Bioinformatics, 2026, **42**(5), btag187
https://doi.org/10.1093/bioinformatics/btag187
Published: 29 April 2026
Applications Note | Genome analysis

OXFORD

nf-core/viralmetagenome: A novel pipeline for untargeted viral genome reconstruction

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

[1]Rega Institute for Medical Research, Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium
[2]Leiden University Center of Infectious Diseases (LUCID), Leiden University Medical Center, Leiden, 2333 ZA, the Netherlands
*Corresponding author. Joon Klaps, Rega Institute for Medical Research, Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium. E-mail: joon.klaps@kuleuven.be
[†]A full list of contributors can be found at https://nf-co.re/community
Associate Editor: Christina Kendziorski

Abstract

Motivation: **Reconstructing eukaryotic viral genomes from metagenomic data is challenging due to their extensive diversity and potential genome segmentation. Current approaches often rely on labor-intensive manual curation for reference selection and scaffolding, limiting scalability for large studies or rapid outbreak response. We address the critical need for an automated, scalable pipeline for efficient viral metagenomic analysis without manual intervention.**

Results: **We present nf-core/viralmetagenome, a comprehensive Nextflow pipeline for the untargeted reconstruction and variant analysis of eukaryotic DNA and RNA viruses from short-read metagenomic or hybridisation capture enriched samples. The pipeline automates the entire process from read preprocessing to consensus generation, integrating multiple *de novo* assemblers, automated reference selection, and iterative consensus refinement. It features robust quality control, extensive documentation, and seamless portability via Docker and Singularity. We validated the pipeline on diverse simulated and real datasets, demonstrating its ability to recover high-quality genomes from complex metagenomic samples and resolve co-infections, making it a powerful tool for viral surveillance.**

Availability: **nf-core/viralmetagenome is freely available at https://github.com/nf-core/viralmetagenome with comprehensive documentation at https://nf-co.re/viralmetagenome. Archival code repository snapshots are published at zenodo with doi: https://doi.org/10.5281/zenodo.17524074.**

1. Introduction

Reconstructing viral genomes from metagenomic sequencing data presents considerable computational challenges, particularly for viruses exhibiting extensive genetic diversity (Baaijens et al. 2017, Meleshko et al. 2021). This challenge is amplified in segmented viruses such as influenza, rotavirus, and bunyaviruses, where individual segments can evolve under distinct selective pressures and can undergo reassortment. While some pipelines target specific viruses and their subtypes (Shepard et al. 2016), accurate genome reconstruction from samples with unknown viral strains typically requires time-consuming manual curation of contigs and reference matching (de Vries et al. 2021). This makes large-scale metagenome studies or rapid outbreak responses impractical.

To address these limitations, we developed nf-core/viralmetagenome, a comprehensive pipeline specifically designed for untargeted viral genome reconstruction using short read sequences. The pipeline is developed using Nextflow (Di Tommaso et al. 2017) within the nf-core framework (Langer et al. 2025), ensuring reproducibility through containerization with Docker (Merkel 2014) and Singularity (Kurtzer et al. 2017), as well as portability across computational platforms such as local desktops, high-performance clusters and cloud environments.

2. Pipeline description

nf-core/viralmetagenome implements a workflow for *de novo* assembly, reference matching, and iterative consensus refinement to reconstruct viral genomes without prior target knowledge. The pipeline comprises five stages: read preprocessing, metagenomic diversity assessment, contig assembly and scaffolding, iterative consensus refinement with variant analysis, and quality control (Fig. 1). Unless otherwise specified, the pipeline offers multiple options to accommodate established user workflows and preferences. Detailed tool descriptions are provided in Table 1, available as supplementary data at *Bioinformatics* online.

