# Exploring the landscape of the genomic wastewater surveillance ecosystem: a roadmap towards standardization

Fotis Psomopoulos, Konstantinos Kyritsis, Ivan Topolsky, Bérénice Batut, Amy Heather Fitzpatrick, Gabriele Leoni

BioHackrXiv

# Exploring the landscape of the genomic wastewater surveillance ecosystem: a roadmap towards standardization

**Fotis Psomopoulos[1], Konstantinos Kyritsis[1], Ivan Topolsky[2], Bérénice Batut[3], Amy Heather Fitzpatrick[4], and Gabriele Leoni[5]**

**1** Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece **2** Computational Biology Group, SIB Swiss Institute of Bioinformatics, Basel,Switzerland **3** University of Freiburg, Germany **4** University College Dublin **5** European Commission, Joint Research Centre (JRC), Ispra, Italy

**BioHackathon series:**
BioHackathon Europe 2022
Paris, France, 2022
*Wastewater Surveillance*

**Submitted:** 01 Nov 2023



Published by BioHackrXiv.org

## Introduction

Nearly two years after the initial report of SARS-CoV-2 in Wuhan, China, the COVID-19 pandemic has affected over 485 million individuals. Wastewater surveillance has garnered substantial attention as a passive monitoring system to complement clinical genomic surveillance activities during the SARS-CoV-2 pandemic. Several effective methods are now in place for detecting and quantifying viral RNA in wastewater samples, and it is evident that RNA concentrations in wastewater correlate with reported case trends.

Exploratory projects have demonstrated the potential of Wastewater Based Epidemiology (WBE), nevertheless it is imperative to coordinate efforts, establish standards, and create a catalog of available software tools and services. This coordination will streamline the deployment of end-to-end genomic wastewater surveillance pipelines and promote the adoption of these monitoring methods within the broader scientific community. The initial step involves identifying and cataloging the challenges of working with wastewater HTS and the pertinent methodologies and bioinformatics workflows essential for managing genomic data from wastewater samples, thus forming a coherent structure.

The primary objective of this project was to systematically review, compile, and initiate the integration, standardization, and documentation of diverse approaches for genomic wastewater surveillance. Drawing on the expertise of the ELIXIR Wastewater Surveillance Working Group, our focus was to create a comprehensive framework of components, including modules and tools, to facilitate the practical implementation of end-to-end genomic wastewater surveillance pipelines.

## Defining the key elements of an omic Wastewater Surveillance framework

One of the project's initial tasks involved a comprehensive review to identify the pertinent questions that omic Wastewater Surveillance should aim to answer. A fundamental criterion for this system is its ability to generate actionable results, which are insights that can be readily comprehended and utilized by public health experts. To illustrate, `VCF` files, with their specialized content, cannot be categorized as actionable since they necessitate specific expertise for interpretation. In contrast, variant frequencies within a given sample represent actionable data, as they provide public health experts with readily understandable information for further actions and decisions.

An indicative list of these questions are:

1. Is virus v present in the wastewater sample?

Essentially this is a question of presence/absence that is traditionally addressed through RT-qPCR or equivalent methods.

2. Does the virus 'v' vary seasonally?

For example, many gastrointestinal viruses exhibit a strong seasonal peak during the winter in the northern hemisphere. Conversely, enteroviruses exhibit a strong peak during the summer period. This information can be used to inform public health authorities when specific intervention measures should be implemented.

3. Is the variant vv present in the wastewater sample?

4. What is the longitudinal change of the variants?

This question seeks to determine how the frequency of variants changes across different time points and samples.

5. What is the fitness advantage?

This is one of the more advanced questions that could be addressed by wastewater surveillance, and involves the projection of the trajectory of a given variant in time and the calculation of a possible $R_e$.

6. What are potentially emerging variants?

# Standardizing definitions

## Variants

Looking into viruses in particular, one of the main challenges is still the definition of variants. Typically, for each different viral family, a different approach is taken with directions coming from consortium groups such as the International Committee on Taxonomy of Viruses (ICTV) or sub-specialist groups at genus level. Generally, variants have one to multiple different Single Nucleotide Variations (SNVs) and short Insertion-Deletion mutations (InDels) compared to the closest known strain within a viral species. Nevertheless, variant indicates that there are no known phenotypic differences between the closest known strain and the variant. The challenge is in defining how many SNVs and InDels result in variant designation, and this has lead to varying definitions and applications of the term. Phenotypic differences should theoretically result in the designation of new strain.

