Introduction

The detection of outbreaks caused by a novel pathogen usually depends on healthcare professionals noticing and reporting unusual cases. However, this approach can be inefficient and slow, particularly for diseases with a high proportion of asymptomatic infections (e.g. poliovirus) or delayed symptom onset (e.g. HIV/AIDS). Such “stealth outbreaks” can spread undetected for a longer period, delaying public health response and complicating containment efforts.

To address the challenges posed by stealth outbreaks, it is crucial to develop early warning systems that do not depend on symptomatic patient reports. One promising approach is to leverage environmental surveillance. Even asymptomatic carriers often shed pathogen particles, which can be detected in samples from strategically selected locations, such as metropolitan areas or high-volume transit hubs. Sample types may include wastewater, air, or pooled specimens like nasal swabs and blood samples. Sequencing these samples produces metagenomic profiles detailing the composition of organism communities, enabling the identification of outbreaks by detecting a pathogen's increasing abundance over time, even without prior knowledge of the pathogen.

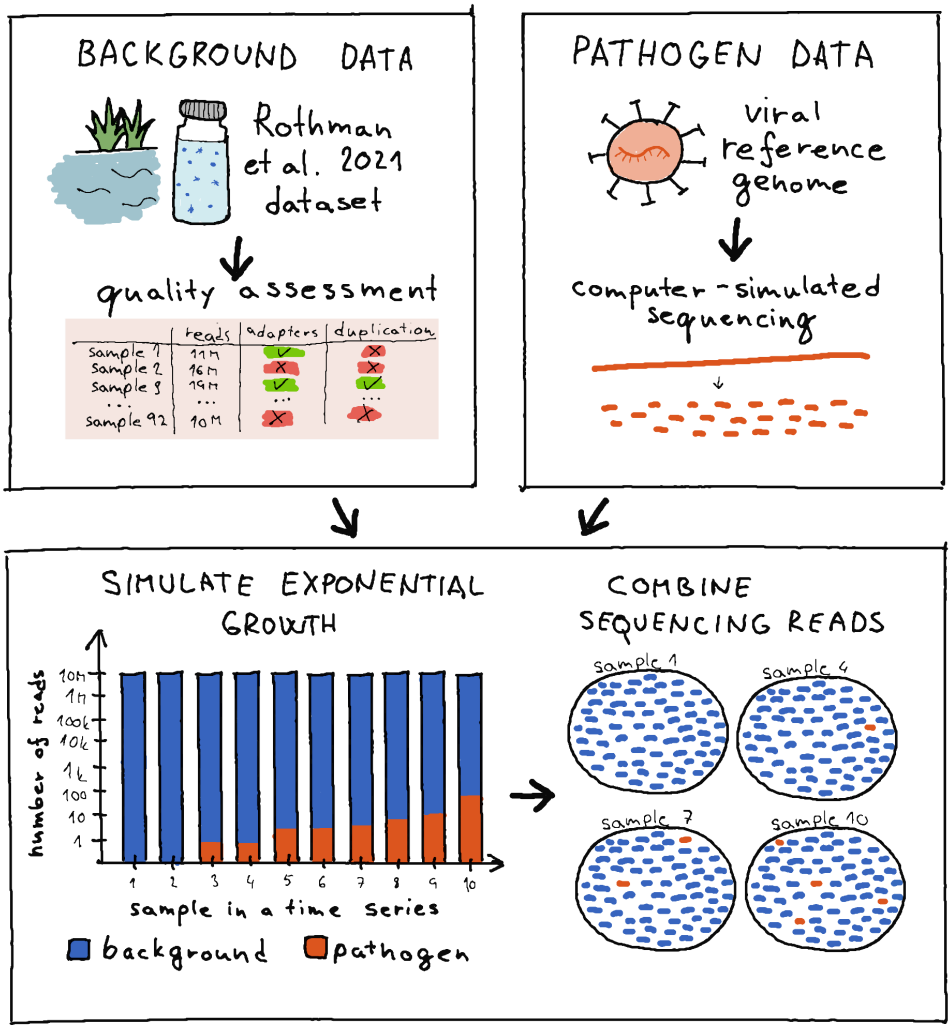
Developing reliable workflows to detect such signals requires datasets where outbreak scenarios are explicitly modeled. However, obtaining such datasets from natural conditions presents significant challenges: it demands extensive sampling, and the exact dynamics of an outbreak—such as the proportion of infected individuals contributing to the samples—are often unknown. Moreover, to ensure the robustness of detection workflows, it is necessary to test them under diverse conditions, including varying pathogen characteristics and transmission parameters.