Received: 23 July 2025. **Revised:** 17 March 2026. **Accepted:** 10 April 2026

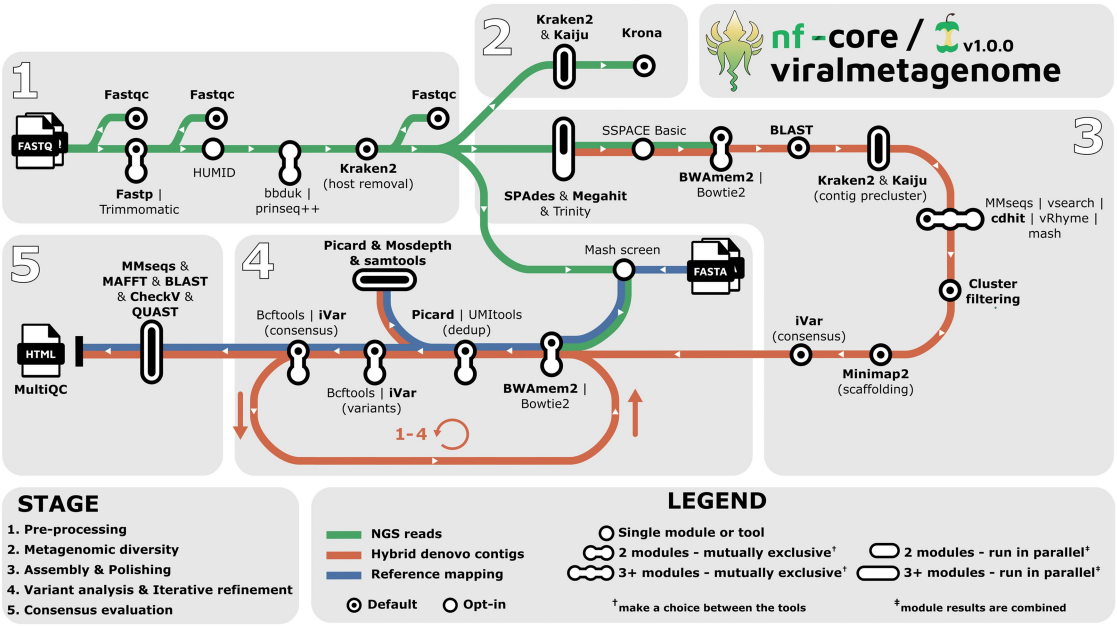

Figure 1 Overview of the nf-core/viralmetagenome pipeline for untargeted viral genome reconstruction. Default options are highlighted in bold. nf-core/viralmetagenome processes short-read data through pre-processing, metagenomic diversity assessment, *de novo* assembly with multiple assemblers, scaffolding with automated reference identification and taxonomy-guided contig clustering, and iterative consensus refinement. Quality-control metrics, assembly statistics, and coverage data are summarised in MultiQC reports and overview tables for downstream analysis.

2.1. Read preprocessing

Input reads are provided via sample sheets containing sample names and paths to short-read FASTQ files. These are preprocessed using FastQC and trimmed for adapters using fastp (Chen et al. 2018) (default) or Trimmomatic (Bolger et al. 2014). Fastp offers automated adapter detection and faster processing (Chen et al. 2018). UMI deduplication is implemented using HUMID (GitHub repository: https://github.com/jfjlaros/HUMID) for raw reads, or UMI-tools after mapping (Smith et al. 2017). Multiple sequencing runs can be merged after trimming by specifying merge group identifiers in the sample sheet. Optional filtering with bbduk (Bushnell 2014) or PRINSEQ++ (Cantu *et al.* 2019) removes low-complexity sequences, while host reads are removed with Kraken2 (Wood et al. 2019).

2.2. Metagenomic diversity assessment

Preprocessed reads are taxonomically classified with Kaiju (Menzel et al. 2016) and Kraken2 (Wood et al. 2019) to maximise sensitivity across diverse viral families. Results from both classifiers are visualised using Krona (Ondov et al. 2011).