Public databases (i.e. public health registries and specialized repositories such as outbreak.info or CovSpectrum) can be used to monitor variants for some viruses such as SARS-CoV-2 or Mpox. Nevertheless, given the multitude of potential variants and respective definitions, more often than not, tools use a custom list of pre-selected definitions (or even simplified versions of the same definitions). This approach implies that there is always a human factor in the process, assessing which definitions (or parts of definitions) will be used.

In either case, it is important to identify what would be the minimal information expected as input to a federated omics wastewater surveillance system, i.e. reaching a consensus on how to tackle common definitions, for example FAIR definitions.

## Antimicrobial Resistance

Antimicrobial resistance (AMR) is the phenomenon where microbes, either through mutations or the transfer of genes, acquire the capability to withstand the effects of antimicrobial agents that were previously effective in treating them. In simpler terms, AMR is the process by which microbes become resistant to drugs that used to work against them.

Antimicrobial resistance (AMR) in bacteria arises through several mechanisms. Bacteria can naturally develop resistance by accumulating genetic mutations, particularly in the genes encoding the target proteins of antimicrobial agents. These mutations can render the drugs less effective. Additionally, bacteria can share resistance genes through horizontal gene transfer mechanisms such as conjugation, transformation, and transduction, allowing the rapid spread of resistance within bacterial populations. Plasmids, small circular DNA molecules, often carry these resistance genes and can be readily transferred between bacteria. Some bacteria possess efflux pumps that actively expel antimicrobial agents from within the bacterial cell, reducing drug concentrations and efficacy. Bacterial biofilm formation provides a protective matrix that shields bacteria from antimicrobial agents. Enzymatic inactivation, wherein bacteria produce enzymes that degrade antibiotics, is another strategy they employ. Therefore, the challenge within a metagenomics framework is recognizing the potential AMR mechanisms that a genome or assembly could employ to evade antimicrobials, without the complete environmental context.

Furthermore, presence of specific mutations or plasmids does not infer AMR as it is not always clear from shotgun metagenomics whether those specific regions/genes can be transcribed. Metatranscriptomics can provide a clearer resolution on potential AMR, but the gold standard remains phenotypic assays in the wet-lab.

### Viral fitness

Viral fitness is a concept in virology that refers to the ability of a virus to successfully replicate and propagate within a host organism or population of hosts. It encompasses various aspects of a virus's biological characteristics, such as its reproductive rate, transmission potential, and adaptability to changing environments. A virus with high fitness can efficiently reproduce, spread from host to host, and adapt to selective pressures, often resulting in a more successful and prevalent infection.

Assessing viral fitness through shotgun metagenomics and metatranscriptomics provides valuable insights into the dynamics of viral populations within complex microbial communities. By tracking changes in viral genomic content, such as the accumulation of mutations or the presence of fitness-enhancing mutations, viral fitness can be indirectly inferred. Moreover, metatranscriptomics is particularly well-suited for studying RNA viruses as RNA is the primary target rather than DNA. Once again, the challenge is the definition of what is fitness-enhancing mutations. In order to interpret genomic data for viral-fitness, mutations need to be studied within a wet-lab context and documented within literature. During specific scenarios, such as pandemics or epidemics these wet-lab studies are challenging to conduct within short time-frames and results often lag behind the evolving evolutionary landscape of the virus. Inference through phylogenomics has improved our understanding of which mutations may play a role in viral-fitness but this is a field lacking curated databases for broad annotation of viral genomes.

## Considerations on Wastewater data analyses

Wastewater data analysis presents unique challenges due to the partial genomic data often found in samples. In many cases, the complete genome sequence of the target organism may not be available, making reliable variant identification challenging. The detection of specific mutations with potential implications for the organism's viral fitness can serve as a valuable signal for public health measures. For instance, it can aid in the identification of significant Spike variants in viruses like SARS-CoV-2. Nevertheless, care should be taken to implement quality control parameters that limit the missed or mis designation of variants i.e. X coverage, quality score cut-offs, contig length, N repeated detection. Specialised tools for variant designation and lineage assignment should be used for highly fragmented data from wastewater, water or food samples.