Simulated datasets offer a practical alternative. By controlling the parameters of the outbreak and the dataset, it should be possible to systematically explore different scenarios and validate data analysis workflows. This work outlines and demonstrates a methodology for creating a simulated dataset of a viral outbreak. First, we selected and processed a time series of metagenomic wastewater data from a public database to use as a realistic background. Next, we chose a pathogen genome from a public repository, and generated computer-simulated sequencing reads of the pathogen. Finally, we modeled pathogen abundance over time based on an exponential growth curve, and combined the background and pathogen reads in the calculated ratios.

Beyond practically implementing these steps, we prioritized outlining the general guidelines and key considerations of simulating such metagenomic datasets. This facilitates further development of the workflow and adaptation of the inputs and the methodology for simulation of various outbreak scenarios.

Methods and Results

Fig 1. Overview of the workflow for generating simulated outbreak metagenomic dataset

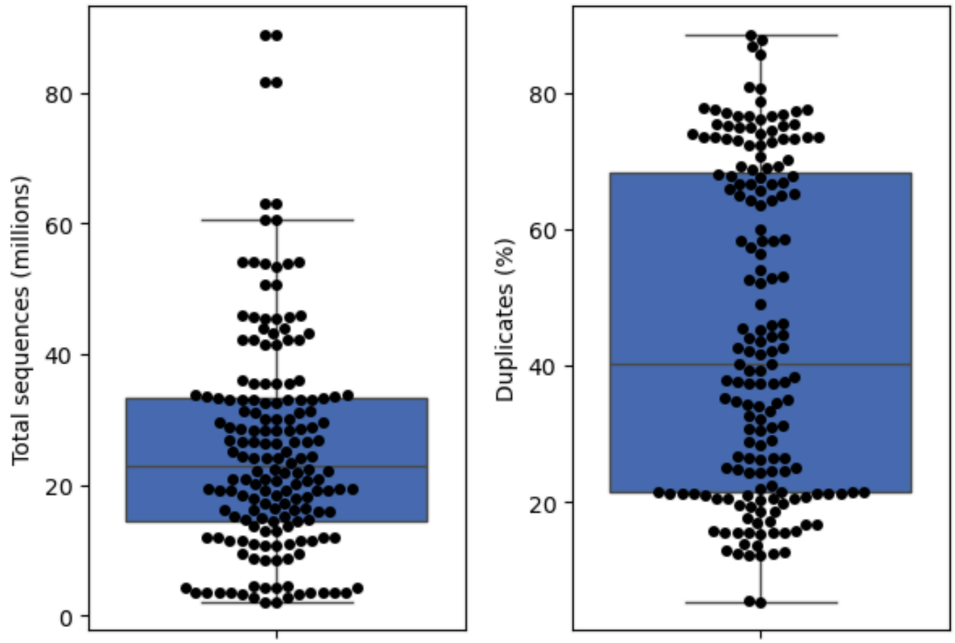


Background metagenomic data

We downloaded wastewater metagenomic dataset published in Rothman *et al.* 2021 (1) mapping RNA virome in Southern California from August 2020 to January 2021. Since this dataset contains samples from multiple water treatment plants and even different methods of sequencing library preparation, we use only data from one of the plants (HTP) and the library that underwent only non-specific PCR amplification (isn't enriched for any particular panel of sequences). We downloaded selected samples using the NCBI SRA Toolkit (2) and then assessed the quality of the reads using FastQC and MultiQC tools (3, 4). We found that the number of reads ranged from 2 to 89 M pairs among the 92 samples. As expected, these unprocessed reads contain sequencing adapters. Non-specific amplification during library preparation also produced significant content of duplicated reads (Fig. 2). Four samples

encountered unexpected memory or parsing errors when processed by quality control tools and were excluded from further processing. Code for this step is available in the project repository (5) at rules/download_background.smk. For the preparation of the simulated dataset, we further excluded samples with fewer than 5 million reads and then selected 20 consecutive samples with less than 7-day intervals to ensure that all stages of the exponential growth curve within the given timeframe are sufficiently covered.

