



香港城市大學
City University of Hong Kong

專業 創新 胸懷全球
Professional · Creative
For The World

CityU Scholars

ViromeFlowX

a Comprehensive Nextflow-based Automated Workflow for Mining Viral Genomes from Metagenomic Sequencing Data

Wang, Xiaokai; Ding, Zhimin; Yang, Ying; Liang, Lifeng; Sun, Yingshuai; Hou, Chaojian; Zheng, Yuning; Xia, Yan; Dong, Lixin

Published in:
Microbial Genomics

Published: 01/01/2024

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
[Go to record](#)

Published version (DOI):
[10.1099/mgen.0.001202](https://doi.org/10.1099/mgen.0.001202)

Publication details:
Wang, X., Ding, Z., Yang, Y., Liang, L., Sun, Y., Hou, C., Zheng, Y., Xia, Y., & Dong, L. (2024). ViromeFlowX: a Comprehensive Nextflow-based Automated Workflow for Mining Viral Genomes from Metagenomic Sequencing Data. *Microbial Genomics*, 10(2), Article 001202. <https://doi.org/10.1099/mgen.0.001202>

Citing this paper

Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights

Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission

Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy

Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.

ViromeFlowX: a Comprehensive Nextflow-based Automated Workflow for Mining Viral Genomes from Metagenomic Sequencing Data

Xiaokai Wang^{1,†}, Zhimin Ding^{2,†}, Ying Yang², Lifeng Liang², Yingshuai Sun², Chaojian Hou¹, Yuning Zheng², Yan Xia^{2,*} and Lixin Dong^{1,*}

Abstract

Understanding the link between the human gut virome and diseases has garnered significant interest in the research community. Extracting virus-related information from metagenomic sequencing data is crucial for unravelling virus composition, host interactions, and disease associations. However, current metagenomic analysis workflows for viral genomes vary in effectiveness, posing challenges for researchers seeking the most up-to-date tools. To address this, we present ViromeFlowX, a user-friendly Nextflow workflow that automates viral genome assembly, identification, classification, and annotation. This streamlined workflow integrates cutting-edge tools for processing raw sequencing data for taxonomic annotation and functional analysis. Application to a dataset of 200 metagenomic samples yielded high-quality viral genomes. ViromeFlowX enables efficient mining of viral genomic data, offering a valuable resource to investigate the gut virome's role in virus-host interactions and virus-related diseases.

DATA SUMMARY

This study demonstrates the use of openly available BioProject accession PRJNA557323 raw metagenomic sequence reads as a representative dataset. The ViromeFlowX tool, which can be accessed by all researchers on GitHub at (<https://github.com/01life/ViromeFlowX>), provides detailed instructions. Besides, two additional publicly available datasets, one consisting of virus-free genomes (<https://zenodo.org/records/4297575>) [1] and the other containing virus-infected samples (PRJNA698986) [2] are further tested by ViromeFlowX, and all corresponding details and results are available at https://github.com/01life/ViromeFlowX_V1_DATA/tree/main/verified_result.

INTRODUCTION

The gut microbiome dramatically impacts human health and disease, and its infecting viruses are likely equally important [3–6]. Emerging views suggest that the gut virome plays a vital role in regulating homeostasis and disease progression through interactions with the bacteriome and the human immune system [7]. These studies have revealed significant associations between changes in phage composition and several diseases [8–11], such as metabolic syndrome [12, 13], necrotizing enterocolitis [14], inflammatory bowel disease [15–18], and type two diabetes mellitus [19, 20]. Recent cross-cohort meta-analyses have identified specific gut viral biomarkers, such as phages of *Porphyromonas*, *Fusobacterium*, and *Hungateella*, enriched in colorectal cancer patients, suggesting their potential as treatment targets [21]. Johansen *et al.*'s research has highlighted the gut virome's impact on longevity. They found that centenarians have a more diverse and slower-working gut virome, which could potentially affect their metabolism [22]. These studies highlight the significance of exploring viral communities founded in metagenomes to understand their effects on human health.

Received 15 August 2023; Accepted 02 February 2024; Published 21 February 2024

Author affiliations: ¹Department of Biomedical Engineering, City University of Hong Kong, Hong Kong 999077, PR China; ²01Life Institute, Shenzhen, PR China.

***Correspondence:** Lixin Dong, l.x.dong@cityu.edu.hk; Yan Xia, xiayan@01lifetech.com

Keywords: Nextflow-based; Viral Genomes Mining; ViromeFlowX.

[†]These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Nine supplementary tables are available with the online version of this article.

001202 © 2024 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author's institution.

Impact Statement

Gut viruses are a topic of significant interest for their roles in shaping the composition of the gut microbiome and regulating human homeostasis. Mining viral genomes from shotgun metagenomic sequencing data is vital for understanding viral communities, virus-bacterium-host interactions, and disease associations. The current tools for viral genome analysis still exhibit certain constraints. Therefore, we have developed ViromeFlowX, a Nextflow-based, highly flexible, modular, customizable, and easy-to-use automatic monitoring pipeline with various applications and databases. ViromeFlowX seamlessly processes raw reads, assembles viral genomes, and provides taxonomic annotations and gene functional analyses. Validated using real data, the pipeline efficiently reduces analysis time, obtains high-quality viral contigs, and offers comprehensive taxonomic and functional annotations. The pipeline's user-friendly nature, scalability, and customizable execution make it a potent tool for unravelling the secrets of viral communities, paving the way for breakthroughs in virome research.