# Omic Wastewater Surveillance

We identify two stages of omic Wastewater Surveillance workflows; with stage two consisting of two approaches:

- Design and in-silico validation of primers and/or benchmarking bioinformatic pipelines
- Real world application
- specific target (e.g. single virus and variants), typically amplicon based, and
- unknown target metagenomics / metatranscriptomics - Shotgun metagenomics or metratranscriptomics can include virus enrichment or host depletion steps prior to High Throughput Sequencing

For each of these stages and approaches, we have identified the key phases involved (not including the sampling and wet-lab partis of the process).

## In-silico validation

Slightly different approaches are required for in-silico validation for targeted and non-targeted approaches. These are clearly outlined in two separate tables, whilst the final table in this section contains the relevant steps that are common to both approaches.

### Specific target (tiling amplicon-based)

| Nb | Step | Objective | Actionable |
|---|---|---|---|
| 0 | Generation of comprehensive viral database | Complete representation of genetic diversity for target virus(es) | No |
| 1 | *In-silico* PCR with primers | FASTA file of amplicons | No |
| 2 | Classification accuracy | Comparison of performance of primers for specific target(s) | Yes |

### Specific target (shotgun metagenomics)

| Nb | Step | Objective | Actionable |
|---|---|---|---|
| 0 | Generation of comprehensive viral database | Complete representation of genetic diversity for target virus(es) | No |
| 1 | Generation of background DNA/RNA wastewater database | Accurate representation of matrix composition ( i.e. influent or effluent) | No |
| 2 | Model wastewater sample composition | Accurately represent the viral composition in a wastewater dataset prior to sequencing, incorporating relevant information such as enrichment/depletion steps, composite or passive sample type | No |

**Final approach for both targeted and non-targeted *in-silico* approaches**

| Nb | Step | Objective | Actionable |
|---|---|---|---|
| 3 | *In-silico* simulation of platform specific sequencing reads | FASTQ/FAST5 files | no |
| 4 | Application of bioinformatic work-flows/pipelines | Various outputs - virus classification variants detected and composition and presence/absense of clinically relevant mutations | No |
| 5 | Comparison to ground truth - Classification accuracy and compositional similar-ity/dissimilarty | Yes | |

## Omics WW Surveillance System - specific target (e.g. single virus + variants) / tiling amplicon based

| Nb | Step | Objective | Actionable |
|---|---|---|---|
| 0 | RT-PCR | Rough presence/absence of virus. Plus variants depending on primer. | Yes (basic) |
| 1 | Sequencing | Produce raw reads data in `FASTQ/FAST5` format | No |
| 2 | Quality control step | trimming/reads quality assesment | No |
| 3 | Alignment to reference (incl. host removal and cleaning) | `BAM` file | No |
| 4 | Identify mutations | `VCF` files, list of mutations, pileups, etc | No |
| 5 | Detect variants based on definitions | List of variants | Yes (variant level) |
| 6 | Quantify variants based on definitions | List of variants with frequencies (should add up to 100% per sample) | Yes (variant trends) |
| 7 | Define new variants based on mutations | List of mutations not mapping to known definitions | Yes (emerging variants) |

| Nb | Step | Objective | Actionable |
|---|---|---|---|
| 8 | Clinically relevant mutation | Interpret mutations from VCF | Tracking of clinically relevent mutations, independent of variant designation - e.g. antibody evasion, drug resistance |

## Unknown target meta-genomics / -transcriptomics workflow

Omics WW Surveillance System - unknown target metagenomics/metatranscriptomics (depending on the type of target you are looking for)

| Nb | Step | Objective | Actionable |
|---|---|---|---|
| 0 | RT-qPCR | **Not applicable** | No |
| 1 | Sequencing | Produce raw reads data in `FASTQ/FAST5` | No |
| 2 | QC on reads | trimming/reads quality assesment | No |
| 3a | Alignment to references (incl. Host removal and cleaning; alignment vs known targets, e.g. `FASTA` files of the top 10 viruses /bacteria you are checking against) | `BAM` files and simple visualizations. | Yes (basic). As a next step, this can connect to the amplicon pipeline (for each `BAM` file), starting from *Step 3*. |
| 3b | Alignment/matching against agnostic databases (such as Kraken, NR, k-mer based search etc) | Table of matches (`BLAST`-like tables) and simple visualizations (such as by taxonomy). | Yes (basic) |
| 3c | AMR detection and virulence factor, using specific AMR databases. | List of AMR and virulence factor genes detected | Yes |
| 3d | Assembly based approaches (SPades) and DIAMOND/BLAST DB search for taxonomy classification | Assembled virus with taxonomic classification | yes |

