scientific reports

Check for updates

OPEN

Accessible viral metagenomics for public health and clinical domains with Jovian

Dennis Schmitz[1,2]✉, Florian Zwagemaker[1], Sam Nooij[1,3], Thierry K. S. Janssens[1], Jeroen Cremer[1], Robert Verhagen[1], Harry Vennema[1], Annelies Kroneman[1], Marion P. G. Koopmans[2], Jeroen F. J. Laros[4,5,6] & Miranda de Graaf[2,6]

The integration of next-generation sequencing into clinical diagnostics and surveillance initiatives is impeded by the lack of data analysis pipelines that align with privacy legislation and laboratory certification protocols. To address these challenges, we developed Jovian, an open-source, virus-focused, metagenomic analysis workflow for Illumina data. Jovian generates scaffolds enriched with pertinent annotations, including taxonomic classification, combined with metrics needed for quality assessment (coverage depth, average GC content, localization of open reading frames, minority single nucleotide polymorphisms), and incorporates host and disease metadata. Interactive web-based reports with an audit trail are generated. Jovian was employed on four systems, hosted by three institutes, utilizing grid-computers, a high-performance compute singular server, and a Windows10 laptop. All systems yielded identical results with matching MD5sums. Comparison with a commercial online reference tool using viral gastroenteritis samples confirmed the identification of the same pathogens. Jovian provides comparable results to a commercially available online reference tool and generates identical results at different institutes with different IT architectures, proving it is portable and reproducible. Jovian addresses bottlenecks in the deployment of metagenomics within public health and clinical laboratories and has the potential to enhance the breadth of surveillance and testing programs, thereby fostering more effective public health interventions.

Keywords Viromics, Public health, Clinics, Diagnostics, Surveillance, Next generation sequencing

Abbreviations

EBI	European Bioinformatics Institute
FAIR	Findability, accessibility, interoperability and reusability
GDPR	General Data Protection Regulation
HIV	Human immunodeficiency virus
HPC	High performance computing
LCA	Lowest common ancestor
LSF	Load sharing facility
MD5	Message Digest 5
NGS	Next generation sequencing
QC	Quality control
RHEL	Red Hat Enterprise Linux
SLURM	Simple linux utility for resource management
SNPs	Single nucleotide polymorphisms
WGS	Whole genome sequencing
WSL	Windows subsystems for linux

[1]National Institute of Public Health and the Environment, Center for Infectious Disease Control, 3720BA Bilthoven, The Netherlands. [2]Viroscience, Erasmus University Medical Center, 3015GB Rotterdam, The Netherlands. [3]Center for Infectious Diseases, Leiden University Medical Center, 2333ZA Leiden, The Netherlands. [4]Department of Bio-Informatics and Computational Services, National Institute of Public Health and the Environment, 3720BA Bilthoven, The Netherlands. [5]Department of Human Genetics, Leiden University Medical Center, 2333ZA Leiden, The Netherlands. [6]These authors contributed equally: Jeroen F.J. Laros and Miranda de Graaf. ✉email: Dennis.Schmitz@rivm.nl

Next-generation sequencing (NGS) metagenomics applied to virus-enriched samples (viromics) offers a comprehensive approach for identifying all viral agents, including potential pathogens, and concurrently generating complete viral genomes with a single assay. These attributes hold great promise for clinical and public health laboratories, where they have the potential to improve or even replace current diagnostic protocols. In addition to diagnostics, the integration of full pathogen genome sequencing has become pivotal in outbreak response and management, as exemplified by its role in the coronavirus disease 2019 (COVID-19) pandemic and in a range of pathogen surveillance programs[1,2]. The use of genetic information contained within complete viral genomes helps in gaining a better understanding of sources and modes of transmission in outbreaks, transmission dynamics, the detection of virulence markers and drug resistance mutations, as well as establishing the relatedness to vaccine strains[2–4].

However, the practical application of viromic data analysis faces obstacles, one of which is the lack of bioinformatics expertise in many institutes. Addressing this challenge necessitates an automated and accessible workflow catering to end-users with varying levels of expertise. For improved applicability, workflows need to be compatible with accreditation procedures, requiring proper documentation of workflows, updates, the validation process, workflow reproducibility, and alignment with patient privacy regulations[5]. Although commercial and cloud-based initiatives offer accessible interfaces[6–8], they often necessitate substantial annual licensing fees or are not compatible with clinical settings due to stringent privacy and security concerns. Moreover, reliance on cloud-based solutions introduces dependencies that add vulnerability to the clinical service, such as potential inaccessibility and extended waiting times resulting from heightened demand. To fulfill the requirements of ISO15189 certification and to provide reliable backup mechanisms and scalability options, an automated analysis should give the same results across institutes. Additionally, relying on external services poses risks when these services go offline[8,9].

The taxonomic annotation of scaffolds via public databases is a central part of viromic analyses. While BLAST[10] is one of the most well-known algorithms, tools such as Kraken2[11] and Centrifuge[12] are also frequently used[13]. These algorithms depend on NCBI databases, which support taxonomic classifications down to the species level, but do not provide reliable subspecies-level assignments. As virus nomenclature evolves over time, inferring subspecies assignments from the names of database matches can be misleading. However, outbreak investigations often require subspecies-level resolution and therefore rely on purpose-built (geno)typing tools based on official nomenclature (Table 1)[14–17].

In response to these challenges and guided by feedback from professionals in public health and clinical research, we have developed a comprehensive metagenomics workflow called Jovian. This workflow is purpose-built for specific use-cases, including the identification of specific (pathogenic) viral taxa, subspecies level (geno) typing for outbreak investigations and the investigation of genetic variants (quasispecies) in both consensus and minority single nucleotide polymorphisms (SNPs).

Jovian is designed to be accessible to non-bioinformaticians after installation, ensuring compliance to data protection regulations, privacy mandates, reproducibility requirements, and quality standards. It is an open source, locally installable and self-contained analytical platform tailored for processing Illumina viromics data. The workflow spans the entire process, from raw Illumina paired-end data to an annotated, interactive, and accessible metagenomics web report.

Jovian uses widely recognized algorithms and databases and was developed with findability, accessibility, interoperability and reusability (FAIR) in mind[18]. It aims to address the limitations of external platforms by providing end users with a robust solution that conforms to established standards such as ISO15189 for medical laboratories and ISO23418 for whole genome sequencing (WGS) of foodborne bacteria. Although the latter focuses on bacteria, its technical aspects are applicable to virus WGS.