2.3. *De novo* assembly and clustering

The assembly workflow employs a multi-assembler approach where contigs generated by all tools are pooled and processed jointly. *De novo* assembly uses SPAdes (Meleshko et al. 2021) (RNAviral mode), MEGAHIT (Li et al. 2016), and Trinity (Grabherr et al. 2011), capitalizing on their distinct algorithmic strengths to maximise genome recovery across diverse viral families and read depths. Subsequently, a BLASTn (Altschul et al. 1990) search against the Reference Viral Database (RVDB) (Chin et al. 2025) identifies candidate reference genomes for each contig. All contigs and their identified reference genomes are subjected to clustering. This strategy enables contig scaffolding against the cluster centroid whether a database reference genome or a *de novo* contig, thereby combining outputs from multiple assemblers into unified scaffolds.

Clustering is performed in two sequential stages. First, taxonomic pre-clustering groups contigs based on taxonomic classification using Kraken2 (Wood et al. 2019) and Kaiju (Menzel et al. 2016), with optional filtering to focus on specific taxonomic clades for more targeted analyses. Second, nucleotide similarity clustering within taxonomic groups is performed using CD-HIT-EST (Li and Godzik 2006) (default), VSEARCH (Rognes et al. 2016), MMseqs2 (Steinegger and Söding 2017), vRhyme (Kieft et al. 2022), or Mash (Ondov et al. 2019). All tools are valid options, though performance may vary by dataset (Steinegger and Söding 2017, Zielezinski et al. 2025).

Optionally, following *de novo* assembly and extension, reads can be mapped to all contigs using BWA-MEM2 (Vasimuddin et al. 2019) (default), or Bowtie2 (Langmead et al. 2019). Clusters are filtered based on the total percentage of reads mapping to all contigs within a cluster, removing assembly artefacts. For the final scaffolding step, all cluster members are mapped to the cluster representative or centroid using Minimap2 (Li 2018), followed by consensus calling with iVar (Grubaugh et al. 2019) to generate reference-assisted scaffolds.

2.4. Consensus generation and variant calling

The consensus module supports reference-guided analysis and scaffold refinement using user-defined reference genomes.

Users can provide one or multiple references; when multiple are provided, Mash (Ondov et al. 2019) selects the most similar one. The selected reference then undergoes a single refinement round to generate the consensus.

In contrast, scaffold refinement employs an iterative approach, performing up to 4 cycles (default 2) to polish the consensus sequence. Both processes map reads using BWA-MEM2 (Vasimuddin et al. 2019) (default), or Bowtie2 (Langmead et al. 2019), followed by variant calling with BCFtools (Danecek et al. 2021) or iVar (Grubaugh et al. 2019) (default). Benchmarking (Bassano et al. 2022) indicated that BCFtools outperforms iVar in precision and recall, while iVar identified more low-frequency variants. Users can prioritise sensitivity or specificity when selecting the variant caller. Final variant calling is performed on the polished consensus to characterise single nucleotide variants.

3. Consensus quality control

Quality control employs CheckV (Nayfach et al. 2021) to assess completeness, BLASTn (Altschul et al. 1990) for reference similarity, and MMseqs2 (Steinegger and Söding 2017) against the annotated database Virosaurus (Gleizes et al. 2020). These analyses enable species identification, genomic segment classification, host prediction, and retrieval of additional metadata embedded within the reference databases.

The consensus refinement is evaluated through sequence alignment with MAFFT (Katoh et al. 2002), 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. 2016). Additionally, key metrics are extracted from the MultiQC report and compiled into overview tables for downstream analysis.

4. Applications

We validated the pipeline's performance using a diverse set of real metagenomic samples containing human and plant pathogens. nf-core/viralmetagenome successfully reconstructed high-quality or near-complete consensus genomes across various viral families, including segmented viruses (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; Supplementary methods S1, available as supplementary data at *Bioinformatics* online). The pipeline proved efficient, processing 28 samples in 412 CPU hours with a maximum memory footprint of 79GB on a high-performance cluster. To facilitate local deployment, we implemented a `-profile local` option, which utilizes a smaller viral RefSeq database instead of the full RVDB database for Kaiju, reducing the pipeline's peak RAM usage to approximately 18GB.