Despite the rapid development of metagenomics techniques and the burgeoning availability of microbial sequencing data, previous research has predominantly focused on bacteria, ignoring the intricate interactions between bacterial and viral communities [23]. However, the advent of shotgun sequencing data offers a valuable avenue for comprehending the dynamic relationship between phages and bacterial ecosystems, eliminating the necessity for viral particle-specific enrichment. While several pipelines have emerged to uncover viral communities from metagenomic data, they still exhibit certain constraints. For instance, ViWrap lacks provisions for quality control and assembly of raw reads, gene prediction, and functional annotation of viral contigs [24]. Similarly, while VirMAP facilitates iterative assembly enhancement and the detection of viral sequences from metagenomic datasets, it does not encompass gene prediction and functional annotation of viral contigs [25].

In response to these challenges and to optimize the analysis of viral genomes within metagenomic data, we have developed ViromeFlowX, a Nextflow-based pipeline. This pipeline efficiently handles viral genomes, from processing raw next-generation sequencing reads to providing viral taxonomic annotation and gene functional analysis. Leveraging the power of Nextflow [26], ViromeFlowX guarantees separate module execution, efficient parallel processing, error resilience, and traceable execution history. It effortlessly adapts to various environments, including local machines, high-performance computing, and cloud infrastructures, enabling swift pipeline development and parameter customization.

ViromeFlowX leverages the capabilities of Virfinder [27], Virsorter2 [2], and robust viral identification tools, to efficiently detect viral contigs. These contigs undergo taxonomic classification through four distinct strategies. The pipeline calculates the abundance of viral contigs, gene functional levels, and a seven-level hierarchical taxonomy. Kraken2 [28], a swift taxonomic classifier, is also integrated for viral taxonomy identification and quantification to avoid overlooking low-abundance viruses. The user-friendly interface, automated task tracking, and well-organized result directory minimize usability barriers. Using ViromeFlowX, researchers can harness its capabilities to effectively mine virus data, deepen our understanding of viral populations, and elucidate the intricate relationship between human health and viruses.

METHODS AND RESULTS**ViromeFlowX workflow**

ViromeFlowX is an easy-to-install, user-friendly, parameter-clear, and highly scalable virome analysis tool. It consists of various software and is divided into five modules, each focusing on an important step in the analysis process. With a single command, it orchestrates quality control and genome assembly (1), identification and taxonomic assignment of viral contigs (2) and (3), gene prediction and functional annotation (4), and viral taxonomic classification via Kraken2 (5) (Fig. 1). Generally, the software employs default parameters, while pivotal parameters can be conveniently accessed and modified within the Nextflow config file. This modular arrangement simplifies parameter manipulation, enhancing user control. Resource allocation for each software is optimized, and outcomes are thoughtfully organized within the default 'work' output directory. This thoughtful structuring eases user examination, circumventing the necessity of delving into intricate intermediate files. As a result, the threshold for using the workflow is greatly reduced.

ViromeFlowX offers two execution modes. The recommended approach launches the complete pipeline via a single command, such as 'nextflow run nf-core-virome --input sample.csv --outdir <OUTDIR>'. However, users seeking tailored analyses can selectively omit certain modules by appending '--skip_<module name>' to the command. Notably, ViromeFlowX encapsulates all essential tools and third-party dependencies within a conda environment, harmoniously coexisting with existing programmes. This environment harnesses Python version 3.10.8 and aligns with the R statistical package version 4.2.2. A comprehensive inventory of software and databases leveraged by ViromeFlowX can be found in Table 1, backed by meticulous benchmarking informed by published literature.