Here, we present Jovian's technical underpinnings and compare various taxonomic annotation tools to the BLAST implementation used within Jovian. We validate Jovian's results using a publicly available metagenomic dataset of gastroenteritis cases to a widely used commercially available online workflow[6,19,20] and test its portability and reproducibility by analyzing an identical dataset at three different institutes using four different hardware platforms.

Typing-tool name and input requirements	Typing output	Example output
Norovirus typing-tool[14,15]: *Caliciviridae* scaffolds	Genogroup, polymerase genotype and variant[a], capsid genotype and variant	Norovirus GII.4 Sydney[P4 New Orleans] Sapovirus GII.4
Enterovirus typing-tool[15]: *Picornaviridae* scaffolds	VP1 type and subtype	Enterovirus C, PV-1, Mahoney/Sabin Enterovirus A, CV-A4, (VP1: CV-A4)
Rotavirus A typing-tool: *Rotavirus A* scaffolds	Segment number and cluster type	Rotavirus A, segment 1, R2 Rotavirus A, segment 11, H1
Hepatitis A typing-tool[16]: *Hepatovirus A* scaffolds	Genotype and subtype	Hepatitis A, II.A
HPV typing-tool: *Papillomaviridae* scaffolds	Clustertype	HPV16
Hepatitis E typing-tool[17]: *Orthohepevirus A* scaffolds	Genotype and subtype	Hepatitis E, 3.a
Flavivirus typing-tool: *Flaviviridae* scaffolds	Clustertype and subclustertype	Dengue virus 1, genotype 1

Table 1. Summary of the different virus typing-tools, listing their input, output and some examples. Human papillomavirus is abbreviated to HPV. [a]Exclusively for norovirus.

Implementation

Jovian orchestrates the transformation of raw, demultiplexed, Illumina paired-end data into annotated metagenomics reports (Fig. 1). It uses Snakemake[21] as a workflow engine, and BioConda[22] and Singularity[23] to manage dependencies and computational mobility, respectively. The workflow was optimized to allow the installation of the complete workflow in a single operation. Here we describe Jovian with a default database setup via the `--install-databases` flag, as described on GitHub (https://github.com/DennisSchmitz/jovian). While Jovian is primarily intended to be used via the interactive report, a summary of Jovian's output file locations, formats and a brief description of their intent and meaning is shown in Supplementary Table S1.

Quality control (QC) is initiated by eliminating subpar data using Trimmomatic[24], with a sliding window of five nucleotides and an average Phred 20 quality threshold, and a minimal read length of 50 nucleotides (default values, configurable via the command-line). QC metrics are visually presented through the combined use of FastQC and MultiQC[25,26]. To remove human-related data and ensure general data protection regulation (GDPR) compliance, alignment to human genome build GRCh38 (the Epstein-Barr-virus FASTA record was manually removed)[27] is performed by Bowtie2[28], after which aligned reads are discarded from the input data via samtools[29]. Summary statistics of library fragment lengths are obtained through Picard[30].

The sequenced reads subsequently undergo assembly into scaffolds by metaSPAdes with kmers of 21, 33, 55, 77 and "--only-assembler" settings[31]. Scaffolds smaller than 250 nucleotides (default value, configurable via the command-line) are excluded from further analyses. Taxonomic labels up to species level are assigned by BLAST, employing the NCBI BLAST nucleotide (nt) database[32], with a maximum allowed E-value of 0.05 and $\geq$ 50% coverage of the scaffold by the BLAST hit. This database is downloaded locally, allowing offline analysis. BLAST results with a bitscore $<$ 100 (default value, configurable via the command-line), as well as those assigned with taxid "81077" and "12908"—artificial and unclassified sequences, respectively—or containing the word "construct" or "synthetic" are filtered out by MGKit[33]. Ambiguous taxonomic labeling is corrected via Lowest Common Ancestor (LCA) analysis by MGKit at a 0.97 quantile threshold. To achieve a comprehensive taxonomic annotation up to the superkingdom level, NCBI "new_taxdump/rankedlineage.dmp"[34] is used. Scaffolds from a selection of clinically relevant viral taxa are sent to online typing tools that employ a phylogenetic algorithm for (geno)typing to the subspecies level, as listed in Table 1[14–17].

The subsequent alignment of all reads to the generated scaffolds is performed by BWA[35], with PCR-duplicate identification through samtools. Scaffold alignment metrics are calculated using BBTools[36], and GC content within 50-nucleotide windows is determined by Bedtools[37] and Picard. LoFreq[38] is used for the identification of minority variants. To enhance the scaffolds' contextual understanding, annotations related to host and disease

