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

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

[1]Rega Institute for Medical Research, Department of Microbiology, Immunology and Transplantation, KU Leuven, Belgium

[2]A full list of contributors can be found at https://nf-co.re/community.

*Correspondence: joon.klaps@kuleuven.be

July 23, 2025

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.

Keywords: viralmetagenome, bioinformatic pipeline, nextflow, viral metagenomics, viral assembly, viral variant analysis

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 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. 2016), accurate and complete genome reconstruction of samples with unknown references typically requires manual curation of contigs and reference matching (de Vries et al. 2021). 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. 2017) within the nf-core framework (Ewels et al. 2020), ensuring reproducibility through containerization with Docker (Merkel 2014) and Singularity (Kurtzer et al. 2017), and enabling portability across computational platforms such as local desktops, high-performance clusters and cloud environments.

2 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 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, unless otherwise specified, the pipeline offers multiple options to accommodate established user workflows and preferences. In depth tool details are available in Supplementary Table 1.

2.1 Read preprocessing

Input reads provided via sample sheets containing sample names and short-read FASTQ paths are preprocessed using FastQC and have their adapters trimmed with fastp (Chen et al. 2018) (default) or Trimmomatic (Bolger et al. 2014). Fastp is overall faster and has automated adapter detection and trimming (Chen et al. 2018). UMI deduplication is implemented using HUMID (Laros 2025) and once reads are mapped to a reference with UMI-tools (Smith et al. 2017). Multiple sequencing runs are merged after trimming by specifying merge group identifiers in the sample sheet. Complexity filtering with bbduk (Bushnell 2022) or PRINSEQ++ (Cantu et al. 2019) removes low-complexity sequences containing repetitive elements where host reads are removed with Kraken2 (Wood et al. 2019).