Fig. 2 Boxplot of sample read counts and proportion of duplicated reads in samples



Pathogen data

Next, we downloaded a reference genome of a suitable pathogen. For this demonstration, we chose Nipah virus, since this RNA virus causing lethal encephalitis is not expected to be present in the background data. We generated 100’000 computer-simulated Illumina reads using InSilicoSeq (6). This tool does not allow modification of the read size, defaulting to 150 base pairs, which corresponds to the read length of a portion of the chosen background datasets. These computer-simulated reads lack adapter sequences. For this demonstration, we did not attempt to include adapters, but this might be a useful feature to be added to the workflow later (7) to better simulate the background dataset sequencing. Code for this step is available in the project repository (5) at scripts/simulate_pathogen_reads.sh

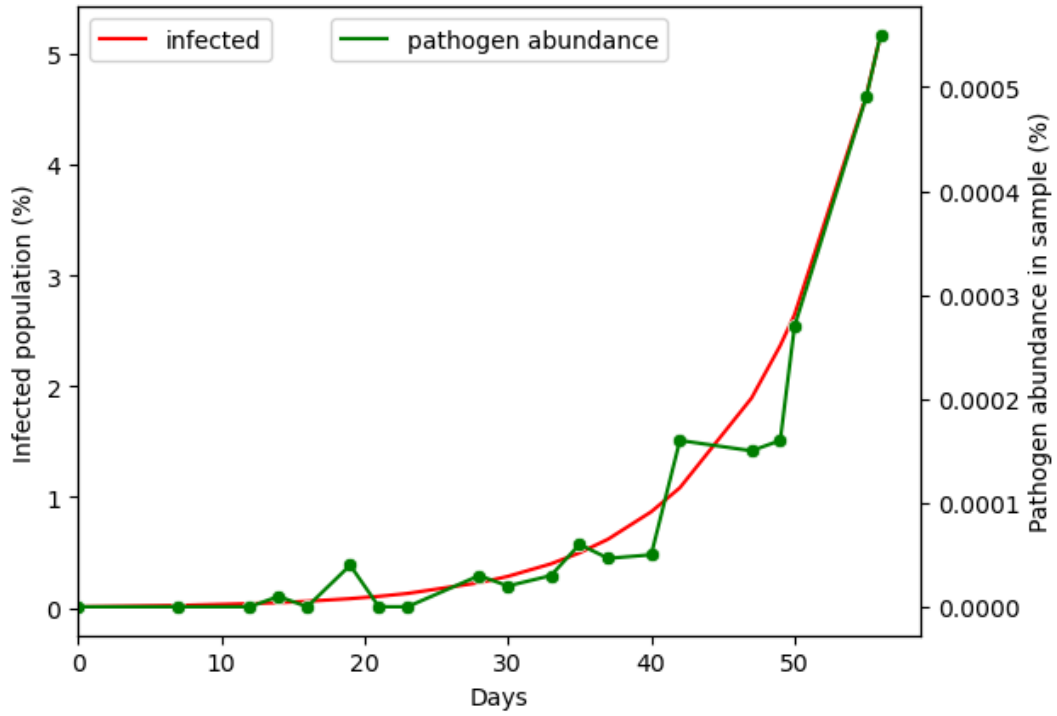
Growth calculation and subsampling

Then, we calculated the expected abundance of pathogen reads in the metagenomic data time series. The amount of viral reads is low in shotgun metagenomic data from wastewater samples. While methodologies to increase the proportion by PCR amplification exist, they can make the method less sensitive to very diverged (e.g. engineered) viruses. Previous studies

have shown that the estimated hypothetical abundance of viruses in wastewater samples if 100% of the contributing population was infected at the same time is about 0.0006% (6 pathogen reads per 1’000’000 total) for SARS-CoV-2 and 0.1% (1 pathogen read per 1’000 total) for a norovirus (8). For this demonstration, we chose the value of this parameter (also referred to as coefficient B) 0.01%, which is inside the previously reported range.
Further, we selected 10 million reads as an expected sample size for the simulated dataset, since this is a common size for wastewater metagenomic samples, but still reasonable in terms of computational resource requirements and processing speed. Additional parameters such as the reproduction number and generation time were manually selected for achieving ~5% infection prevalence in the simulated dataset, but can be easily modified to simulate scenarios with variable outbreak dynamics. This simplified model does not take into account parameters such as the reproduction number changing over time as the number of susceptible hosts decreases or the duration it takes an infected person to recover, making this model less realistic but still sufficient for modeling low prevalence stages (<0.5% infected), where researchers are the most likely to use early detection workflows.