To evaluate the pipeline's robustness in complex scenarios, we conducted two targeted experiments. First, we assessed its ability to resolve co-infections using a simulated mixture of viral genomes (Supplementary Methods S2, available as supplementary data at *Bioinformatics* online). The pipeline successfully distinguished and reconstructed distinct co-infecting strains, provided the genetic divergence was sufficient ($\lesssim$ 88.7% average nucleotide identity (ANI); Figure 1, available as supplementary data at *Bioinformatics* online).

The second part of our benchmarking explored the critical role of the scaffolding reference in consensus genome reconstruction. We selected sequencing data from seven Lassa virus-positive human clinical samples with diverse viral loads (Supplementary Methods S3; Figure 2, available as supplementary data at *Bioinformatics* online). In these samples, a significantly higher sequencing depth was observed for the S segment than for the L segment (Figure 4, available as supplementary data at *Bioinformatics* online). To establish a benchmark, a baseline consensus genome was generated for each sample using automated reference selection from the full unclustered-RVDB (U-RVDB) reference pool. We then systematically challenged the scaffolding process by providing reference genomes (19–30 per sample; Table 2, available as supplementary data at *Bioinformatics* online) spanning a range of similarity to the baseline consensus (64–100% ANI; Figure 3, available as supplementary data at *Bioinformatics* online). For each supplied reference, a new consensus genome was generated using it as the scaffolding backbone when clustered with the *de novo* contigs.

Comparing these generated sequences against their respective baseline consensus highlights the utility of reference scaffolding (Fig. 2), particularly in low read depth scenarios when *de novo* assembly was challenging. Here, we defined consensus accuracy as the Average Nucleotide Identity between each test consensus and its baseline (Supplementary Methods S3, available as supplementary data at *Bioinformatics* online). For the high-coverage S segment (Figure 4, available as supplementary data at *Bioinformatics* online), *de novo* assembly was sufficient to reconstruct the full genome, rendering the choice of scaffolding reference largely redundant (Fig. 2A). In contrast, the low-coverage L segment proved more challenging: *de novo* assembly produced shorter, incomplete fragments, demonstrating that the pipeline's ability to identify and use a closely related database reference genome was essential to bridge gaps and recover a complete segment consensus sequence.

These findings emphasise the need to keep databases like RVDB (Chin et al. 2025) up-to-date. Since nf-core/viralmetagenome is primarily designed for eukaryotic viruses, users interested in bacteriophage analysis are encouraged to explore pipelines designed to target phages such as VIRify (Rangel-Pineros et al. 2022), VIBRANT (Kieft et al. 2020), VirSorter2 (Guo et al. 2021).

5. 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 process of viral genome reconstruction 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 results indicate that the pipeline is suitable for a wide range of eukaryotic viral families, it removes the burden of extensive manual curation, and enables consistent, high-quality genome reconstruction with reproducible results across varied