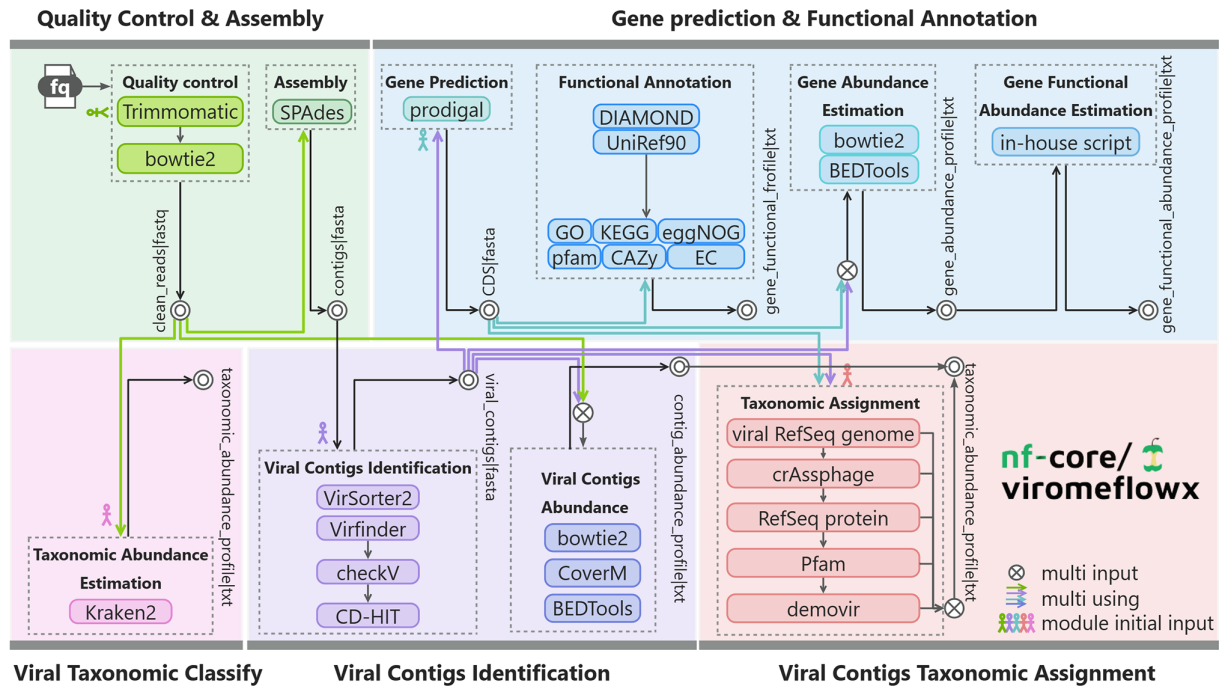


Fig. 1. Schematic overview of the ViromeFlowX. The general flowchart of ViromeFlowX is colour-coded by modules. Coloured lines signify that the file flow will serve as input for multiple subprograms. Intersecting circles indicate that the subprogram requires multiple input files, while concentric circles indicate that the programme at this point has published result files.

Reads quality control and genome assembly

In the initial stage of our analysis, we applied Trimmomatic [29], a tool utilized for filtering low-quality reads, reads containing adapters, or reads that are too short. Subsequently, our custom-built in-house kit, a tool used to quality control developed by ourselves, selectively removed host genome sequences identified by Bowtie2 [30]. This remarkable kit proved to be more efficient and saved valuable resources, surpassing the capabilities of the KneadData software (<https://github.com/biobakery/kneaddata>). Each metagenomic sample underwent individual contig assembly using SPAdes [31] with the ‘-meta’ option and default parameters. Contigs below 1 kb were excluded from further analysis, ensuring a focus on more substantial genetic components.

Viral contig extracted from raw assemblies

The extraction of viral contigs from metagenomic data was a meticulous process achieved through the use of two cutting-edge tools: VirSorter2 [31] and Virfinder [27]. VirSorter2 seamlessly integrates sequence similarity and distinctive viral-like features to classify entire contigs within the dataset. Simultaneously, Virfinder relies on the nuanced analysis of k-mer signatures to pinpoint viral contigs exclusively. The union of the results obtained from both tools were then merged to ensure comprehensive coverage of potential virus sequences. To ensure the quality and reliability of the identified viral genomes, we employed CheckV [32], a powerful tool that assesses various aspects of single-contig viral genomes. This process entails detecting host contaminations in integrated proviruses, gauging the completeness of genome fragments, and recognizing fully closed genomes. The amalgamation of CheckV results provided a comprehensive view of the potential viruses harboured within each sample. We employed CD-HIT [33], a highly efficient clustering algorithm that curates a non-redundant collection of virus contig sequences to streamline the analysis. This step significantly optimized the analytical process, especially when dealing with multiple samples, by reducing redundancy and conserving computational resources. Notably, this approach alleviated resource demands for subsequent analysis, potentially lowering the threshold for large-scale studies. With the aim of quantification, we determined the contigs' coverage using the complementary capabilities of CoverM (<https://github.com/wwood/CoverM>) and BEDTools [34]. Further refinement involved a custom Perl script, engineered to convert coverage values into an abundance profile, quantified as Reads Per Kilobase per Million mapped reads (RPKM). This meticulous process underpinned the accuracy and robustness of our subsequent viral analysis.

Gene prediction and functional annotation

To uncover the genetic structure of viral contigs, we initiated gene prediction by utilizing Prodigal [35], which expertly identifies open reading frames (ORFs). These ORFs were subsequently aligned with precision to the UniProt Reference Clusters (UniRef)

Table 1. Comprehensive list of third-party tools and databases used in ViromeFlowX

Software or database	Version	Summary	Links
Software			
Trimmomatic	0.36	Read trimmer for Illumina NGS data	https://github.com/timflutre/trimmomatic
Bowtie2	2.5.0	A fast and sensitive gapped read aligner	https://github.com/BenLangmead/bowtie2
SPAdes	3.11	<i>De novo</i> assembly of genome sequencing data	https://github.com/ablab/spades
VirSorter2	2.1	Pipeline to identify viral sequences from (meta)genomic data	https://github.com/jiarong/VirSorter2
VirFinder	1.1	A novel k-mer-based tool for identifying viral sequences from assembled metagenomic data	https://github.com/jessieren/VirFinder
CD-HIT	4.8.1	Accelerated clustering of the next-generation sequencing data	http://cd-hit.org .Contact
checkV	1.0.1	Assesses the quality and completeness of metagenome-assembled viral genomes	https://bitbucket.org/berkeleylab/CheckV
blastn, blastp	2.11.0	A web interface for sequence similarity search	http://www.ncbi.nlm.nih.gov/blast
taxonkit	0.7.2	A practical and efficient NCBI taxonomy toolkit, also supports creating NCBI-style taxdump files for custom taxonomies like GTDB/ICTV	https://github.com/shenwei356/taxonkit
CoverM	0.6.1	Read coverage calculator for metagenomics	https://github.com/wwood/CoverM
bedtools	2.30.0	A flexible suite of utilities for comparing genomic features	https://github.com/arq5x/bedtools2
Prodigal	2.6.3	Prokaryotic gene recognition and translation initiation site identification	https://github.com/hyattpd/Prodigal
Prokka	1.14.6	Rapid prokaryotic genome annotation	https://github.com/tseemann/prokka
diamond	2.0.6	Sensitive protein alignments at tree-of-life scale	https://github.com/bbuchfink/diamond
demovir		Taxonomic classification of viruses at Order and Family level	https://github.com/feargalr/Demovir
Database			
uniref90	01/2019	A database consisting of clustered sets of sequences from UniProtKB and selected UniParc records	http://www.uniprot.org/uniref
NCBI RefSeq sequence	10/6/2021	A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein	https://www.ncbi.nlm.nih.gov/refseq/
crAssphage	NC_024711.1	The most abundant virus in the human gut	https://www.sciencedirect.com/science/article/pii/S1931312818305249
eggNOg	5.0	A hierarchical, functionally and phylogenetically annotated orthology resource	http://eggnog5.embl.de/#/app/home
pfam	11/1/2021	A database of conserved protein families and domains.	http://pfam.xfam.org/