| Nb | Step | Objective | Actionable |
|---|---|---|---|
| 4 | Specific target (amplicon-based) workflow (from point 4) | If Specific target can be identified by 3a/3b/3d, Specific target (amplicon-based) workflow can be used from point 4 | yes |

## List of relevant software tools

In order to have a better assessment of the landscape, it's important to be aware of the various bioinformatic tools that exist, and fit in the above steps. We have generated two tables to outline the tools useful for *in-silico* validation, as well as those tools used for the bioinformatic processing and analysis of High Throughput Sequencing data from wastewater samples.

We identified 12 tools for *in-silico* validation

| Tool name | Citation |
|---|---|
| InsilicoSeq | (Gourlé, 2023) |
| ART | (Huang et al., 2012) |
| BEAR | (sej917, 2023) |
| Grinder | (Xue, 2023) |
| NanoSim2 | (*NanoSim*, 2023) |
| NEAT | (*The NEAT Project V4.0*, 2023) |
| BadRead | (Wick, 2019) |
| PBSIM | (Faucon, 2023) |
| longsiland | (*Longislnd*, 2020) |
| *In-silico*-pcr | (Ozer, 2023) |
| Silica | (*Silica*, 2023) |
| SWAMPy | (*SWAMPy*, 2023) |

We identified 36 tools/workflow for bioinformatic processing and analysis of HTS data from wastewater

Here's the table reordered by action category:

| Tool/Workflow name | Action | Citation |
|---|---|---|
| Hostile | Host removal | (Constantinides, 2023) |
| viralrecon | Workflow for assembly and variant detection | (*Nf-Core/Viralrecon*, 2023) |
| Viral-ngs | Workflow for assembly and taxonomic classification | (*Viral-Ngs*, 2023) |
| VIRify | Detection, annotation, and taxonomic classification | (*VIRify*, 2023) |
| Galaxy wastewater amplicon workflow | | |

| Tool/Workflow name | Action | Citation |
|---|---|---|
| Wastewater_surveillance_pipeline | Variant composition & lineage assignment | (N'Guessan et al., 2022) |
| coronaSPades | Virus assembler | (Meleshko et al., 2021) |
| Kraken | taxonomic classification | (Wood & Salzberg, 2014) |
| Kraken2 | taxonomic classification | (Wood et al., 2019) |
| VirSorter2 | taxonomic classification | (jiarong, 2023) |
| KRONA | taxonomic visulisation | (*Marbl/Krona*, 2023) |
| Abricate | AMR or virulence gene prediction | (Seemann, 2023) |
| staramr | AMR prediction | (*Staramr*, 2023) |
| Mykrobe | AMR prediction & phylogenetics | (*Mykrobe*, 2023) |
| Pathofact | AMR or virulence gene prediction | (Nies et al., 2021) |
| ariba | AMR prediction and MLST typing | (*ARIBA*, 2023) |
| ViralClust | Virus specific clustering | (Lamkiewicz, 2023) |
| ViralMSA | Multiple sequence alignment | (Moshiri, 2023) |
| DeepVirFinder | Virus detection | (Ren, 2023) |
| Virulign | Codon-correct pairwise alignments | (*Rega-Cev/Virulign*, 2023) |
| Palmscan | Virus detection and taxonomic classification | (Edgar, 2023) |
| SAM refiner | Variant extraction | (Gregory et al., 2021),(Gregory et al., 2022),(Yaglom et al., 2022) |
| Virstrain | Variant detection | (Ray, 2023) |
| COJAC | Variant detection | (Jahn et al., 2022) |
| Lineagespot | Variant composition & lineage assignment | (Pechlivanis et al., 2022) |
| Alcov | Variant composition | (Ellmen et al., 2021) |
| Gromstole | Variant composition | https://github.com/PoonLab/gromstole |
| LolliPop | Variant Visualization | (Dreifuss et al., 2022) |
| Freyja | Lineage assignment | (Karthikeyan et al., 2022) |
| LCS | Lineage composition | (Valieris et al., 2022), (Karthikeyan et al., 2022) |
| Kallisto | Fast pseudoalignment quantification | (Baaijens et al., 2021) |
| ViralFlye | long-read Metagenome Assembled Viruses (MAVs) | (Antipov, 2023) |