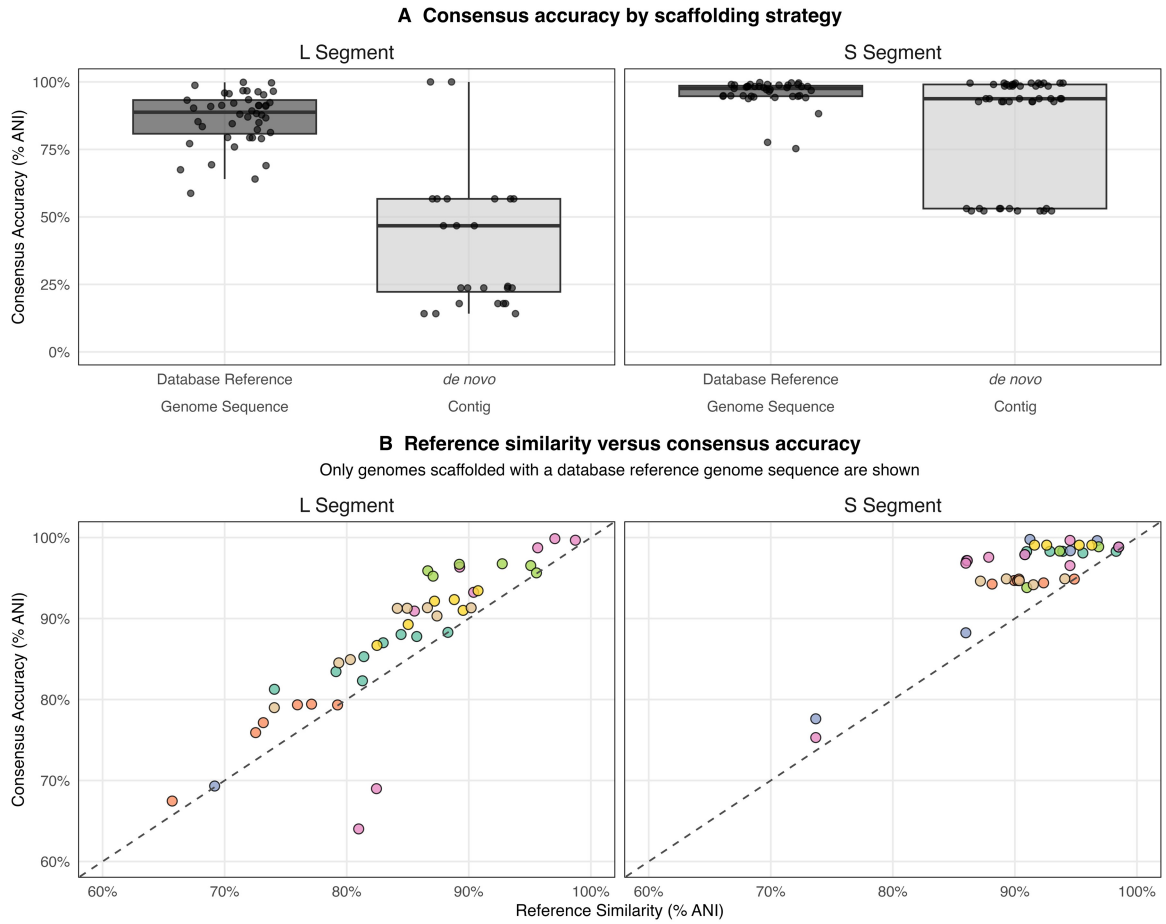

Figure 2 Impact of scaffolding reference similarity on consensus reconstruction accuracy. Consensus accuracy is defined as the ANI (Eq. 1 in Supplementary Methods, available as supplementary data at *Bioinformatics* online) between the test consensus and the baseline consensus (generated using the pipeline's automated reference selection from the full unclustered-RVDB). Test consensus genomes were scaffolded using either a supplied database reference genome or a *de novo* contig. (A) Boxplots comparing Consensus Accuracy (% ANI) for the L and S segments. The plots contrast the performance of using a database reference genome as the scaffolding backbone (dark grey) against relying solely on the *de novo* contig (light grey). For the low-coverage L segment, using the database reference genome sequences significantly improves accuracy, whereas for the high-coverage S segment, the *de novo* route performs comparably to the database reference genome scaffolding (sequencing depths detailed in Supplementary Figure S4, available as supplementary data at *Bioinformatics* online). (B) Scatter plots illustrate the relationship between reference similarity and consensus accuracy (% ANI) for the L and S segments. Here, reference similarity is defined as the ANI between the supplied database reference genome sequence and the baseline consensus genome. The dashed line represents the line of identity ($y = x$), and points are coloured by sample source (Supplementary Figure S2, available as supplementary data at *Bioinformatics* online). Only consensus genomes that used a database reference genome sequence as the scaffold reference are shown in panel B.

computational settings. 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.