90 database (release 01/2019) [36] using DIAMOND [37], a robust alignment tool. For exact results, our Perl script identified the best match based on the results from DIAMOND. The selection criteria were strict. Only records with a coverage of over 80% and an identity of more than 50% were included as annotation records. We continued transforming UniRef90 database items into actionable annotation insights for Gene Ontology (GO) [38, 39], EGGNOG [40], Kyoto Encyclopaedia of Genes and Genomes (KEGG) [41], PfamA [42], EC [41], and the Carbohydrate-Active EnZymes database (CAZy) [43] databases. In-house scripts were used to map UniRef90 entries to these annotations. The Bowtie2 [30] tool and BEDTools [34] have proven helpful in measuring the number of genes present. Bowtie2 efficiently quantified gene content, while bedtools accurately gauged the coverage of each cluster across multiple samples. The resulting gene abundance table underwent normalization via the RPKM method, yielding a comprehensive gene abundance profile table. Our in-house R script was employed to enhance further functionality to convert

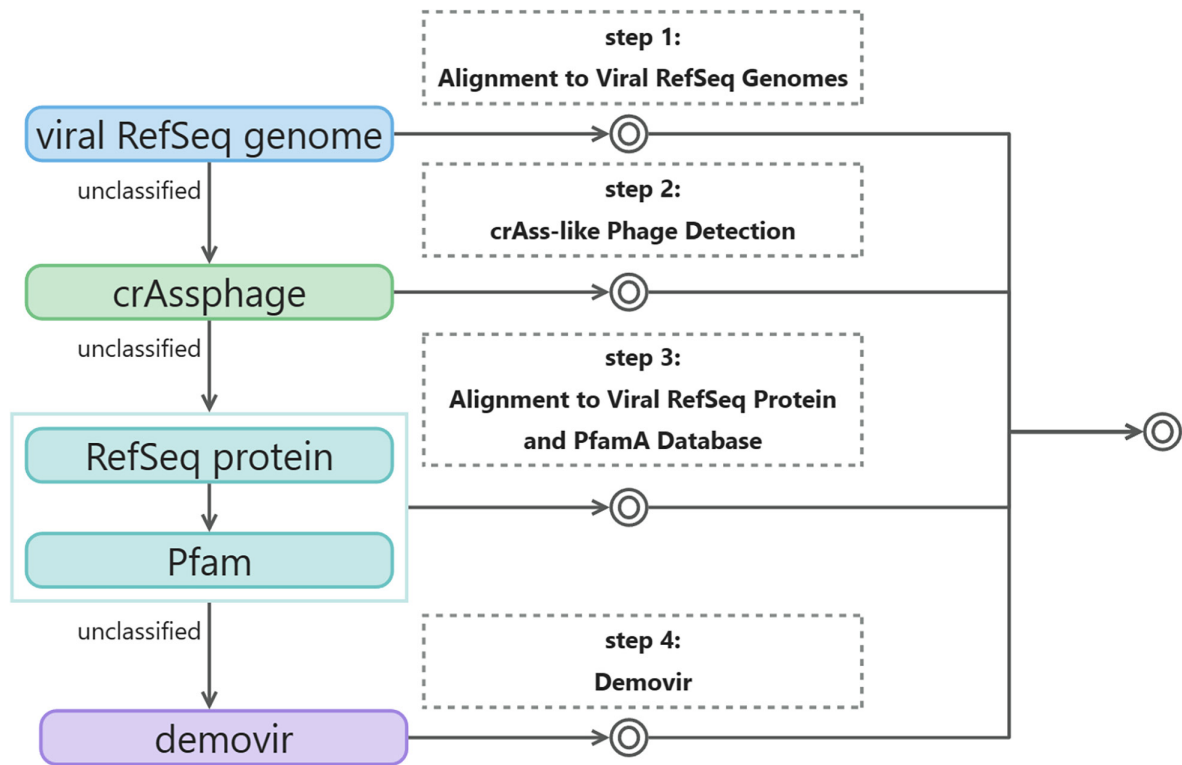


Fig. 2. Workflow of four-step viral taxonomic classification of viruses. Extracted viral contigs were sequentially assigned to a seven-level hierarchical taxonomy. The flowchart is colour-coded by steps, concentric circles indicate that the programme at this point has published result files.

UniRef90 gene abundance into actionable profiles within the functional database. This intricate gene prediction and available annotation process greatly enriched the contextual understanding of the viral contigs and their potential roles.