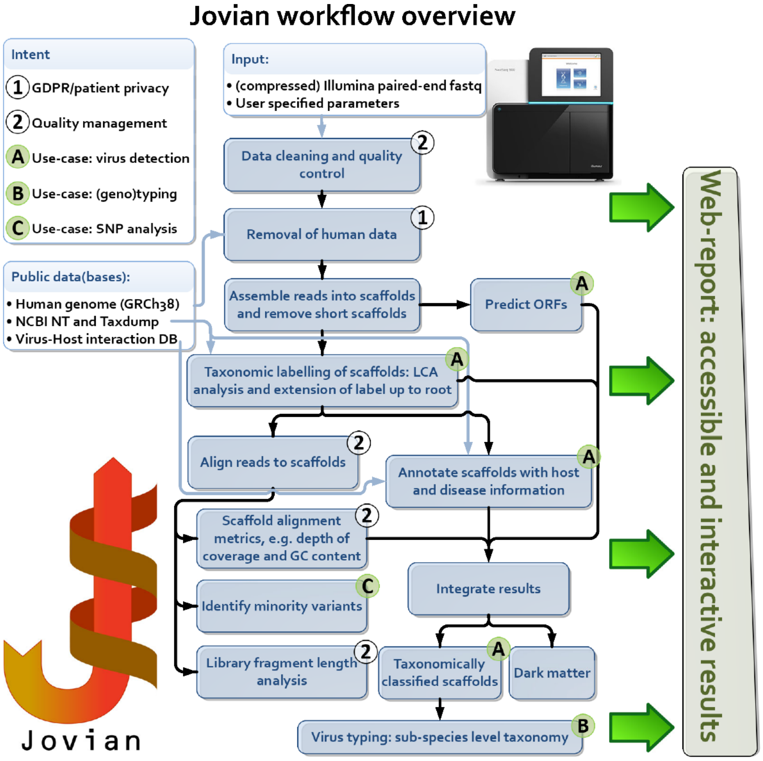

Jovian workflow overview
Intent
1 GDPR/patient privacy
2 Quality management
A Use-case: virus detection
B Use-case: (geno)typing
C Use-case: SNP analysis
Public data(bases):
• Human genome (GRCh38)
• NCBI NT and Taxdump
• Virus-Host interaction DB
Input:
• (compressed) Illumina paired-end fastq
• User specified parameters
Data cleaning and quality control 2
Removal of human data 1
Assemble reads into scaffolds and remove short scaffolds
Predict ORFs A
Taxonomic labelling of scaffolds: LCA analysis and extension of label up to root A
Align reads to scaffolds 2
Annotate scaffolds with host and disease information A
Scaffold alignment metrics, e.g. depth of coverage and GC content 2
Identify minority variants C
Integrate results
Library fragment length analysis 2
Taxonomically classified scaffolds A
Dark matter
Virus typing: sub-species level taxonomy B
Web-report: accessible and interactive results
Jovian

Fig. 1. Schematic representation of Jovian's functional subprocesses. Subprocesses are annotated with their intent as listed in the top left box. Analyses are started via the command-line interface, after which Jovian transforms raw Illumina paired-end FASTQ files into structured tabular flat-files, which are subsequently visualized within the Jovian web-report. It relies on the following public data: NCBI nt, NCBI Taxdump, HuGo build GRCh38 and Virus-Host DB.

information are integrated through Virus-Host DB[39] and NCBI "new_taxdump/host.dmp", while Prodigal[40] predicts Open Reading Frames (ORFs).

Audit trail

A key component of accreditation schemes is the logging of process steps to ensure transparency and reproducibility of any procedure used for generating results that are shared with clinicians and patients. Lists of essential software versions are logged alongside Jovian, and pre-validated and immutable Singularity containers are automatically installed. Singularity recipes are stored on Jovian's GitHub repository, and pre-built containers are publicly available on Sylabs or available on request (https://cloud.sylabs.io/library/ds_bioinformatics). Both the overarching pipeline and the software-version specifications are placed under version control (Git), with versioned releases and a publicly available GitHub repository (https://github.com/DennisSchmitz/jovian).

Each analysis is accompanied by a brief methodological fingerprint, the Git hash, which reproducibly provides a methodological audit trail for the intermediate subprocesses from the input data to the final report. Augmenting this approach, a comprehensive log is maintained that captures every facet ranging from pipeline settings (both default and user-defined) to subprocesses and database timestamps. The audit-trail is contained in the final web report.

Since database versions are dependent on the experimental design (e.g., databases that are continuously updated versus a fixed-date snapshot, a syndrome specific database versus a database such as NCBI nt, etc.), only their filenames and timestamps are logged. Downloading these databases is described in the documentation (https://github.com/DennisSchmitz/jovian), and versioning them is the responsibility of the end-user. Only the names of input files are logged since assessing proper demultiplexing or removal of custom designs depends on lab-specific operating protocols. Virus-typing results are generated by a public-private service without detailed versioning, so timestamps are logged instead.

Intuitive visual output

Tabular data are presented via interactive, user-friendly and sortable Qgrid spreadsheets[41]. For the taxonomic overview, interactive heatmaps are made with Bokeh[42]. Metrics of the different subprocesses are collated in a MultiQC report. Krona[43] plots provide a per-sample taxonomic overview. Scaffold alignment and annotations are visualized, and interactively analyzed, via IGVjs[44]. All visualizations are embedded into an interactive web-report based on Jupyter Notebooks[45], as shown in Fig. 2. An example notebook with mock data is available on MyBinder[46] and has been archived on Zenodo (https://doi.org/10.5281/zenodo.13371083).

Patient-privacy

Although a web browser is used as an interface, Jovian is a locally installed workflow, which means that the interface does not require an internet connection after the initial installation and downloading of databases, except for virus typing. Therefore, no patient data is sent over the internet. To enable further GDPR compliant dissemination, human data is removed by discarding any reads that match to human genome version GRCh38.

Availability, installation, usage instructions and documentation

Jovian is available via https://github.com/DennisSchmitz/jovian under the AGPLv3 Free Open Source license. For future continuity, installation and usage instructions are described on GitHub. Please refer to the instructions and documentations provided there. For accessibility, a default database configuration suitable for public health and clinical applications can be downloaded and installed using the '--install-databases' flag.

