoption of metagenomic approaches in viral research and public health applications.

Acknowledgments

J.K, P.L. and L.E.K acknowledge support from the Research Foundation - Flanders (Fonds voor Wetenschappelijk Onderzoek – Vlaanderen, G005323N and G051322N, 1SH2V24N, 12X9222N).

Author Contributions

J.K. designed and implemented the pipeline, performed validation analyses, and wrote the manuscript. P.L. and L.E.K. supervised the project and provided critical feedback. The nf-core community contributed to maintaining the pipeline. All authors reviewed and approved the final manuscript.

Data Availability

The nf-core/viralmetagenome pipeline is freely available at https://github.com/nf-core/viralmetagenome. The raw data and analysis code is available on https://github.com/Joon-Klaps/nf-core-viralmetagenome-manuscript/.

Conflict of Interest

The authors declare no competing interests.

References

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. 215:403–410. doi: 10.1016/S0022-2836(05)80360-2.

J. A. Baaijens, A. Z. E. Aabidine, E. Rivals, and A. Schönhuth. De novo assembly of viral quasispecies using overlap graphs. 27:835–848. doi: 10.1101/gr.215038.116.

I. Bassano, V. K. Ramachandran, M. S. Khalifa, C. J. Lilley, M. R. Brown, R. van Aerle, H. Denise, W. Rowe, A. George, E. Cairns, C. Wierzbicki, N. D. Pickwell, M. Wilson, M. Carlile, N. Holmes, A. Payne, M. Loose, T. A. Burke, S. Paterson, M. J. Wade, and J. M. S. Grimsley. Evaluation of variant calling algorithms for wastewater-based epidemiology using mixed populations of SARS-CoV-2 variants in synthetic and wastewater samples. doi: 10.1101/2022.06.06.22275866.

M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano. Scaffolding pre-assembled contigs using SSPACE. 27:578–579. doi: 10.1093/bioinformatics/btq683.

A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. 30:2114–2120. doi: 10.1093/bioinformatics/btu170.

Broad Institute. Picard toolkit.

B. Bushnell. BBMap.

V. A. Cantu, J. Sadural, and R. Edwards. PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets. doi: 10.7287/peerj.preprints.27553v1.

S. Chen, Y. Zhou, Y. Chen, and J. Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. 34:i884–i890. doi: 10.1093/bioinformatics/bty560.

P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. Twelve years of SAMtools and BCFtools. 10. doi: 10.1093/gigascience/giab008.

Z.-L. Deng, A. Dhingra, A. Fritz, J. Götting, P. C. Münch, L. Steinbrück, T. F. Schulz, T. Ganzenmüller, and A. C. McHardy. Evaluating assembly and variant calling software for strain-resolved analysis of large DNA viruses. 22. doi: 10.1093/bib/bbaa123.

P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame. Nextflow enables reproducible computational workflows. 35:316–319. doi: 10.1038/nbt.3820.

P. Ewels, M. Magnusson, S. Lundin, and M. Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. 32:3047–3048, a. doi: 10.1093/bioinformatics/btw354.

P. A. Ewels, A. Peltzer, S. Fillinger, H. Patel, J. Alneberg, A. Wilm, M. U. Garcia, P. Di Tommaso, and S. Nahnsen. The nf-core framework for community-curated bioinformatics pipelines. 38:276–278, b. doi: 10.1038/s41587-020-0439-x.

M. Forbes, D. Y. K. Ng, R. M. Boggan, A. Frick-Kretschmer, J. Durham, O. Lorenz, B. Dave, F. Lassalle, C. Scott, J. Wagner, A. Lignes, F. Noaves, D. K. Jackson, K. Howe, and E. Harrison. Benchmarking of human read removal strategies for viral and microbial metagenomics. page 2025.03.21.644587. doi: 10.1101/2025.03.21.644587.

A. Gleizes, F. Laubscher, N. Guex, C. Iseli, T. Junier, S. Cordey, J. Fellay, I. Xenarios, L. Kaiser, and P. L. Mercier. Virosaurus a reference to explore and capture virus genetic diversity. 12. doi: 10.3390/v12111248.

N. Goodacre, A. Aljanahi, S. Nandakumar, M. Mikailov, and A. S. Khan. A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. 3. doi: 10.1128/mSphereDirect.00069-18.

M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-seq data without a reference genome. 29:644–652. doi: 10.1038/nbt.1883.

N. D. Grubaugh, K. Gangavarapu, J. Quick, N. L. Matteson, J. G. De Jesus, B. J. Main, A. L. Tan, L. M. Paul, D. E. Brackney, S. Grewal, N. Gurfield, K. K. A. Van Rompay, S. Isern, S. F. Michael, L. L. Coffey, N. J. Loman, and

K. G. Andersen. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. 20:8. doi: 10.1186/s13059-018-1618-7.

J. Guo, B. Bolduc, A. A. Zayed, A. Varsani, G. Dominguez-Huerta, T. O. Delmont, A. A. Pratama, M. C. Gazitúa, D. Vik, M. B. Sullivan, and S. Roux. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. 9:37. doi: 10.1186/s40168-020-00990-y.