Viral taxonomic classification

Building upon K. Fujimoto *et al.*'s [44] viral taxonomic classification method, we refined and enhanced viral contig annotations to better understand their biological significance. We have created viral genome and protein databases to classify the viral contigs. The process of virus classification annotation consists of four sequential steps (Fig. 2). First, we aligned the contigs with viral RefSeq genomes [45]. If the contigs did not align, we conducted further analysis using crAss-like phage detection [46] to try and obtain classification information. For contigs that were still not classified, we attempted to identify similar viral proteins at the protein level to infer taxonomic classification. Lastly, contigs that showed no similarity to known proteins were annotated using Demovir as a final step. Here are the specific methods used for each of the four steps:

Step 1: Alignment to Viral RefSeq Genomes

We harnessed the viral RefSeq sequences (downloaded on 10 June 2021) [45], encompassing 14679 genomes and 546060 proteins from NCBI. Leveraging blastn [47], we aligned viral contig sequences against these genomes, employing an E-value threshold of $1e-10$. The taxonkit [48] coupled with the NCBI taxonomy pinpointed the lowest common ancestor (LCA) based on the top five hits, thus attributing classification to the contigs.

Step 2: crAss-like Phage Detection

The annotation of contig ORFs facilitated crAss-like phage detection. We employed blastp [37] to query amino acid sequences of prototypical crAssphage genetic signatures (p-crAssphage, NC_024711.1) against contig ORFs, using an E-value cutoff of $1e-10$.

Step 3: Alignment to Viral RefSeq Protein and PfamA Database

Predicted ORFs lacking classification in previous steps underwent a blastp search against the viral RefSeq protein database from Step 1, using an E-value threshold of $1e-10$ and a bitscore exceeding 50. Leveraging phage structural proteins in the PfamA [42] annotation, viral contigs were classified based on specific proteins. This PfamA family included phage tail protein, phage tail sheath protein, or Microviridae capsid protein gene, attributing contigs to Caudovirales, Myoviridae, or Microviridae.

Step 4: Demovir

Contigs not classified earlier received family-level annotations via the Demovir script (<https://github.com/feargalr/Demovir>) with default parameters and database. This script hunted for homologies in amino acid sequences between contig-encoded proteins and a viral subset of the TrEMBL database. Taxonomic assignment employed a voting approach.

These four steps were ultimately amalgamated into comprehensive taxonomic classifications for each contig. The results, encompassing taxonomic levels from phylum to species, were exported in BIOM-format [49]. Kraken2 [28] was harnessed using a viral database to address low-abundance viruses, which are often overlooked. Bringing Kraken2 into our workflow enabled an alternate perspective on viral sample composition, thus broadening researchers' insights into viral diversity.

Results organization

The final output directory is meticulously organized, ensuring easy access to comprehensive results. Here's an overview of the layout:

01. QC:

This section showcases the outcomes of the read quality control step. Inside, you'll discover the clean reads for each sample, formatted in fastq. These reads are the result of Trimmomatic application.

02. assembly:

Here, the results of viral assemblies are stored in the directory. This folder contains the contig fasta files for each sample.

03. identify:

This dedicated folder encompasses distinct sub-folders for VirFinder and VirSorter2 outcomes. Moreover, a consolidated folder brings together results from both runs.

04. predict:

This directory contains a refined subset of viral contigs post-CD-HIT deduplication. It also houses intermediary CD-HIT results. Expect to find Prodigal gene predictions on deduplicated contigs enriched with nucleotide sequences, protein sequences, and GFF format annotations.

05. classify:

This directory is the result of viral taxonomic classification insights. It incorporates the results from the four classification steps, leading to the final merging of classification outcomes.

06. abundance:

Abundance profiling at both contig and gene levels. Every sample has unique characteristics, and when all the samples are combined, it enhances the dataset even more.

07. functional:

Annotations abound in this repository, featuring a symphony of results against uniref90 and GO, EggNOG, KEGG, PfamA, EC, and CAZy databases.

08. profile:

Witness the profiles come to life! This folder contains abundance profiles at various taxonomic levels, spanning viral entities and functional databases. The Kraken results are also in this folder.

pipeline_info:

This directory encapsulates Nextflow's insights into the workflow run. It offers a window into job execution status and the detailed utilization of computing resources, guiding users on the journey.

This meticulous arrangement empowers you to seamlessly navigate through the intricacies of your results, enhancing your analytical prowess.

Performance excellence

Leveraging nf-core's adept management of task execution and optimal allocation of computing resources, ViromeFlowX works well across datasets of varying dimensions. We delved into three metagenomic samples through meticulous evaluation, each harbouring an average raw data size of 6G. Employing a robust elastic computing cluster boasting 32 CPUs and a generous 64G memory capacity, our analysis executed seamlessly, culminating in completion within a mere 6h.

Table 2. Comparison of ViromeFlowX functionality with commonly used metagenome assembly and binning pipelines

Functionality	nf-core/ViromeFlowX	nf-core/viralrecon	VirWrap
Support Nanopore Sequencing data	×	√	×
QC	√	√	×
Assembly	√	√	×
Taxonomic Abundance Estimation	√	×	×
Virus identification and annotation	√	×	√
Virus quality characterization	√	×	√
Functional Annotation	√	×	×
Virus binning and clustering	×	×	√
Virus taxonomy classification	√	×	√
Virus host prediction	×	×	√
Variant calling	×	√	×

As we took on the task of testing bulk metagenomes, we discovered the increased strength and reliability of ViromeFlowX. With a comprehensive study of 200 human gut microbiome samples [50], each with an average of 6G sequencing data, totaling 658G, we conducted this analysis using the Amazon Cloud platform, fortified with a CPU limit 3200. The journey concluded within 3 days and 12 h, underlining the workflow's efficiency and batch-processing prowess. This impressive performance slashes analysis time and grants analysts respite from continuous job monitoring.