Technical details

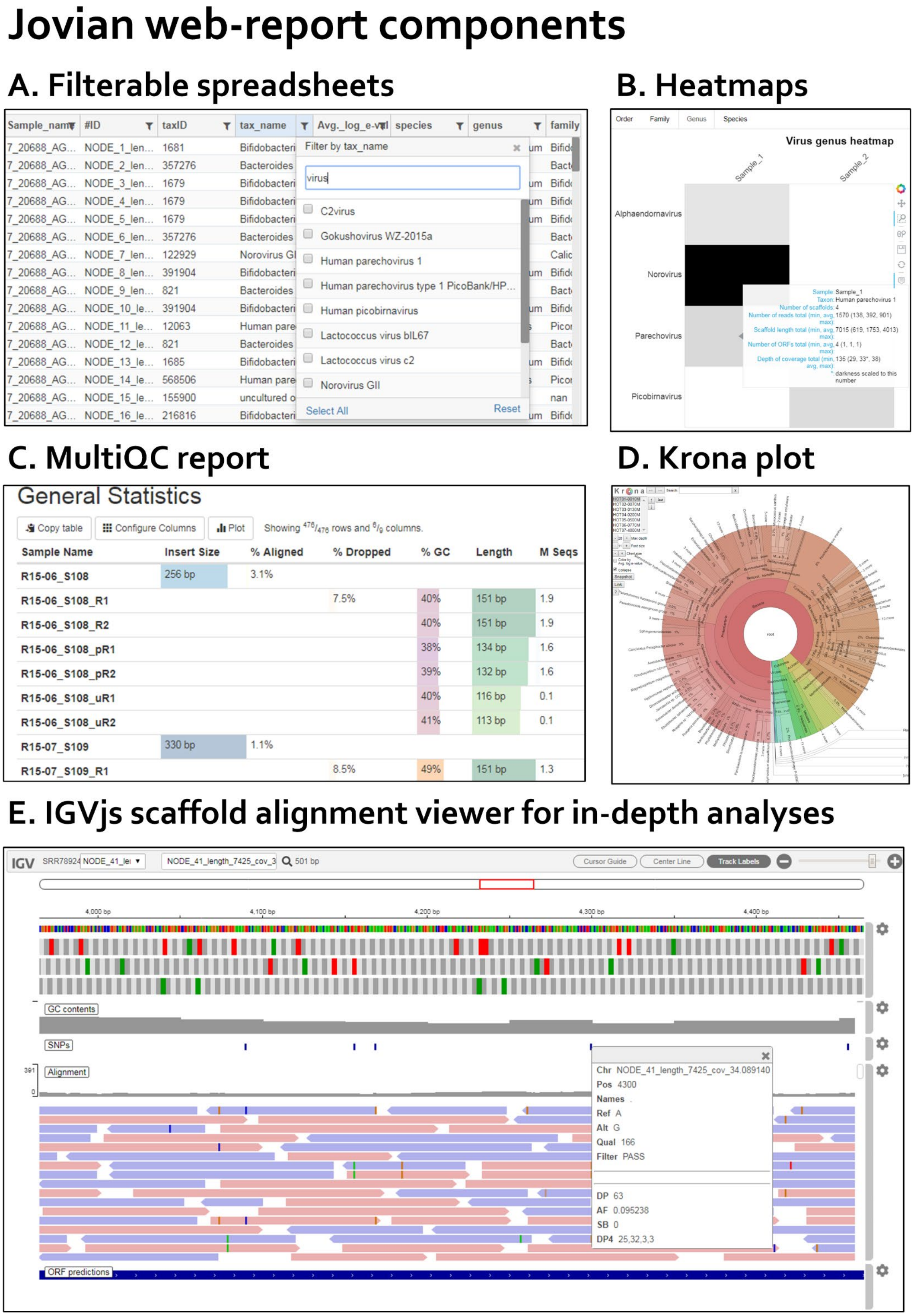
Jovian has been developed for high-end consumer-grade laptops/PCs, servers and grid-computers running the Red Hat Enterprise Linux (RHEL), CentOS or Ubuntu Linux distributions, with a recommended minimum of 16 cores and 24GB RAM. A single process will use at most 14 threads. It works on single machines by specifying "--local" and high-performance computing (HPC) grids (default value) using Load Sharing Facility (LSF) or Slurm[47,48].

Comparison algorithms for taxonomic classification

BLAST with the NCBI nt database was used for taxonomic annotation and compared to Kraken2 (version 2.1.3) and Centrifuge (version 1.0.4) using both the nt and the standard databases, as recommended in their respective user manuals. Scaffolds generated by Jovian, based on study PRJEB54724[49], were used as input for Kraken2 and Centrifuge. To ensure comparability, scaffolds assigned to the taxa listed in Table 2 were processed through their respective typing tools. All analyses were conducted on the same RHEL cluster using 12 threads. The same reporting threshold as the original manuscript was applied[49].

Assessment of portability and reproducibility

Jovian was independently installed and tested on four systems, hosted by three different institutes, utilizing grid-computers, a high-performance singular server and a Windows10 laptop using Windows Subsystems for Linux version 2 (WSL2) to determine its portability and reproducibility by analyzing samples R01-06, R02-13 and R04-15 from the publicly available study PRJEB54724[49]. These samples are from norovirus-positive human feces, subjected to Illumina shotgun metagenomic sequencing. The FASTQ files did not contain human reads, as these were removed before submission to the European Nucleotide Archive (ENA). Henceforth these systems will be referred to as "CentOS", "RHEL", "RockyLinux" and "Windows10-Ubuntu" with their details outlined below:



Jovian web-report components

A. Filterable spreadsheets

Sample_name	#ID	taxID	tax_name	Avg._log_e-val	species	genus	family
7_20688_AG...	NODE_1_len...	1681	Bifidobacteri			um	Bifid
7_20688_AG...	NODE_2_len...	357276	Bacteroides				Bact
7_20688_AG...	NODE_3_len...	1679	Bifidobacteri			um	Bifid
7_20688_AG...	NODE_4_len...	1679	Bifidobacteri			um	Bifid
7_20688_AG...	NODE_5_len...	1679	Bifidobacteri			um	Bifid
7_20688_AG...	NODE_6_len...	357276	Bacteroides				Bact
7_20688_AG...	NODE_7_len...	122929	Norovirus GI				Calic
7_20688_AG...	NODE_8_len...	391904	Bifidobacteri			um	Bifid
7_20688_AG...	NODE_9_len...	821	Bacteroides				Bact
7_20688_AG...	NODE_10_le...	391904	Bifidobacteri			um	Bifid
7_20688_AG...	NODE_11_le...	12063	Human pare			s	Picor
7_20688_AG...	NODE_12_le...	821	Bacteroides				Bact
7_20688_AG...	NODE_13_le...	1685	Bifidobacteri			um	Bifid
7_20688_AG...	NODE_14_le...	568506	Human pare			s	Picor
7_20688_AG...	NODE_15_le...	155900	uncultured o				nan
7_20688_AG...	NODE_16_le...	216816	Bifidobacteri			um	Bifid

Filter by tax_name

virus

- □ C2virus
- □ Gokushovirus WZ-2015a
- □ Human parechovirus 1
- □ Human parechovirus type 1 PicoBank/HP...
- □ Human picobirnavirus
- □ Lactococcus virus bIL67
- □ Lactococcus virus c2
- □ Norovirus GII

Select All Reset

B. Heatmaps

Sample: Sample_1
Taxon: Human parechovirus 1
Number of scaffolds: 4
Number of reads total (min, avg, max): 1570 (138, 392, 901)
Scaffold length total (min, avg, max): 7015 (619, 1753, 4013)
Number of ORFs total (min, avg, max): 4 (1, 1, 1)
Depth of coverage total (min, avg, max): 135 (29, 33*, 38)
*: darkness scaled to this number

C. MultiQC report

General Statistics

Copy table | Configure Columns | Plot | Showing 476/476 rows and 6/9 columns.