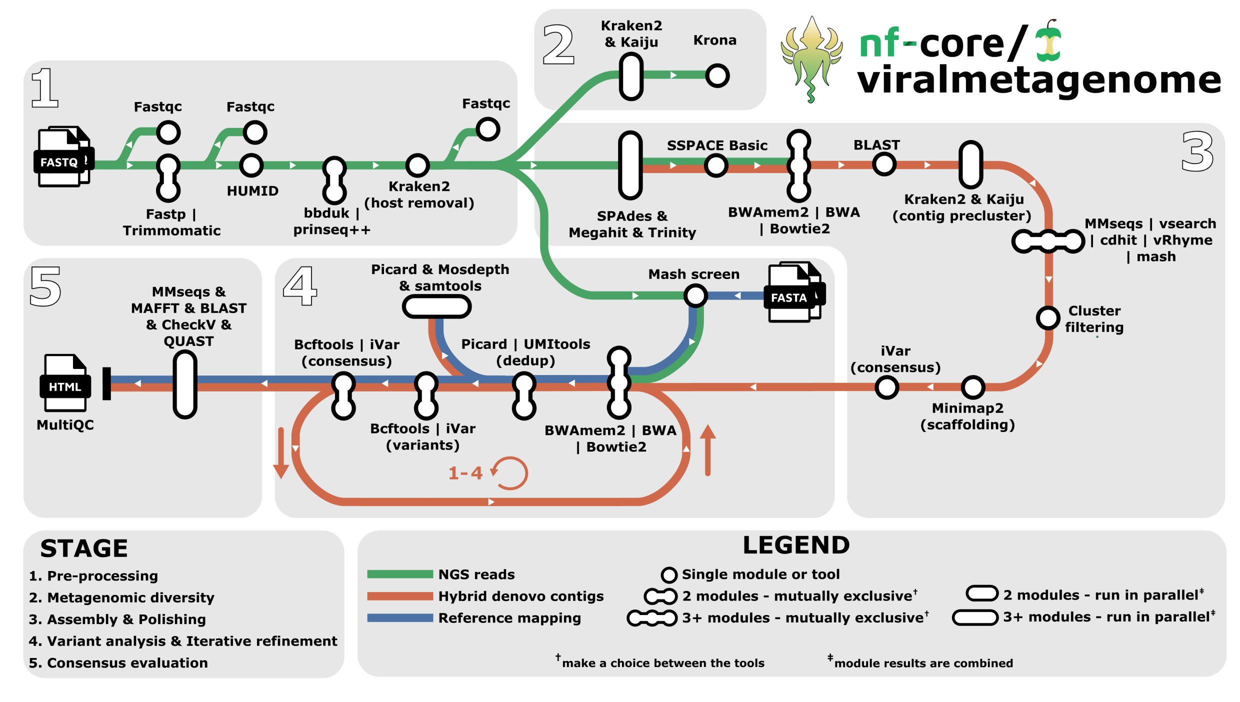

Figure 1: Visual overview of the nf-core/viralmetagenome pipeline for untargeted viral genome reconstruction. nf-core/viralmetagenome processes short-read data through pre-processing, metagenomic diversity assessment, *de novo* assembly with multiple assemblers, scaffolding with automated reference identification 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.

2.2 Metagenomic diversity assessment

Taxonomic classification of preprocessed reads is performed using two complementary approaches - Kaiju (Menzel et al. 2016) and Kraken2 (Wood et al. 2019) - to maximise detection 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 implements multi-assembler approaches followed by clustering and scaffolding. *De novo* assembly is performed using SPAdes (Meleshko et al. 2021) (RNAviral mode), MEGAHIT (Li et al. 2016), and Trinity (Grabherr et al. 2011), capitalizing on distinct algorithmic strengths to maximise genome recovery across diverse viral families and variable read depths.

Reference identification uses BLASTn (Altschul et al. 1990) against the Reference Viral Database (RVDB) (Goodacre et al. 2018), retaining top five hits to facilitate identification of related genomic segments and appropriate reference sequences for contig scaffolding and clustering.

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), 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 depending on the dataset (Zielezinski et al. 2025, Steinegger and Söding 2017).

As an optional filtering step of contig clusters, after assembly and extension, reads can be mapped to all contigs using BWA-MEM2 (Vasimuddin et al. 2019) (default), BWA (Li 2013), or bowtie2 (Langmead et al. 2019). Clusters are filtered based on the total percentage of reads mapping to all contigs within a cluster, allowing identification and removal of clusters that likely represent assembly artefacts resulting from low read coverage.

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 assemblies. Regions with zero coverage depth can optionally be represented by the reference genome to produce a more complete scaffold genome for consensus calling.

2.4 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. 2019).

Scaffold refinement performs up to 4 iterative cycles (default 2). Each iteration maps reads using BWA-MEM2 (Vasimuddin et al. 2019), BWA (Li 2013), or bowtie2 (Langmead et al. 2019) to the consensus, followed by variant calling with BCFtools (Danecek et al. 2021) or iVar (Grubaugh et al. 2019). Benchmarking by Bassano et al. (Bassano et al. 2022) showed that BCFtools outperformed iVar in precision and recall, where iVar identified more low frequency variants. Users are recommended to consider prioritising sensitivity or specificity when selecting the variant caller.

2.5 Consensus Quality Control

Quality control employs CheckV (Nayfach et al. 2021) for completeness estimates, 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 any other additional metadata embedded within the reference databases.

The refinement progression 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 standalone overview tables to facilitate downstream analysis across all processed samples.

3 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%), 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%).

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 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 the scaffolding reference was closer to the true viral genome. This emphasizes the need to keep databases like RVDB (Goodacre et al. 2018) and Virosaurus (Gleizes et al. 2020) 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. 2022), VIBRANT (Kieft et al. 2020), VirSorter2 (Guo et al. 2021).

4 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.

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. *J. Mol. Biol.*, 215(3):403–410, 1990. doi: 10.1016/S0022-2836(05)80360-2.

Jasmijn A Baaijens, Amal Zine El Aabidine, Eric Rivals, and Alexander Schönhuth. De novo assembly of viral quasispecies using overlap graphs. *Genome Res.*, 27(5):835–848, 2017. doi: 10.1101/gr.215038.116.