Beyond streamlined analysis, ViromeFlowX offers a panoramic view of metagenomic virome analysis. From the inception of read quality control to the culmination of taxonomy abundance assignment and multi-faceted database annotation, the workflow's coverage is comprehensive. Table 2 illustrates that ViromeFlowX may presently lack modules for viral host prediction. Its advantage lies in taxonomy annotation and abundance profiling, surpassing other commonly employed metagenomic virome analysis suites. With an elegantly structured workflow and an expansive spectrum of functionalities, ViromeFlowX is a beacon of cutting-edge technology for the meticulous dissection of viral communities embedded within metagenomic landscapes.

Validation and visualization on real dataset

To thoroughly validate the efficacy of ViromeFlowX, we conducted an in-depth analysis of a dataset comprised of 200 published faecal metagenomes (Table S1, available in the online version of this article) from healthy individuals in Hong Kong. These data were assembled into 7930827 raw contigs, with an average assembly of 39654 contigs per sample, and a mean contig length of 4871 kb (Fig. 3a, b; Tables S2 and S3). With the integration of VirSorter and Virfinder, our analysis identified 598566 viral contigs, yielding a recall rate of 7.54% (Fig. 3c and Table S4). These viral contigs were refined to 3007 per sample, boasting an average length of 3477. A rigorous CheckV evaluation highlighted that 65.71% of the contigs could not be conclusively determined. Among the remainder, 33.06% were categorized as low-quality, while the prevalence of medium-quality, high-quality, and complete contigs stayed below 1%. Importantly, our analysis solidified 1066 complete and 2340 high-quality viral contigs (Fig. 3d and Table S5). During virus classification annotation, the contigs were assigned to 14 phyla, 21 classes, and 53 families (Fig. 3e and Table S6). A notable domain class emerged as Caudoviricetes, with the leading genera or species affiliated with Nucleocytoviricota (Fig. 3f) and Uroviricota (Fig. 3G) phyla. Then, Prodigal was utilized to predict open reading frames (ORFs). A total of 1548275 ORFs were categorized into 267525 gene clusters, incorporating alignment thresholds of 80% for length and 50% for identity. Our analysis delved into virome functions, anchored in COG, KEGG and GO databases (Tables S7–S9), complete with corresponding gene count fractions (Fig. 3f–h). These visually impactful findings provided an instinctive grasp of the quantified characteristics of the viral community, virus quantities, and gene counts of various databases.

DISCUSSION

ViromeFlowX is a Nextflow pipeline that streamlines the process of mining viral genomes from metagenomic data. What sets ViromeFlowX apart from other pipelines is its ability to not only assemble viral genomes but also perform comprehensive viral taxonomy classification and abundance estimation using state-of-the-art taxonomy classifiers. This feature-rich pipeline is specifically designed to handle large metagenomic samples and consistently deliver reproducible results, ensuring reliable outcomes for diverse research questions and experimental setups.