Calculated pathogen abundance (Fig 3.) was then used to estimate the number of background and pathogen reads in the sample of a given size using selection from a Poisson distribution. The number of pathogen reads in the samples ranges from 0 to 55 reads. Finally, we subsampled the background and pathogen reads based on this model using seqtk tool (9). The dataset created in this demonstration is available at Zenodo (10). Code for this step is available in the project repository (5) at [notebooks/calculate_outbreak_growth.ipynb] and [scripts/subsample_and_combine_reads.sh]

Fig. 3 Simulated pathogen abundance in the metagenomic time series dataset



Discussion

In this project, we demonstrated a methodology for creating simulated metagenomic datasets with exponentially increasing pathogen abundance. These datasets can be used to develop and test workflows to detect novel pathogen outbreaks. Simulated metagenomic datasets also allow convenient performance comparison of different detection workflow variants such as reference-based versus reference-free approaches (11). This facilitates development of a workflow that can reliably detect outbreaks of a wide range of pathogens even in datasets with limited sample size and avoid false positives animal pathogens or other natural sequences.

To enhance our workflow and enable simulation of a wider range of outbreak scenarios, several improvements and extensions can be considered (described in more detail in the project repository “Issues” section):

1. Unprocessed background metagenomic datasets contain technical artifacts such as read duplication and presence of adapters, which would be typically addressed during the preprocessing stage of detection workflows. Adding simulated pathogen sequences (which do not contain these artifacts) before the preprocessing might result in a skewed sample composition profile. Adding artifacts to the pathogen reads could address this issue.

2. In the computer-simulated sequencing reads, parameters such as the read length, fragment length, and sequencing error rate should correspond to the background data. We could develop methods to infer these parameters from background data and apply them to the computer-simulated sequencing.

3. Due to low abundance of viral reads in shotgun wastewater samples, PCR amplification is often used to aid detection. Simulation of this kind of datasets should be considered.

4. Our demonstration uses a simplified model of a pathogen outbreak. Including variables such as gradual decrease of viral particle shedding by recovering individuals and fluctuations of the reproduction number would make the simulations more realistic. Additionally, introducing random noise to the model could also further increase real-world faithfulness of the simulated datasets.

Developing functional growth detection workflows is a complex, multifaceted task. I hope researchers tackling this challenge will find my work on dataset simulation valuable as a practical starting point. By focusing on workflow documentation, customization, and potential enhancements, I aimed to provide a solid foundation for generating simulated metagenomic datasets, ultimately supporting advances in outbreak detection and public health response.

Acknowledgement

I'd like to thank Mike McLaren, Evan Fields, and Dan Rice from the Nucleic Acid Observatory project for their valuable and abundant guidance and feedback.

References

1. Rothman, J. A., Loveless, T. B., Kapcia III, J., Adams, E. D., Steele, J. A., Zimmer-Faust, A. G., ... & Whiteson, K. L. (2021). RNA viromics of Southern California wastewater and detection of SARS-CoV-2 single-nucleotide variants. *Applied and environmental microbiology*, *87*(23), e01448-21.
2. https://github.com/ncbi/sra-tools
3. Andrews, S. (2010, April). *FastQC: a quality control tool for high throughput sequence data*.
4. Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047-3048.
5. https://github.com/437364/outbreak_metagenomic_sim
6. Gourlé, H., Karlsson-Lindsjö, O., Hayer, J., & Bongcam-Rudloff, E. (2019). Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*, *35*(3), 521-522.
7. https://github.com/437364/outbreak_metagenomic_sim/issues/1
8. Grimm, S. L., Kaufman, J. T., Rice, D. P., Whittaker, C., Bradshaw, W. J., & McLaren, M. R. (2023). Inferring the sensitivity of wastewater metagenomic sequencing for pathogen early detection. *medRxiv*, 2023-12.
9. Li, H. (2023). seqtk Toolkit for processing sequences in FASTA/Q formats. *GitHub*, *767*, 69.
10. https://zenodo.org/records/14635401

11. Consortium, T. N. A. O. (2021). A global nucleic acid observatory for biodefense and planetary health. *arXiv preprint arXiv:2108.02678*.