K. Katoh, K. Misawa, K.-I. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. 30:3059–3066. doi: 10.1093/nar/gkf436.

K. Kieft, A. Adams, R. Salamzade, L. Kalan, and K. Anantharaman. vRhyme enables binning of viral genomes from metagenomes. 50:e83, a. doi: 10.1093/nar/gkac341.

K. Kieft, Z. Zhou, and K. Anantharaman. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. 8:90, b. doi: 10.1186/s40168-020-00867-0.

G. M. Kurtzer, V. Sochat, and M. W. Bauer. Singularity: Scientific containers for mobility of compute. 12:e0177459. doi: 10.1371/journal.pone.0177459.

B. Langmead, C. Wilks, V. Antonescu, and R. Charles. Scaling read aligners to hundreds of threads on general-purpose processors. 35:421–432. doi: 10.1093/bioinformatics/bty648.

J. F. J. Laros. HUMID: HUMID: reference free FastQ deduplication.

D. Li, R. Luo, C.-M. Liu, C.-M. Leung, H.-F. Ting, K. Sadakane, H. Yamashita, and T.-W. Lam. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. 102:3–11, a. doi: 10.1016/j.ymeth.2016.02.020.

H. Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. a.

H. Li. Minimap2: pairwise alignment for nucleotide sequences. 34:3094–3100, b. doi: 10.1093/bioinformatics/bty191.

W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. 22: 1658–1659. doi: 10.1093/bioinformatics/btl158.

Y. Li, M. Polychronopoulou, I. Boonen, A. Fikatas, S. Gryseels, A. Laudisoit, J. G. de Bellocq, B. Vrancken, G. Magiorkinis, P. Lemey, and M. Bletsa. Evaluation of metatranscriptomic sequencing protocols to obtain full-length RNA virus genomes from mammalian tissues. 20:e0324537, b. doi: 10.1371/journal.pone.0324537.

D. Meleshko, I. Hajirasouliha, and A. Korobeynikov. coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies. doi: 10.1093/bioinformatics/btab597.

P. Menzel, K. L. Ng, and A. Krogh. Fast and sensitive taxonomic classification for metagenomics with kaiju. 7: 11257. doi: 10.1038/ncomms11257.

D. Merkel. Docker: lightweight linux containers for consistent development and deployment. 2014:2. doi: 10.5555/2600239.2600241.

S. Nayfach, A. P. Camargo, F. Schulz, E. Eloe-Fadrosh, S. Roux, and N. C. Kyrpides. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. 39:578–585. doi: 10.1038/s41587-020-00774-7.

B. D. Ondov, N. H. Bergman, and A. M. Phillippy. Interactive metagenomic visualization in a web browser. 12:385, a. doi: 10.1186/1471-2105-12-385.

B. D. Ondov, G. J. Starrett, A. Sappington, A. Kostic, S. Koren, C. B. Buck, and A. M. Phillippy. Mash screen: high-throughput sequence containment estimation for genome discovery. 20:232, b. doi: 10.1186/s13059-019-1841-x.

B. S. Pedersen and A. R. Quinlan. Mosdepth: quick coverage calculation for genomes and exomes. 34:867–868. doi: 10.1093/bioinformatics/btx699.

G. Rangel-Pineros, A. Almeida, M. Beracochea, E. Sakharova, M. Marz, A. R. Muñoz, M. Hölzer, and R. D. Finn. VIRify: an integrated detection, annotation and taxonomic classification pipeline using virus-specific protein profile hidden markov models. page 2022.08.22.504484. doi: 10.1101/2022.08.22.504484.

T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé. VSEARCH: a versatile open source tool for metagenomics. 4:e2584. doi: 10.7717/peerj.2584.

S. S. Shepard, S. Meno, J. Bahl, M. M. Wilson, J. Barnes, and E. Neuhaus. Viral deep sequencing needs an adaptive approach: IRMA, the iterative refinement meta-assembler. 17:708. doi: 10.1186/s12864-016-3030-6.

T. Smith, A. Heger, and I. Sudbery. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. 27:491–499. doi: 10.1101/gr.209601.116.

M. Steinegger and J. Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. 35: 1026–1028. doi: 10.1038/nbt.3988.

C. Tomkins-Tinch, S. Ye, H. Metsky, I. Jungreis, R. Sealfon, X. Yang, K. Andersen, M. Lin, and D. Park. broadinstitute/viral-ngs: v1.14.0.

M. Vasimuddin, S. Misra, H. Li, and S. Aluru. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 314–324. IEEE. doi: 10.1109/IPDPS.2019.00041.

D. E. Wood, J. Lu, and B. Langmead. Improved metagenomic analysis with kraken 2. 20:257. doi: 10.1186/s13059-019-1891-0.

A. Zielezinski, A. Gudyś, J. Barylski, K. Siminski, P. Rozwalak, B. E. Dutilh, and S. Deorowicz. Ultrafast and accurate sequence alignment and clustering of viral genomes. 22: 1191–1194. doi: 10.1038/s41592-025-02701-7.