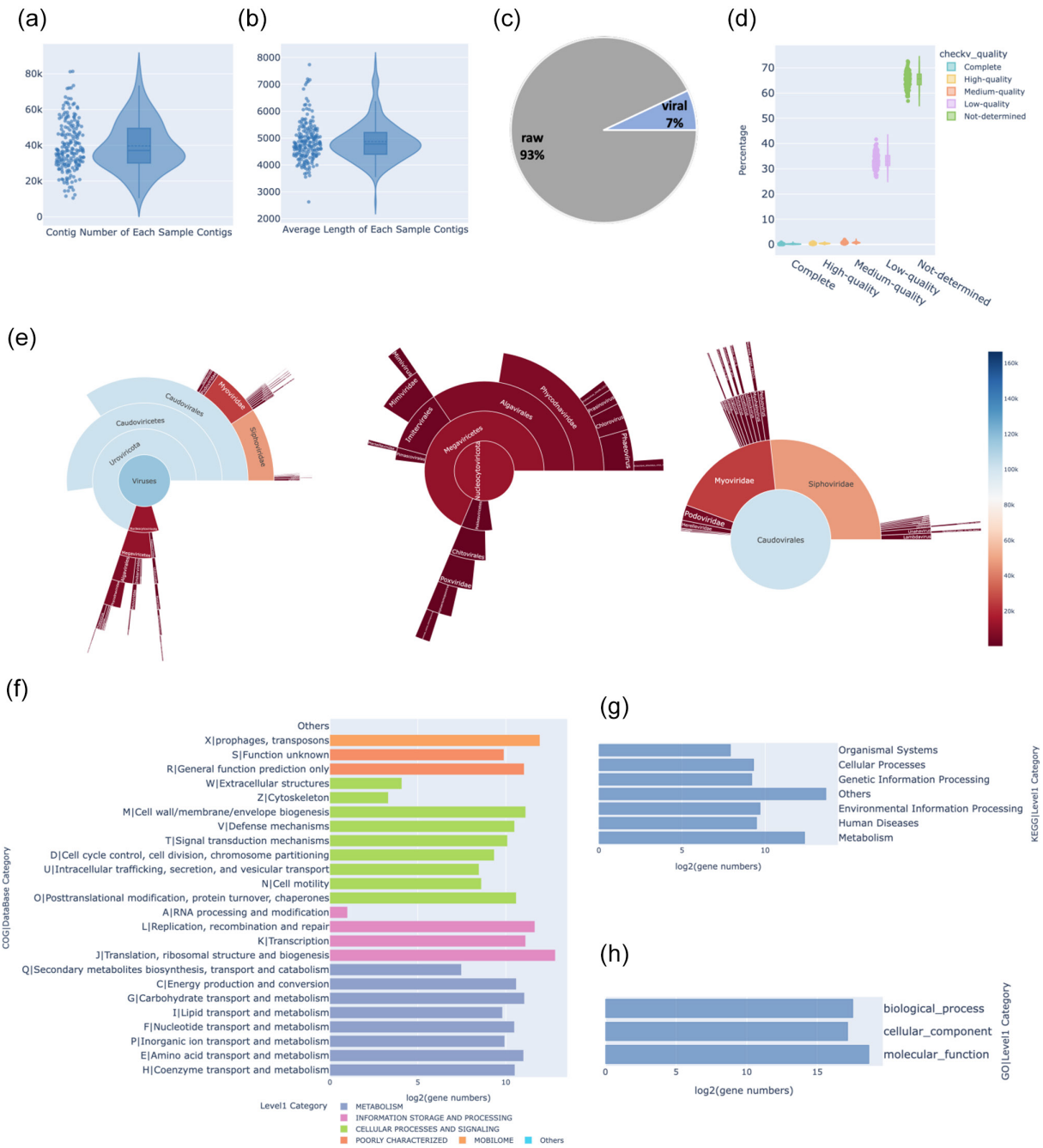


Fig. 3. Visualisations of virus statistics. The violin chart (a) showcases contig distribution across samples, while (b) portrays average contig length. The pie chart (c) visually encapsulates viral contig presence. The violin plot (d) reveals the quality distribution of viral contigs. The sunburst chart (e) maps viral taxonomy and highlights Nucleocytoviricota and Uroviricota. Bar charts (f–h) depict gene counts in COG, KEGG, and GO databases.

By leveraging Nextflow, ViromeFlowX automates the execution of analyses with built-in error recovery and checkpoint resumption capabilities. The visual run report provides users with a transparent view of the workflow execution process, promoting ease of interpretation and monitoring.

The modular organization of ViromeFlowX allows users to customize their analyses by skipping unnecessary steps and focusing on specific functionalities. This flexibility empowers researchers to tailor the pipeline to their unique requirements and compare different settings effectively.

While ViromeFlowX represents a significant advancement in viral genomic analysis, continuous maintenance and development are essential to keep pace with evolving analytical approaches. As part of future improvements, the upcoming version, ViromeFlowX2, is poised to introduce additional features such as binning techniques [51] and in-depth analysis of viral genomes, including host prediction, further enhancing its utility and applicability. ViromeFlowX2 is committed to achieving a custom expandable virus database, allowing users to add new knowledge of virus data to prepare for the potential emergence of new mutated viruses in the future.

One aspect that requires attention in the current code is the use of intermediate ‘work’ directories, leading to a fourfold increase in storage requirements for running a single sample. This storage inefficiency arises from Nextflow’s retention of all result files, including temporary ones, unless manually deleted. To mitigate this issue, the upcoming ViromeFlowX version will implement an automated mechanism to remove unnecessary intermediate files, optimizing storage usage and streamlining the overall workflow execution process.

CONCLUSION

ViromeFlowX is a robust and user-friendly solution that empowers researchers to investigate viral communities in metagenomic datasets. Its automation, adaptability, and integration of advanced tools make it a valuable resource for viral genomic data analysis. By using ViromeFlowX, the scientific community can better understand the complex relationships between viruses and their hosts, leading to discoveries and advancements in virome research.

Funding information

This work was partially supported by the National Natural Science Foundation of China (grant no. 62127810) and the University Grants Committee (project nos. 11213720, 11219419, and 11217221).

Author contributions

Methodology: X.W., Y.Y., Y.S., Y.Z. Writing – Original Draft: X.W., Z.D., L.L. Data Curation: X.W., Z.D., L.L., C.H. Writing – Review and Editing: all authors. Supervision: Y.X., L.D. Project administration: Y.X., L.D. Funding acquisition: Y.X., L.D.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