Sample Name	Insert Size	% Aligned	% Dropped	% GC	Length	M Seqs
R15-06_S108	256 bp	3.1%				
R15-06_S108_R1			7.5%	40%	151 bp	1.9
R15-06_S108_R2				40%	151 bp	1.9
R15-06_S108_pR1				38%	134 bp	1.6
R15-06_S108_pR2				39%	132 bp	1.6
R15-06_S108_uR1				40%	116 bp	0.1
R15-06_S108_uR2				41%	113 bp	0.1
R15-07_S109	330 bp	1.1%				
R15-07_S109_R1			8.5%	49%	151 bp	1.3

D. Krona plot

E. IGVjs scaffold alignment viewer for in-depth analyses

Chr NODE_41_length_7425_cov_34.089140
Pos 4300
Names .
Ref A
Alt G
Qual 166
Filter PASS

DP 63
AF 0.095238
SB 0
DP4 25,32,3,3

Fig. 2. Static overview of key components within the Jovian web report. An interactive demonstration is available on MyBinder[45,46]. (**A**) Interactive and filterable Qgrid spreadsheets visualize tabular flat files[41]. (**B**) Bokeh[42] heatmaps offer an interactive multi-sample perspective, highlighting different taxa (bacteria, phages, viruses) across distinct taxonomic levels (up to species). (**C**) The aggregation of quality control metrics from various subprocesses is presented through a MultiQC report[26]. (**D**) Per-sample taxonomic insights are conveyed via Krona (version 2.7.1; https://github.com/marbl/Krona) plots[43]. (**E**) Detailed alignment investigation is facilitated through IGVjs[44] which presents depth of coverage, average GC content, predicted open reading frames and identified minority variants.

Sample name	Virus name	Read-count Jovian	Read-count commercial workflow	Fraction (percentage)
R01-01	Norovirus GII.4[P4]	626.832	495.943	79.1
R01-02	Norovirus GII.7[P7]	498.706	419.530	84.1
R01-06	Norovirus GI.7[P7]	8.570	7.334	85.6
	Sapovirus GI.7	764.574	658.311	86.1
	Coxsackievirus A2	614	534	87.0
R02-06	Norovirus GII.17[P17]	27.574	23.460	85.1
R02-13	Norovirus GII.17[P17]	98.879	87.349	88.3
	Norovirus GII.4[P31]	1.931	1.689	87.5
R02-18	Norovirus GII.2[P16]	506	461	91.1
R04-14	Norovirus GII.4[P16]	3.022.977	2.526.414	83.6
	Coxsackievirus A4	163.425	137.926	84.4
	Bocaparvovirus 1	64.501	54.575	84.6
	Bocaparvovirus 3	3.754	3.243	86.4
R04-15	Norovirus GI.5[P4]	10.154	8.558	84.3
	Norovirus GI.6[P6]	1.045	903	86.4
	Aichivirus 1	6.236.103	5.343.266	85.7

Table 2. Comparison of jovian versus a commercial online reference workflow[6] for eight samples from study PRJEB54724[49] which focused on viruses with public health relevance. Both workflows identified the same strains as full genomes.

1. A server using the CentOS version 7 (Core) Linux distribution, hosted by the Erasmus University Medical Center, employed Jovian in "--local" mode, with kernel "3.10.0-1160.49.1.el7.x86_64".
2. A Load Sharing Facility (LSF) grid-computer using the Red Hat Enterprise Linux (RHEL) version 7.9 (Maipo) distribution, hosted at the Dutch National Institute for Public Health and the Environment (RIVM), employed Jovian with kernel "3.10.0-1160.88.1.el7.x86_64".
3. A Simple Linux Utility for Resource Management (Slurm) grid-computer using the Rocky Linux version 8.8 distribution, hosted by the Leiden University Medical Center, employed Jovian in "--slurm" compute mode with kernel "4.18.0-477.15.1.el8_8.x86_64".
4. A Windows10 laptop using WSL2 with Ubuntu version 20.04.6 LTS Linux distribution employed Jovian in "--local" compute mode with kernel "5.15.90.1-microsoft-standard-".

Message Digest 5[50] (MD5) hashes were generated to compare the results generated by these different systems. When these hashes were not identical, the differences were inspected using daff (version 1.3.46)[51] and diff (version 3.3) using the '-y –suppress-common-lines' flags.

Results

Validation of jovian: comparative analysis with a commercial reference workflow

To validate Jovian, we compared it to an online commercial reference workflow[6]. This workflow was selected due to its similar scope and frequent usage in public health. For this purpose, we used eight samples from study PRJEB54724[49]: R01-01, R01-02, R01-06, R02-06, R02-13, R02-18, R04-14 and R04-15. The original study used a viromics approach for norovirus surveillance, benchmarking it against conventional RT-qPCR and Sanger sequencing methodologies. The samples were metagenomically sequenced on an Illumina platform. Jovian version 1.0.0, with default settings and databases downloaded on May 22nd 2022, was compared against the online commercial reference workflow[6] on November 3rd 2023. In Table 2, both workflows were compared and identified the same genomes of viruses with public health relevance and the same full genomes.

Performance evaluation of taxonomic annotation algorithms

Parallel analysis of the benchmark dataset using Jovian took 103 min on an RHEL cluster, with the overall runtime determined by the most computationally demanding sample. To explore potential reductions in runtime, the time-consuming taxonomic annotation step using BLAST, was compared to Kraken2 and Centrifuge. Scaffolds generated by Jovian from the benchmark dataset were used as input for Kraken2 and Centrifuge (Fig. 3).

We applied the same reporting threshold as the original manuscript to evaluate the accuracy of virus detection in the benchmark dataset. The F1-score, which balances precision and recall by accounting for both false positives and false negatives, was used as the primary performance metric. BLAST correctly identified all scaffolds (100% F1-score). Kraken2 with the standard database misclassified norovirus and sapovirus scaffolds of R01-06 as 'unclassified', resulting in an F1-score of 93.3%, whereas Kraken2 with the nt database correctly identified all expected viruses. Centrifuge, using both the standard and nt database identified all expected viruses but reported false positives: a 2.8 kilobase scaffold was incorrectly labeled as Herpesvirus, and an 8.3 kilobase scaffold as human immunodeficiency virus (HIV), leading to an F1-score of 97.0% for both databases.