| Tool/Workflow name | Action | Citation |
|---|---|---|
| CheckV | Quality control of MAVs | (Nayfach et al., 2021) |

## Use-case

Integrated workflows or pipelines offer a streamlined and cohesive approach to data analysis, enhancing the efficiency and reliability of the entire process, in contrast to the fragmented nature of working with independent tools. The advantage of integrated workflows lies in their ability to seamlessly connect different stages of analysis, ensuring data consistency and reducing the risk of errors or data loss that can occur when manually transferring data between separate tools. In contrast, relying on independent tools often involves intricate data handling and increased risk of compatibility issues, which can lead to time-consuming and error-prone tasks in data management and analysis. Integrated workflows address these challenges, providing a more robust and efficient solution for complex data processing and analysis tasks. Furthermore, this reduces the specialist knowledge required to run specific bioinformatics pipelines, a important component for standardised, accredited WBE or genomic surveillance in general.

We have demonstrated two case studies of integrated pipelines below relevant to viral surveillance within a WBE context.

### V-pipe: A Comprehensive Bioinformatics Workflow

In response to the evolving landscape of viral genomic diversity analysis, we have developed an end-to-end bioinformatics workflow, V-pipe, now hosted on Elixir's WorkflowHub platform. This integration brings together essential components for the analysis of sequencing data derived from viral samples, whether they exhibit genomic diversity within clinical hosts or across hosts in environmental samples.

#### Key Components

- **V-pipe**: An Integrated Bioinformatics Workflow V-pipe is an integrated bioinformatics workflow designed to address the evolving challenges in the analysis of viral genomic diversity. It has been a pivotal resource within the SIB Swiss Institute of Bioinformatics (Swiss Elixir Node) since 2017(Posada-Céspedes et al., 2021). V-pipe covers the initial preprocessing steps, including quality control using prinseq, configurable alignment employing bwa2 for SARS-CoV-2, and mutation identification via pileup generation and results summarization in TSV format.
- **COJAC**: Early Detection of Emerging Variants COJAC is a crucial tool for the early detection of emerging variants, particularly useful for monitoring the spread of viral variants. It leverages the co-occurrence of signature mutations and has been instrumental in tracking the Omicron variant's distribution across 450 sites in the UK, as documented in a technical briefing (UK Health Security Agency, 2022).
- **LolliPop**: Variant Deconvolution and Relative Abundance Estimation LolliPop is specifically developed for variant deconvolution and the estimation of relative abundances, even when dealing with shared mutations among variants and complex sequencing data. It utilizes a kernel-based deconvolution approach and capitalizes on time series information in the sample set.

These three key components were seamlessly integrated during a recent biohackathon, with real-world data from the Swiss variant surveillance program in wastewater serving as a benchmark for parameterization and prototype stability. The result is an integrated, end-to-end workflow that simplifies bioinformatics analysis, replacing the need for multiple manual steps and varying

tool dependencies. This integrated environment enhances version control and standardization, catering to the needs of public health surveillance. The Swiss variant surveillance was updated to leverage this workflow.

Since its introduction at the Biohackathon 2022, the finalized version of this integrated wastewater workflow, incorporated into V-pipe version 3.0 (Fuhrmann et al., 2023), has been made accessible to a wider audience and has been successfully adopted by similar surveillance projects, underlining its value and effectiveness in the field (Zhakparov et al., 2023).

### Galaxy

During the biohackathon, we engaged in productive discussions with researchers working on the Galaxy platform, exploring ways to make various analytical tools available within Galaxy's ecosystem.