Acknowledgements

We thank our collaborators at the Bernhard Nocht Institute for Tropical Medicine (BNITM), Germany, and their partners at the Irrua Specialist Teaching Hospital (ISTH), Nigeria, for valuable support from which the pipeline development benefited. Assistance from AI language models (e.g., ChatGPT) was used to enhance the clarity of the manuscript and to support code development.

Author contributions

Joon Klaps (Conceptualization [Lead], Funding acquisition [Equal], Software [Lead], Writing—original draft [Lead]), Philippe Lemey (Conceptualization [Supporting], Funding acquisition [Equal], Supervision [Supporting], Writing—review & editing [Equal]), Magda Bletsa (Conceptualization [Supporting], Methodology [Supporting], Writing—review & editing [Supporting]), and Nf-core Community (Resources [Supporting], Software [Supporting], Validation [Supporting]), Liana Eleni

Kafetzopoulou (Conceptualization [Equal], Funding acquisition [Equal], Supervision [Equal], Writing—review & editing [Equal])

Supplementary material

Supplementary material is available at *Bioinformatics* online.

Conflicts of interest

None declared.

Funding

J.K, M.B, P.L. and L.E.K acknowledge support from the German Research Foundation [Deutsche Forschungsgemeinschaft 511260616, STELLAR] as well as Research Foundation—Flanders [Fonds voor Wetenschappelijk Onderzoek—Vlaanderen, G005323N and G051322N, 1SH2V24N, 12X9222N].

Data availability

nf-core/viralmetagenome is freely available at https://github.com/nf-core/viralmetagenome with comprehensive documentation at https://nf-co.re/viralmetagenome. Archival code repository snapshots are published at zenodo with doi: https://doi.org/10.5281/zenodo.17524074.

References

Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2

Baaijens JA, Aabidine AZE, Rivals E *et al.* De novo assembly of viral quasispecies using overlap graphs. *Genome Res* 2017;**27**:835–48. https://doi.org/10.1101/gr.215038.116

Bassano I *et al.* Evaluation of variant calling algorithms for wastewater-based epidemiology using mixed populations of SARS-CoV-2 variants in synthetic and wastewater samples. *bioRxiv*2022. https://doi.org/10.1101/2022.06.06.22275866

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 2014;**30**:2114–20. https://doi.org/10.1093/bioinformatics/btu170

Brian Bushnell. *BBMap: A Fast, Accurate, Splice-Aware Aligner*. Berkeley, CA: Ernest Orlando Lawrence Berkeley National Laboratory, 2014.

Cantu VA, Sadural J, Edwards R. PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets. *PeerJ* 2019;**7**:e27553v1. https://doi.org/10.7287/peerj.preprints.27553v1

Chen S, Zhou Y, Chen Y *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90. https://doi.org/10.1093/bioinformatics/bty560

Chin P-J, Bhavsar JD, Bosma TJ *et al.* Refinement of the reference viral database (RVDB) for improving bioinformatics analysis of virus detection by high-throughput sequencing (HTS). *mSphere* 2025;**10**:e0028625. https://doi.org/10.1128/msphere.00286-25

Danecek P, Bonfield JK, Liddle J *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 2021;**10**:giab008. https://doi.org/10.1093/gigascience/giab008

de Vries JJC, Brown JR, Fischer N *et al.* Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. *J Clin Virol* 2021;**141**:104908. https://doi.org/10.1016/j.jcv.2021.104908

Di Tommaso P, Chatzou M, Floden EW *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;**35**:316–9. https://doi.org/10.1038/nbt.3820

Ewels P, Magnusson M, Lundin S *et al.* MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**:3047–8. https://doi.org/10.1093/bioinformatics/btw354

Gleizes A, Laubscher F, Guex N *et al.* Virosaurus a reference to explore and capture virus genetic diversity. *Viruses* 2020;**12**:1248. https://doi.org/10.3390/v12111248