Runtime, maximum memory usage and storage requirements were also assessed (Fig. 3). BLAST with the nt database required approximately 37 min per sample, using up to 9 GB of RAM and 490 GB of storage. Centrifuge with the nt databases had similar runtimes but required more memory. Kraken2 with the nt database

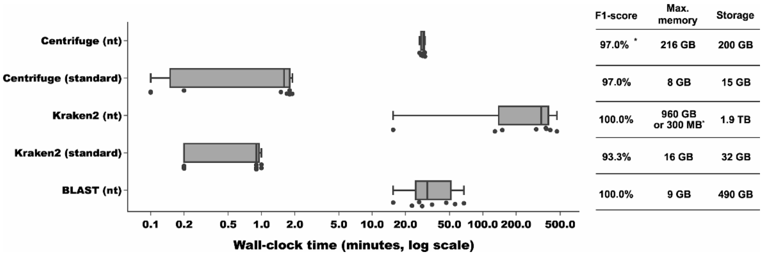


Fig. 3. Comparison of taxonomic annotation algorithms based on F1-score, runtime, maximum memory usage and required storage space. BLAST with the nt database was compared to Centrifuge and Kraken2 using the nt and their standard databases. *For Kraken2 with the nt database, the analysis required ~ 960GB RAM. Due to insufficient available memory, the "--memory-map" flag was used, leading to a bottleneck and extended runtimes.

Filename	Content description	Identical MD5 checksums (last six digits)
all_taxClassified.tsv	Scaffold metadata, incl. taxonomy	Yes (e30ea7)
all_taxUnclassified.tsv	Scaffold metadata, excl. taxonomy	Yes (b0c79e)
all_virusHost.tsv	Scaffold host and disease metadata	Yes (7f0568)
all_filtered_SNPs.tsv	Minority variants	Yes (73ba51)

Table 3. Reproducibility of jovian analysis across four different machines, hosted by independent institutes, when analyzing an identical dataset (PRJEB54724[49]) with identical databases. Jovian's four main output files are listed with a short description of its contents and whether the result was identical based on MD5 checksums of output files generated by all respective institutes.

required 960 GB RAM, since this was not available, the '--memory-map' flag was used which increased runtime significantly but lowered maximum memory usage to 300 MB. However, when using their respective standard databases, both algorithms achieved shorter runtimes and lower memory and storage usage compared to BLAST with the nt database.

Overall, while the nt database provided the most reliable results across all taxonomic algorithms, it also resulted in longer runtimes regardless of the algorithm used.

Reproducibility of jovian across diverse computing environments and institutes

To assess the reproducibility of Jovian, we installed the workflow on four distinct machines—CentOS, RHEL, RockyLinux and Windows10-Ubuntu—at three different institutes. All installations used identical databases downloaded on May 22nd, 2022, and analyzed the same dataset within an hour (study PRJEB54724, samples R01-06, R02-13 and R04-15[49]), which is a subset of the dataset used in the previous comparisons. Importantly, on the CentOS and Windows10-Ubuntu machines, Jovian was employed in "--local" mode, operating on a single computer, while the analyses on the RHEL and RockyLinux machines were performed on LSF and SLURM grids, respectively. MD5 hashes were used to compare the output files generated by different institutes. Identical hashes would confirm that files are identical, thereby demonstrating that Jovian was successfully deployed and that its results were reproducibly generated.

Initially, discrepancies in MD5 hashes were observed for the output files generated by the different institutes, despite identical input files and database versions. Analysis using daff[51] revealed many small discrepancies between the output files, primarily stemming from minor variations in SNP-calling metrics and differences in scaffold names and the sorting of scaffolds. These variations were caused by small rounding differences in the scaffold names, which contains the scaffolds' average depth-of-coverage (Supplementary Table S2). We identified variations in the number of computational threads used by SPAdes and LoFreq as the cause of these differences.

To achieve more efficient parallelization for a shorter turn-around time, Jovian had initially been optimized to the number of threads the host-computer could provide. Due to the effect it had on reproducibility, we adapted this and fixed the number of computational threads used per algorithm. After fixing the number of threads, analyses were repeated at three different institutes with different IT architectures. The resulting MD5sum hashes, shown in Table 3, were all identical. Furthermore, in all replicates the expected viral pathogens, as listed in Table 2, were accurately identified. This reproducibility is a requirement for diagnostic usage.

Discussion

To validate Jovian, its results were first compared with an online reference workflow, showing that they are comparable in relation to the identification of clinically or public health relevant viruses (Table 2). Jovian was able to identify slightly more reads belonging to these viruses compared to the commercial reference workflow, but this did not impact the interpretation of the data. We hypothesized this was caused by different QC cutoffs. As outlined by De Vries et al.[52] and ISO15189, validation by comparison to conventional methodologies and ring trials is required. Both in-silico and real-sample ring trails show that Jovian's results are reliable and sensitive[53,54] and in a public health surveillance setting, metagenomic sequencing on a Illumina-platform combined with Jovian analysis shows an improvement over Sanger sequencing-based surveillance of noroviruses[49].

We also assessed the performance of BLAST, Kraken2 and Centrifuge on the benchmark dataset with the intent of reducing its runtime. BLAST and Kraken2 (nt) resulted in the highest F1-score (100%). However, Kraken2 (nt) required 960 GB of RAM or 4.5 h runtime per sample. Kraken2 with the standard database resulted in two false negatives, while Centrifuge, using both nt and standard databases, reported one false positive. This led us to select BLAST (nt) as the taxonomic annotation algorithm for Jovian. This decision also aligned with stakeholders' preferences for backward-compatibility and preparedness for emerging viruses by using the most complete nucleotide database.

When we assessed Jovian's reproducibility, we found minimal changes in contig names and SNP-calling metrics such as depth of coverage, quality and allele frequency (Supplementary Table S2) based on the number of logged threads used by the algorithms incorporated in Jovian. While the differences were minor and did not affect the interpretation of this dataset, there is the possibility that this could confound independent replication. In theory, in some instances this variation could result in a SNP not being reported if it falls below the 5% allele frequency cutoff. We removed the inter-replication variability by fixing the thread-counts within Jovian, thereby making hash-based comparison feasible. This allows rapid validation of workflow deployment to different IT architectures. To the best of our knowledge, we have not seen the influence of thread-counts on metagenomic analysis reported in clinical or public health literature.