Since the conclusion of the biohackathon, there have been notable advancements in the integration of wastewater analysis tools into the Galaxy platform. Notably, tools such as COJAC are now accessible via a visual, user-friendly point-and-click interface on Galaxy (accessible at COJAC on Galaxy https://usegalaxy.eu/root?tool_id=cooc_mutbamscan). This integration enhances the accessibility of these tools, providing users with a seamless experience while also enabling their incorporation into comprehensive Galaxy workflows.

## Discussion

The primary objective of this project was to identity the opportunities for standardization within the various bioinformatic methods used in WBE. Through our exploration, we have identified several crucial factors in the analysis of wastewater data, spanning from the unique characteristics of the wastewater sample to the interpretation of results, with a significant emphasis on the pivotal role played by analysis workflows and tools. To address these considerations, we have delineated two distinct stages within omic wastewater surveillance workflows and have constructed a comprehensive framework of components (such as modules and tools) that can be effectively harnessed to create end-to-end genomic wastewater surveillance pipelines. These two omic wastewater surveillance workflows can be categorized as follows:

1. Design and *in-silico* validation of primers and benchmarking bioinformatic pipelines.
2. Application
    1. Specific target analyses, often focused on single viruses and their variants, typically relying on tiling amplicon-based methods.
    2. Analyses of unknown targets using metagenomics and metatranscriptomics approaches.

For each of these stages, we have identified key phases, excluding the sampling and wet-lab portions of the process. In this biohackathon we have demonstrated the advantage of integrated workflows such as V-pipe and the integration of tools into the Galaxy platform. Integrated pipelines/workflows enhance efficiency, reliability, and usability in a domain often fraught with complexities associated with disparate tools and formats. Their capacity to reduce the requirement for specialized knowledge is instrumental in standardizing genomic surveillance practices within public health and scientific research.

Nonetheless, given the diverse array of tools available for processing High-Throughput Sequencing data from wastewater, even within the same phase, there is a critical need for comprehensive bioinformatic benchmarking to compare and contrast the resulting outputs. This need is further underscored by the presence of various methodologies, such as amplicon-based and metagenomics/metatranscriptomics approaches, each with their own specific set of variations from primer choice to rRNA depletion. Furthermore, the outcomes and interpretations of these bioinformatic analyses are highly contingent on the chosen workflow, spanning from the wet-lab to the bioinformatics (dry-lab) stages and the original sample.

In conclusion, ongoing research efforts, with a specific focus on *in-silico* validation and the practical implementation of workflows, remain essential in advancing our comprehension of viral presence and distribution within wastewater systems using bioinformatic approaches for virus surveillance.

## Acknowledgements

Some of the authors were funded by ELIXIR, the research infrastructure for life-science data, to join the BioHackathon Europe.

## References

Antipov, D. (2023). *viralFlye*. https://github.com/Dmitry-Antipov/viralFlye

*ARIBA*. (2023). Pathogen Informatics, Wellcome Sanger Institute. https://github.com/sanger-pathogens/ariba

Baaijens, J. A., Zulli, A., Ott, I. M., Petrone, M. E., Alpert, T., Fauver, J. R., Kalinich, C. C., Vogels, C. B. F., Breban, M. I., Duvallet, C., McElroy, K., Ghaeli, N., Imakaev, M., Mckenzie-Bennett, M., Robison, K., Plocik, A., Schilling, R., Pierson, M., Littlefield, R., ... and, M. B. (2021). *Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-seq quantification*. https://doi.org/10.1101/2021.08.31.21262938

Constantinides, B. (2023). *Hostile*. https://github.com/bede/hostile

Dreifuss, D., Topolsky, I., Icer Baykal, P., & Beerenwinkel, N. (2022). Tracking SARS-CoV-2 genomic variants in wastewater sequencing data with LolliPop. *medRxiv*. https://doi.org/10.1101/2022.11.02.22281825

Edgar, R. (2023). *Rcedgar/palmscan*. https://github.com/rcedgar/palmscan

Ellmen, I., Lynch, M. D. J., Nash, D., Cheng, J., Nissimov, J. I., & Charles, T. C. (2021). *Alcov: Estimating variant of concern abundance from SARS-CoV-2 wastewater sequencing data*. https://doi.org/10.1101/2021.06.03.21258306