Grabherr MG, Haas BJ, Yassour M *et al.* Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol* 2011;**29**:644–52. https://doi.org/10.1038/nbt.1883

Grubaugh ND, Gangavarapu K, Quick J *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* 2019;**20**:8. https://doi.org/10.1186/s13059-018-1618-7

Guo J, Bolduc B, Zayed AA *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 2021;**9**:37. https://doi.org/10.1186/s40168-020-00990-y

Katoh K, Misawa K, Kuma K-I *et al.* MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res* 2002;**30**:3059–66. https://doi.org/10.1093/nar/gkf436

Kieft K, Adams A, Salamzade R *et al.* vRhyme enables binning of viral genomes from metagenomes. *Nucleic Acids Res* 2022;**50**:e83. https://doi.org/10.1093/nar/gkac341

Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 2020;**8**:90. https://doi.org/10.1186/s40168-020-00867-0

Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. *PLoS One* 2017;**12**:e0177459. https://doi.org/10.1371/journal.pone.0177459

Langer BE, Amaral A, Baudement M-O *et al.* Empowering bioinformatics communities with nextflow and nf-core. *Genome Biol* 2025;**26**:228. https://doi.org/10.1186/s13059-025-03673-9

Langmead B, Wilks C, Antonescu V *et al.* Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* 2019;**35**:421–32. https://doi.org/10.1093/bioinformatics/bty648

Li D, Luo R, Liu C-M *et al.* MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 2016;**102**:3–11. https://doi.org/10.1016/j.ymeth.2016.02.020

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100. https://doi.org/10.1093/bioinformatics/bty191

Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9. https://doi.org/10.1093/bioinformatics/btl158

Meleshko D, Hajirasouliha I, Korobeynikov A. coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies. *Bioinformatics* 2021;**38**:1–8. https://doi.org/10.1093/bioinformatics/btab597

Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nat Commun* 2016;**7**:11257. https://doi.org/10.1038/ncomms11257

Merkel D. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal* 2014;**2014**:2. https://doi.org/10.5555/2600239.2600241

Nayfach S, Camargo AP, Schulz F *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* 2021;**39**:578–85. https://doi.org/10.1038/s41587-020-00774-7

Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics* 2011;**12**:385. https://doi.org/10.1186/1471-2105-12-385

Ondov BD, Starrett GJ, Sappington A *et al.* Mash screen: high-throughput sequence containment estimation for genome discovery. *Genome Biol* 2019;**20**:232. https://doi.org/10.1186/s13059-019-1841-x

Rangel-Pineros G *et al.* VIRify: an integrated detection, annotation and taxonomic classification pipeline using virus-specific protein profile hidden markov models. *bioRxiv*, page 2022.08.22.5044842022. https://doi.org/10.1101/2022.08.22.504484

Rognes, Torbjørn, Flouri, Tomáš, Nichols, Ben, *et al.* VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**:e2584, 2016. https://doi.org/10.7717/peerj.2584

Shepard SS, Meno S, Bahl J *et al.* Viral deep sequencing needs an adaptive approach: IRMA, the iterative refinement meta-assembler. *BMC Genomics* 2016;**17**:708. https://doi.org/10.1186/s12864-016-3030-6

Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res* 2017;**27**:491–9. https://doi.org/10.1101/gr.209601.116

Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**:1026–8. https://doi.org/10.1038/nbt.3988

Vasimuddin M, Misra S, Li H *et al.* Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2019, 314–24. https://doi.org/10.1109/IPDPS.2019.00041

Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. *Genome Biol* 2019;**20**:257. https://doi.org/10.1186/s13059-019-1891-0

Zielezinski A, Gudyś A, Barylski J *et al.* Ultrafast and accurate sequence alignment and clustering of viral genomes. *Nat Methods* 2025;**22**:1191–4. https://doi.org/10.1038/s41592-025-02701-7