By aligning Jovian with privacy legislation and laboratory certification protocols we lowered the barrier for integration into clinical and public health surveillance programs. By automatically removing human data, it adheres to the GDPR. All subprocess parameters, user-defined parameters and database information are logged into an audit trail, thus aligning with ISO15189. Furthermore, all software is publicly available and transparent. Third, Jovian is demonstrably portable and reproducible, a requisite for clinical implementation, as shown by generating identical results on four different hardware platforms across three institutes (Table 3). In the comparable field of bacterial WGS of food products, this is a requirement for ISO23418 certification, emphasizing its importance. This offers the possibility to expand across a network of laboratories. To improve this interoperability, and in light of FAIR guidelines[18], all output files, their formats and brief content descriptions are provided in Supplementary Table S1 and the GitHub repository (https://github.com/DennisSchmitz/jovian).

Integrating any technique or tool within a public health or clinical setting requires integration into a quality assurance regimen, which should be provided by the user. This includes steps such as demultiplexing of raw data to remove barcodes or custom designs, establishing a versioning and validation strategy for the employed databases, and implementing a data management ecosystem to store results, metadata and methodological information[5,55]. While Jovian's '--install-databases' flag offers a convenient default by downloading 520 GB of databases in about four hours, this may not always be the optimal choice. Database selection is a critical aspect of experimental design. For example, when working with non-human datasets, GDPR compliance may not be necessary, and aligning against a species-specific genome rather than the default human genome could streamline analysis and reduce runtime. Considering the storage requirements (~ 550 GB) and processing power required, Jovian is best suited for high-performance computing (HPC) clusters, although smaller datasets can be processed on high-end consumer hardware using the '--local' flag.

Contamination in viromics is a known problem[56,57]. Since Jovian can be used to analyze multiple sequencing runs either simultaneously or individually, we recommend users apply filters in the web report based on their negative controls. As a general guideline, scaffolds with an even distribution, an average depth-of-coverage of 3x, and a length of ≥ 250nt would be sufficient for reporting for samples with approximately one million reads. These thresholds should be adjusted according to the total read count per sample, the number of PCR cycles during library preparation, and lab-specific contamination levels. We also recommend that low-coverage or high-impact findings be independently validated using classical molecular methods[52].

Currently, virus (geno)typing relies on a webtool hosted by the RIVM. We send queries to it and receive (geno)typing results, which are integrated into the web report. This process, therefore, relies on an internet connection. Since these webtools are the results of a public-private partnership, they are not open source and cannot be incorporated into Jovian for offline use at this time. This is why only timestamps of the typing results are logged and more details are not included. This reliance affects Jovian's accessibility, reproducibility and long-term sustainability which is why the development of a free and open-source genotyper is ongoing.

Another impediment to its integration into clinical and public health surveillance is its accessibility. Designed for accessible application, Jovian operates through a command-line interface, rendering results accessible via standard web browsers. The interactive web report has been tailored to a set of use-cases, encompassing the identification of specific (pathogen) taxa, virus-typing for outbreak investigations, and the exploration of variants (quasispecies), either present as consensus level SNPs or minority SNPs (Fig. 1). The interactive genome viewer empowers end users to discern potential biases or artifacts inherent in the data. As is characteristic of any automated workflow, manual curation of pertinent information by experts remains essential. In this regard, Jovian provides visualizations that assist in this process (Fig. 2). While the initial setup of Jovian necessitates the involvement of a system administrator or bioinformatician, subsequent utilization of Jovian empowers lab-technicians and (clinical) virologists to engage in independent analyses.

Building upon insights by Nieroda et al.[55], the integration of demultiplexing, Jovian data analysis, and the preservation of an audit trail in an comprehensive quality management system can be accomplished through iRODS[58]. Through this integration, Jovian has supported the surveillance and diagnostic initiatives at the RIVM virology department by successfully analyzing over a thousand surveillance samples, encompassing viruses such as mumps[59], enterovirus E30[60], and norovirus[49]. Its robustness has been underscored through ring trials[53,54] and its implementation in the COMPARE datahubs by the European Bioinformatics Institute (EBI)[61]. Furthermore, Jovian's core principles served as the foundation for the RIVM bacterial workflow, Juno[62], highlighting the essential role of open-source science in advancing cross-domain scientific progress.

We conclude that the development of this workflow is an important step toward the deployment of metagenomic approaches within public health institutes and clinical settings. Its results were validated by comparison to a commercial reference workflow, and it adheres to privacy legislation and laboratory certification protocols. This addresses several impediments to the deployment of metagenomics in these settings and has the potential to enhance the breadth of surveillance and testing programs, thereby fostering more effective public health interventions.

Data availability

The authors declare that no new sequence data was generated for this study. All sequence datasets analyzed in this study are available within this article: Schmitz et al.[49] under study accession number: PRJEB54724. The raw sequence files of human samples for PRJEB54724 were submitted after the removal of human reads.

Received: 25 March 2024; Accepted: 20 September 2024
Published online: 29 October 2024