Faucon, C. (2023). *PBSIM*. https://github.com/pfaucon/PBSIM-PacBio-Simulator

Fuhrmann, L., Jablonski, K. P., Topolsky, I., Batavia, A. A., Borgsmüller, N., Baykal, P. I., Carrara, M., Chen, C., Dondi, A., Dragan, M., Dreifuss, D., John, A., Langer, B., Okoniewski, M., Plessis, L. du, Schmitt, U., Singer, F., Stadler, T., & Beerenwinkel, N. (2023). V-pipe 3.0: A sustainable pipeline for within-sample viral genetic diversity estimation. *bioRxiv*. https://doi.org/10.1101/2023.10.16.562462

Gourlé, H. (2023). *InSilicoSeq*. https://github.com/HadrienG/InSilicoSeq

Gregory, D. A., Trujillo, M., Rushford, C., Flury, A., Kannoly, S., San, K. M., Lyfoung, D. T., Wiseman, R. W., Bromert, K., Zhou, M.-Y., Kesler, E., Bivens, N. J., Hoskins, J., Lin, C.-H., O'Connor, D. H., Wieberg, C., Wenzel, J., Kantor, R. S., Dennehy, J. J., & Johnson, M. C. (2022). Genetic diversity and evolutionary convergence of cryptic SARS-CoV-2 lineages detected via wastewater sequencing. *PLOS Pathogens*, *18*(10), e1010636. https://doi.org/10.1371/journal.ppat.1010636

Gregory, D. A., Wieberg, C. G., Wenzel, J., Lin, C.-H., & Johnson, M. C. (2021). Monitoring SARS-CoV-2 populations in wastewater by amplicon sequencing and using the novel program SAM refiner. *Viruses*, *13*(8), 1647. https://doi.org/10.3390/v13081647

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics*, *28*(4), 593–594. https://doi.org/10.1093/bioinformatics/btr708

Jahn, K., Dreifuss, D., Topolsky, I., Kull, A., Ganesanandamoorthy, P., Fernandez-Cassi, X., Bänziger, C., Devaux, A. J., Stachler, E., Caduff, L., Cariti, F., Corzón, A. T., Fuhrmann, L.,

Chen, C., Jablonski, K. P., Nadeau, S., Feldkamp, M., Beisel, C., Aquino, C., ... Beerenwinkel, N. (2022). Early detection and surveillance of SARS-CoV-2 genomic variants in wastewater using COJAC. *Nature Microbiology*, *7*(8), 1151–1160. https://doi.org/10.1038/s41564-022-01185-x

jiarong. (2023). *VirSorter 2*. https://github.com/jiarong/VirSorter2

Karthikeyan, S., Levy, J. I., Hoff, P. D., Humphrey, G., Birmingham, A., Jepsen, K., Farmer, S., Tubb, H. M., Valles, T., Tribelhorn, C. E., Tsai, R., Aigner, S., Sathe, S., Moshiri, N., Henson, B., Mark, A. M., Hakim, A., Baer, N. A., Barber, T., ... Knight, R. (2022). Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. *Nature*, *609*(7925), 101–108. https://doi.org/10.1038/s41586-022-05049-6

Lamkiewicz, K. (2023). *Klamkiew/viralclust*. https://github.com/klamkiew/viralclust

*Longislnd*. (2020). The Bioinformatics Repository. https://github.com/bioinform/longislnd

*Marbl/Krona*. (2023). MarBL. https://github.com/marbl/Krona

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

Moshiri, N. (2023). *ViralMSA*. https://github.com/niemasd/ViralMSA

*Mykrobe*. (2023). Mykrobe-tools. https://github.com/Mykrobe-tools/mykrobe

N'Guessan, A., Tsitouras, A., Sanchez-Quete, F., Goitom, E., Reiling, S. J., Galvez, J. H., Nguyen, T. L., Nguyen, H. T. L., Visentin, F., Hachad, M., Krylova, K., Matthews, S., Kraemer, S. A., Stretenowich, P., Bourgey, M., Djambazian, H., Chen, S.-H., Roy, A.-M., Brookes, B., ... Shapiro, B. J. (2022). *Detection of prevalent SARS-CoV-2 variant lineages in wastewater and clinical sequences from cities in québec, canada*. https://doi.org/10.1101/2022.02.01.22270170