Irene Bassano, Vinoy K Ramachandran, Mohammad S Khalifa, Chris J Lilley, Mathew R Brown, Ronny van Aerle, Hubert Denise, William Rowe, Airey George, Edward Cairns, Claudia Wierzbicki, Natalie D Pickwell, Myles Wilson, Matthew Carlile, Nadine Holmes, Alexander Payne, Matthew Loose, Terry A Burke, Steve Paterson, Matthew J Wade, and Jasmine 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. *bioRxiv*, 2022. doi: 10.1101/2022.06.06.22275866.

Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014. doi: 10.1093/bioinformatics/btu170.

Brian Bushnell. BBMap, 2022.

Vito Adrian Cantu, Jeffrey Sadural, and Robert Edwards. PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets. *PeerJ*, 2019. doi: 10.7287/peerj.preprints.27553v1.

Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018. doi: 10.1093/bioinformatics/bty560.

Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *Gigascience*, 10(2), 2021. doi: 10.1093/gigascience/giab008.

Jutte J C de Vries, Julianne R Brown, Nicole Fischer, Igor A Sidorov, Sofia Morfopoulou, Jiabin Huang, Bas B Oude Munnink, Arzu Sayiner, Alihan Bulgurcu, Christophe Rodriguez, Guillaume Gricourt, Els Keyaerts, Leen Beller, Claudia Bachofen, Jakub Kubacki, Cordey Samuel, Laubscher Florian, Schmitz Dennis, Martin Beer, Dirk Hoeper, Michael Huber, Verena Kufner, Maryam Zaheri, Aitana Lebrand, Anna Papa, Sander van Boheemen, Aloys C M Kroes, Judith Breuer, F Xavier Lopez-Labrador, and Eric C J Claas. Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. *J. Clin. Virol.*, 141:104908, 2021. doi: 10.1016/j.jcv.2021.104908.

Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, 35 (4):316–319, 2017. doi: 10.1038/nbt.3820.

Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016. doi: 10.1093/bioinformatics/btw354.

Philip A Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.*, 38(3):276–278, 2020. doi: 10.1038/s41587-020-0439-x.

Anne Gleizes, Florian Laubscher, Nicolas Guex, Christian Iseli, Thomas Junier, Samuel Cordey, Jacques Fellay, Ioannis Xenarios, Laurent Kaiser, and Philippe Le Mercier. Virosaurus a reference to explore and capture virus genetic diversity. *Viruses*, 12(11), 2020. doi: 10.3390/v12111248.

Norman Goodacre, Aisha Aljanahi, Subhiksha Nandakumar, Mike Mikailov, and Arifa S Khan. A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. *mSphere*, 3(2), 2018. doi: 10.1128/mSphereDirect.00069-18.

Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.*, 29(7):644–652, 2011. doi: 10.1038/nbt.1883.

Nathan D Grubaugh, Karthik Gangavarapu, Joshua Quick, Nathaniel L Matteson, Jaqueline Goes De Jesus, Bradley J Main, Amanda L Tan, Lauren M Paul, Doug E Brackney, Saran Grewal, Nikos Gurfield, Koen K A Van Rompay, Sharon Isern, Scott F Michael, Lark L Coffey, Nicholas J Loman, and Kristian G Andersen. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.*, 20(1):8, 2019. doi: 10.1186/s13059-018-1618-7.

Jiarong Guo, Ben Bolduc, Ahmed A Zayed, Arvind Varsani, Guillermo Dominguez-Huerta, Tom O Delmont, Akbar Adjie Pratama, M Consuelo Gazitúa, Dean Vik, Matthew B Sullivan, and Simon Roux. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*, 9(1):37, 2021. doi: 10.1186/s40168-020-00990-y.

Kazutaka Katoh, Kazuharu Misawa, Kei-Ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.*, 30 (14):3059–3066, 2002. doi: 10.1093/nar/gkf436.

Kristopher Kieft, Zhichao Zhou, and Karthik Anantharaman. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, 8(1):90, 2020. doi: 10.1186/s40168-020-00867-0.