- Lou YC, Chen L, Borges AL, West-Roberts J, Firek BA, et al. Infant gut DNA bacteriophage strain persistence during the first 3 years of life. *Cell Host Microbe* 2024;32:35–47.
- Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 2021;9:37.
- Cao Z, Sugimura N, Burgermeister E, Ebert MP, Zuo T, et al. The gut virome: A new microbiome component in health and disease. *EBioMedicine* 2022;81:104113.
- Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* 2014;5:4498.
- Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, et al. Rapid evolution of the human gut virome. *Proc Natl Acad Sci U S A* 2013;110:12450–12455.
- Fan Y, Pedersen O. Gut microbiota in human metabolic health and disease. *Nat Rev Microbiol* 2021;19:55–71.
- Qv L, Mao S, Li Y, Zhang J, Li L. Roles of gut bacteriophages in the pathogenesis and treatment of inflammatory bowel disease. *Front Cell Infect Microbiol* 2021;11:755650.
- Zhang Y, Wang R. The human gut phageome: composition, development, and alterations in disease. *Front Microbiol* 2023;14:doi.
- Liang G, Gao H, Bushman FD. The pediatric virome in health and disease. *Cell Host Microbe* 2022;30:639–649.
- Lam S, Bai X, Shkoporov AN, Park H, Wu X, et al. Roles of the gut virome and mycobiome in faecal microbiota transplantation. *Lancet Gastroenterol Hepatol* 2022;7:472–484.
- Hsu CL, Duan Y, Fouts DE, Schnabl B. Intestinal virome and therapeutic potential of bacteriophages in liver disease. *J Hepatol* 2021;75:1465–1475.
- de Jonge PA, Wortelboer K, Scheithauer TPM, van den Born B-JH, Zwinderman AH, et al. Gut virome profiling identifies a widespread bacteriophage family associated with metabolic syndrome. *Nat Commun* 2022;13:3594.
- Bikel S, López-Leal G, Cornejo-Granados F, Gallardo-Becerra L, García-López R, et al. Gut dsDNA virome shows diversity and richness alterations associated with childhood obesity and metabolic syndrome. *iScience* 2021;24:102900.
- Kaelin EA, Rodriguez C, Hall-Moore C, Hoffmann JA, Linneman LA, et al. Longitudinal gut virome analysis identifies specific viral signatures that precede necrotizing enterocolitis onset in preterm infants. *Nat Microbiol* 2022;7:653–662.
- Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 2015;160:447–460.
- Clooney AG, Sutton TDS, Shkoporov AN, Holohan RK, Daly KM, et al. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. *Cell Host Microbe* 2019;26:764–778.
- Zuo T, Kamm MA, Colombel JF, Ng SC. Urbanization and the gut microbiota in health and inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol* 2018;15:440–452.
- Iliev ID, Cadwell K. Effects of intestinal fungi and viruses on immune responses and inflammatory bowel diseases. *Gastroenterology* 2021;160:1050–1066.
- Yang K, Niu J, Zuo T, Sun Y, Xu Z, et al. Alterations in the gut virome in obesity and type 2 diabetes mellitus. *Gastroenterology* 2021;161:1257–1269.
- Fan G, Cao F, Kuang T, Yi H, Zhao C, et al. Alterations in the gut virome are associated with type 2 diabetes and diabetic nephropathy. *Gut Microbes* 2023;15:2226925.
- Chen F, Li S, Guo R, Song F, Zhang Y, et al. Meta-analysis of fecal viromes demonstrates high diagnostic potential of the gut viral

- signatures for colorectal cancer and adenoma risk assessment. *J Adv Res* 2023;49:103–114.
22. Johansen J, Atarashi K, Arai Y, Hirose N, Sørensen SJ, et al. Centenarians have a diverse gut virome with the potential to modulate metabolism and promote healthy lifespan. *Nat Microbiol* 2023;8:1064–1078.
 23. Shkoporov AN, Turkington CJ, Hill C. Mutualistic interplay between bacteriophages and bacteria in the human gut. *Nat Rev Microbiol* 2022;20:737–749.
 24. Zhou Z, Martin C, Kosmopoulos JC, Anantharaman K. ViWrap: a modular pipeline to identify, bin, classify, and predict viral-host relationships for viruses from metagenomes. *Imeta* 2023;2:e118.
 25. Ajami NJ, Wong MC, Ross MC, Lloyd RE, Petrosino JF. Maximal viral information recovery from sequence data using VirMAP. *Nat Commun* 2018;9:3205.
 26. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, et al. Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;35:316–319.
 27. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 2017;5:69.
 28. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;20:257.
 29. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
 30. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.
 31. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes *de novo* assembler. *Curr Protoc Bioinformatics* 2020;70:e102.
 32. Nayfach S, Camargo AP, Schulz F, Eloë-Fadrosch E, Roux S, et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* 2021;39:578–585.
 33. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28:3150–3152.
 34. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–842.
 35. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119.
 36. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015;31:926–932.
 37. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using diamond. *Nat Methods* 2021;18:366–368.
 38. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. Gene Ontology: tool for the unification of biology. *Nat Genet* 2000;25:25–29.
 39. Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The Gene Ontology knowledgebase in 2023. *Genetics* 2023;224.
 40. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;47:D309–D314.
 41. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27–30.
 42. Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 2016;44:D279–85.
 43. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* 2009;37:D233–8.
 44. Fujimoto K, Kimura Y, Shimohigoshi M, Satoh T, Sato S, et al. Metagenome data on intestinal phage-bacteria associations aids the development of phage therapy against pathobionts. *Cell Host Microbe* 2020;28:380–389.
 45. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, et al. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res* 2021;49:D1020–D1028.
 46. Guerin E, Shkoporov A, Stockdale SR, Clooney AG, Ryan FJ, et al. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. *Cell Host Microbe* 2018;24:653–664.
 47. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, et al. NCBI BLAST: a better web interface. *Nucleic Acids Res* 2008;36:W5–9.
 48. Shen W, Ren H. TaxonKit: a practical and efficient NCBI taxonomy toolkit. *J Genet Genomics* 2021;48:844–850.
 49. McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience* 2012;1:7.
 50. Yeoh YK, Chen Z, Wong MCS, Hui M, Yu J, et al. Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor. *Gut* 2020;69:1998–2007.
 51. Johansen J, Plichta DR, Nissen JN, Jespersen ML, Shah SA, et al. Genome binning of viral entities from bulk metagenomics data. *Nat Commun* 2022;13:965.

The Microbiology Society is a membership charity and not-for-profit publisher.

Your submissions to our titles support the community – ensuring that we continue to provide events, grants and professional development for microbiologists at all career stages.

Find out more and submit your article at microbiologyresearch.org