*NanoSim*. (2023). BC Cancer Canada's Michael Smith Genome Sciences Centre. https://github.com/bcgsc/NanoSim

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology*, *39*(5), 578–585. https://doi.org/10.1038/s41587-020-00774-7

*Nf-core/viralrecon*. (2023). nf-core. https://github.com/nf-core/viralrecon

Nies, L. de, Lopes, S., Busi, S. B., Galata, V., Heintz-Buschart, A., Laczny, C. C., May, P., & Wilmes, P. (2021). PathoFact: A pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome*, *9*(1), 49. https://doi.org/10.1186/s40168-020-00993-9

Ozer, E. A. (2023). *IN_SILICO_PCR*. https://github.com/egonozer/in_silico_pcr

Pechlivanis, N., Tsagiopoulou, M., Maniou, M. C., Togkousidis, A., Mouchtaropoulou, E., Chassalevris, T., Chaintoutis, S. C., Petala, M., Kostoglou, M., Karapantsios, T., Laidou, S., Vlachonikola, E., Chatzidimitriou, A., Papadopoulos, A., Papaioannou, N., Dovas, C. I., Argiriou, A., & Psomopoulos, F. (2022). Detecting SARS-CoV-2 lineages and mutational load in municipal wastewater and a use-case in the metropolitan area of thessaloniki, greece. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-06625-6

Posada-Céspedes, S., Seifert, D., Topolsky, I., Jablonski, K. P., Metzner, K. J., & Beerenwinkel, N. (2021). V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btab015

Ray. (2023). *VirStrain*. https://github.com/liaoherui/VirStrain

*Rega-cev/virulign*. (2023). rega-cev. https://github.com/rega-cev/virulign

Ren, J. J. (2023). *DeepVirFinder: Identifying viruses from metagenomic data by deep learning*. https://github.com/jessieren/DeepVirFinder

Seemann, T. (2023). *ABRicate*. https://github.com/tseemann/abricate

sej917. (2023). *BEAR*. https://github.com/sej917/BEAR

*Silica*. (2023). GEAR. https://github.com/gear-genomics/silica

*Staramr*. (2023). National Microbiology Laboratory. https://github.com/phac-nml/staramr

*SWAMPy*. (2023). Goldman Group EBI. https://github.com/goldman-gp-ebi/SWAMPy

*The NEAT Project v4.0*. (2023). NCSA. https://github.com/ncsa/NEAT

UK Health Security Agency. (2022). *Technical briefing 30*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1038404/Technical_Briefing_30.pdf

Valieris, R., Drummond, R. D., Defelicibus, A., Dias-Neto, E., Rosales, R. A., & Silva, I. T. da. (2022). A mixture model for determining SARS-cov-2 variant composition in pooled samples. *Bioinformatics*, *38*(7), 1809–1815. https://doi.org/10.1093/bioinformatics/btac047

*Viral-ngs*. (2023). Broad Institute. https://github.com/broadinstitute/viral-ngs

*VIRify*. (2023). MGnify. https://github.com/EBI-Metagenomics/emg-viral-pipeline

Wick, R. R. (2019). *Badread: Simulation of error-prone long reads*. https://doi.org/10.21105/joss.01316

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46

Xue, Z. (2023). *Grinder*. https://github.com/zyxue/biogrinder

Yaglom, H. D., Maurer, M., Collins, B., Hojnacki, J., Monroy-Nieto, J., Bowers, J. R., Packard, S., Erickson, D. E., Barrand, Z. A., Simmons, K. M., Brock, B. N., Lim, E. S., Smith, S., Hepp, C. M., & Engelthaler, D. M. (2022). One health genomic surveillance and response to a university-based outbreak of the SARS-CoV-2 delta AY.25 lineage, arizona, 2021. *PLOS ONE*, *17*(10), e0272830. https://doi.org/10.1371/journal.pone.0272830

Zhakparov, D., Quirin, Y., Xiao, Y., Battaglia, N., Holzer, M., Bühler, M., Kistler, W., Engel, D., Zumthor, J. P., Caduff, A., & Baerenfaller, K. (2023). Sequencing of SARS-CoV-2 RNA fragments in wastewater detects the spread of new variants during major events. *Microorganisms*, *11*(11). https://doi.org/10.3390/microorganisms11112660