Kristopher Kieft, Alyssa Adams, Rauf Salamzade, Lindsay Kalan, and Karthik Anantharaman. vRhyme enables binning of viral genomes from metagenomes. *Nucleic Acids Res.*, 50(14):e83, 2022. doi: 10.1093/nar/gkac341.

Gregory M Kurtzer, Vanessa Sochat, and Michael W Bauer. Singularity: Scientific containers for mobility of compute. *PLoS One*, 12(5):e0177459, 2017. doi: 10.1371/journal.pone.0177459.

Ben Langmead, Christopher Wilks, Valentin Antonescu, and Rone Charles. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*, 35(3):421–432, 2019. doi: 10.1093/bioinformatics/bty648.

Jeroen F J Laros. HUMID: HUMID: reference free FastQ deduplication, 2025.

Dinghua Li, Ruibang Luo, Chi-Man Liu, Chi-Ming Leung, Hing-Fung Ting, Kunihiko Sadakane, Hiroshi Yamashita, and Tak-Wah Lam. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, 102:3–11, 2016. doi: 10.1016/j.ymeth.2016.02.020.

Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*, 2013.

Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018. doi: 10.1093/bioinformatics/bty191.

Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006. doi: 10.1093/bioinformatics/btl158.

Dmitry Meleshko, Iman Hajirasouliha, and Anton Korobeynikov. coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies. *Bioinformatics*, 2021. doi: 10.1093/bioinformatics/btab597.

Peter Menzel, Kim Lee Ng, and Anders Krogh. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nat. Commun.*, 7:11257, 2016. doi: 10.1038/ncomms11257.

D Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014(239):2, 2014. doi: 10.5555/2600239.2600241.

Stephen Nayfach, Antonio Pedro Camargo, Frederik Schulz, Emiley Eloe-Fadrosh, Simon Roux, and Nikos C Kyrpides. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.*, 39(5):578–585, 2021. doi: 10.1038/s41587-020-00774-7.

Brian D Ondov, Nicholas H Bergman, and Adam M Phillippy. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics*, 12(1):385, 2011. doi: 10.1186/1471-2105-12-385.

Brian D Ondov, Gabriel J Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B Buck, and Adam M Phillippy. Mash screen: high-throughput sequence containment estimation for genome discovery. *Genome Biol.*, 20(1):232, 2019. doi: 10.1186/s13059-019-1841-x.

Guillermo Rangel-Pineros, Alexandre Almeida, Martin Beracochea, Ekaterina Sakharova, Manja Marz, Alejandro Reyes Muñoz, Martin Hölzer, and Robert D Finn. VIRify: an integrated detection, annotation and taxonomic classification pipeline using virus-specific protein profile hidden markov models. *bioRxiv*, page 2022.08.22.504484, 2022. doi: 10.1101/2022.08.22.504484.

Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016. doi: 10.7717/peerj.2584.

Samuel S Shepard, Sarah Meno, Justin Bahl, Malania M Wilson, John Barnes, and Elizabeth Neuhaus. Viral deep sequencing needs an adaptive approach: IRMA, the iterative refinement meta-assembler. *BMC Genomics*, 17(1):708, 2016. doi: 10.1186/s12864-016-3030-6.

Tom Smith, Andreas Heger, and Ian Sudbery. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.*, 27(3):491–499, 2017. doi: 10.1101/gr.209601.116.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, 35(11):1026–1028, 2017. doi: 10.1038/nbt.3988.

Md Vasimuddin, Sanchit Misra, Heng Li, and Srinivas 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, 2019. doi: 10.1109/IPDPS.2019.00041.

Derrick E Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome Biol.*, 20(1):257, 2019. doi: 10.1186/s13059-019-1891-0.

Andrzej Zielezinski, Adam Gudyś, Jakub Barylski, Krzysztof Siminski, Piotr Rozwalak, Bas E Dutilh, and Sebastian Deorowicz. Ultrafast and accurate sequence alignment and clustering of viral genomes. *Nat. Methods*, 22(6):1191–1194, 2025. doi: 10.1038/s41592-025-02701-7.