References

1. Munnink, B. B. O. et al. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nat. Med.* **26**, 1405–1411 (2020).
2. Houldcroft, C. J., Beale, M. A. & Breuer, J. Clinical and biological insights from viral genome sequencing. *Nat. Rev. Microbiol.* **15**, 183–192 (2017).
3. Grubaugh, N. D. et al. Tracking virus outbreaks in the twenty-first century. *Nat. Microbiol.* **4**, 10–19 (2019).
4. Francis, R. V. et al. The impact of real-time whole-genome sequencing in controlling healthcare-associated SARS-CoV-2 outbreaks. *J. Infect. Dis.* **225**, 10–18 (2022).
5. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nat. Rev. Genet.* **20**, 341–355 (2019).
6. Vilsker, M. et al. Genome detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics* **35**, 871–873 (2019).
7. Minot, S. S., Krumm, N. & Greenfield, N. B. One codex: a sensitive and accurate data platform for genomic microbial identification. *BioRxiv* 027607 (2015).
8. Flygare, S. et al. Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol.* **17**, 1–18 (2016).
9. Taxonomer. *Taxonomer Page*. https://taxonomer.iobio.io/ (Accessed 02 November 2023).
10. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
11. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 1–13 (2019).
12. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
13. Cadenas-Castrejón, E., Verleyen, J., Boukadida, C., Díaz-González, L. & Taboada, B. Evaluation of tools for taxonomic classification of viruses. *Brief. Funct. Genom.* **22**, 31–41 (2023).
14. Kroneman, A. et al. Proposal for a unified norovirus nomenclature and genotyping. *Arch. Virol.* **158**, 2059–2068 (2013).
15. Kroneman, A. et al. An automated genotyping tool for enteroviruses and noroviruses. *J. Clin. Virol.* **51**, 121–125 (2011).
16. Kroneman, A., de Sousa, R., Verhoef, L., Koopmans, M. P. & Vennema, H. Usability of the international HAVNet hepatitis a virus database for geographical annotation, backtracing and outbreak detection. *Eurosurveillance* **23**, 1700802 (2018).
17. Mulder, A. C. et al. HEVnet: a one health, collaborative, interdisciplinary network and sequence data repository for enhanced hepatitis E virus molecular typing, characterisation and epidemiological investigations. *Eurosurveillance* **24**, 1800407 (2019).
18. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).
19. Jansen, S. A. et al. Broad virus detection and variant discovery in fecal samples of hematopoietic transplant recipients using targeted sequence capture metagenomics. *Front. Microbiol.* **11**, 560179 (2020).
20. Carbo, E. C. et al. Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics. *J. Clin. Virol.* **130**, 104566 (2020).
21. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
22. Grüning, B. et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **15**, 475–476 (2018).
23. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: scientific containers for mobility of compute. *PLoS ONE* **12**, e0177459 (2017).
24. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
25. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data* (2010).
26. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
27. Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
28. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
29. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
30. Broad Institute. *Broad Institute GitHub Page for Picard*. https://broadinstitute.github.io/picard/ (Accessed 06 October 2023).
31. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
32. Wheeler, D. L. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **36**, D13–D21 (2007).
33. Rubino, F. & Creevey, C. MGkit Metagenomic framework for the study of microbial communities. *Figshare Poste* (2014).
34. NCBI. *NCBI FTP for new_taxdump*. https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/ (Accessed 06 October 2023).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. Bushnell, B. *BBTools Software Package* (2014).
37. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

38. Wilm, A. et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).
39. Mihara, T. et al. Linking virus genomes with host taxonomy. *Viruses* **8**, 66 (2016).
40. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 1–11 (2010).
41. Quantopian. *Quantopian GitHub Page*. https://github.com/quantopian/qgrid (Accessed 06 October 2023).
42. Bokeh. *Bokeh Homepage*. https://bokeh.pydata.org/en/latest/ (Accessed 06 October 2023).
43. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a web browser. *BMC Bioinform.* **12**, 1–10 (2011).
44. Robinson, J. T., Thorvaldsdóttir, H., Turner, D. & Mesirov, J. P. igv.js: an embeddable JavaScript implementation of the integrative genomics viewer (IGV). *Bioinformatics* **39**, 830 (2023).
45. Kluyver, T. et al. *Positioning and Power in Academic Publishing: Players, Agents and Agendas* 87–90 (IOS, 2016).
46. Ragan-Kelley, B. et al. *Proceedings of the 17th Python in Science Conference* 113–120 (eds. Akici, F.).
47. IBM. *IBM Spectrum LSF Suites Homepage*. https://www.ibm.com/products/hpc-workload-management (Accessed 06 October 2023).
48. Sched, M. D. *Slurm Homepage*. https://www.schedmd.com/ (Accessed 06 October 2023).
49. Schmitz, D. et al. Metagenomic surveillance of viral gastroenteritis in a public health setting. *Microbiol. Spectr.* **11**, e05022 (2023).
50. Rivest, R. (Editor RFC, 1992).
51. Fitzpatrick, P. *daff GitHub Page*. https://github.com/paulfitz/daff (Accessed 06 October 2023).
52. de Vries, J. J. et al. Recommendations for the introduction of metagenomic next-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting. *J. Clin. Virol.* **138**, 104812 (2021).
53. Brinkmann, A. et al. Proficiency testing of virus diagnostics based on bioinformatics analysis of simulated in silico high-throughput sequencing data sets. *J. Clin. Microbiol.* **57**, 419. https://doi.org/10.1128/jcm.00466-00419 (2019).
54. de Vries, J. J. et al. Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. *J. Clin. Virol.* **141**, 104908 (2021).
55. Nieroda, L. et al. iRODS metadata management for a cancer genome analysis workflow. *BMC Bioinform.* **20**, 1–8 (2019).
56. Salter, S. J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 1–12 (2014).
57. Kulakov, L. A., McAlister, M. B., Ogden, K. L., Larkin, M. J. & O'Hanlon, J. F. Analysis of bacteria contaminating ultrapure water in industrial systems. *Appl. Environ. Microbiol.* **68**, 1548–1555 (2002).
58. iRODS. *iRODS Homepage*. https://irods.org/ (Accessed 06 October 2023).
59. Bodewes, R. et al. Molecular epidemiology of mumps viruses in the Netherlands, 2017–2019. *PLoS ONE* **15**, e0233143 (2020).
60. Benschop, K. S. et al. Molecular epidemiology and evolutionary trajectory of emerging Echovirus 30, Europe. *Emerg. Infect. Dis.* **27**, 1616 (2021).
61. Amid, C. et al. The COMPARE data hubs. *Database* **2019**, 136 (2019).
62. RIVM Bioinformatics Team. *Juno GitHub Page*. https://github.com/RIVM-bioinformatics/juno-assembly (Accessed 06 October 2023).

Acknowledgements

We would like to thank Divyae Prasad, Bas van der Veer, Anne-Marie van den Brandt, Rogier Bodewes, Kim Benschop, Erwin van Wieringen, Mark Kroon, Sander van Boheemen and Richard Molenkamp for their feedback, and Ed Kuijper for institutional support of the contribution of Sam Nooij.

Author contributions

Conceptualization, D.S., H.V., A.K. and M.K.; methodology, D.S., F.Z., S.N. and T.J.; software, D.S., F.Z., S.N., T.J., J.C., R.V., and J.L.; validation, D.S., F.Z., H.V., J.C. and M.G.; formal analysis, D.S., and F.Z; investigation, D.S. and J.C.; resources, H.V., A.K., R.V., J.L. and M.K.; data curation, all authors; writing—original draft preparation, D.S.; writing—review and editing, all authors; visualization, D.S.; supervision, A.K., M.G., J.L. and M.K.; project administration, D.S., A.K. and M.K.; funding acquisition, M.K. All authors have read and agreed to the published version of the manuscript.

Funding

This study was financed under European Union's Horizon H2020 grants COMPARE and VEO (Grant No. 643476 and 874735) and the NWO Stevin prize (Koopmans).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-73785-y.

Correspondence and requests for materials should be addressed to D.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.