A Workflow for Detecting Cryptic SARS-CoV-2 Lineages Using Sequencing Data from Wastewater Samples

Veedhi Solanki

Student Number - 1301912

Advisors: Dr. Ryan Gregory, Dr. Lawrence Goodridge, Dr. Angela Canovas

BINF*6999 – Bioinformatics Masters Project

August 13, 2024

TABLE OF CONTENTS

ABSTRACT

Wastewater surveillance has become an essential tool in monitoring the spread and evolution of SARS-CoV-2, offering insights into viral prevalence and the emergence of new variants. This study focused on developing a specialized analytical pipeline to detect cryptic SARS-CoV-2 lineages, which are low-abundance variants with significant genetic divergence. These lineages often represent early-stage variants or result from prolonged within-host evolution or recombination events. The pipeline was validated on a dataset from New York City (NYC) and applied to a subset of the Ontario Wastewater Surveillance Initiative's dataset, focusing on samples from the University of Guelph.

The pipeline, incorporating tools such as Minimap2 for alignment, FastTree and IQ-TREE for phylogenetic analysis, and custom scripts for dereplication and clustering, effectively identified cryptic lineages in both datasets. While the NYC dataset revealed several divergent cryptic lineages, the Ontario dataset showed lineages that were still evolving, suggesting that these lineages had not yet reached the level of divergence observed in NYC. The analysis also indicated that within-host evolution was more prevalent than recombination events, particularly in the later stages of the pandemic.

The study highlights the scalability and adaptability of the pipeline for broader applications in viral surveillance, particularly for monitoring the ongoing evolution of SARS-CoV-2. However, the analysis of larger datasets will require significant computational resources, and the findings emphasize the need for further validation through experimental studies. Future research should focus on the temporal and geographical tracking of viral sequences, the investigation of shedding

patterns in the gut, and the ethical considerations associated with viral surveillance. The pipeline developed in this study provides a robust foundation for detecting cryptic lineages and contributes to ongoing efforts to manage the COVID-19 pandemic effectively.

Key words:

SARS-CoV-2, Cryptic Lineages, Wastewater Surveillance, Phylogenetic Analysis, Bioinformatics, Genomic Surveillance

List of abbreviations:

SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2

VCF: Variant Call Format

BAM: Binary Alignment Map

FASTA: Fast-All (text-based format for nucleotide sequences)

MAFFT: Multiple Alignment using Fast Fourier Transform

IQ-TREE: Efficient Tree Reconstruction using Maximum Likelihood

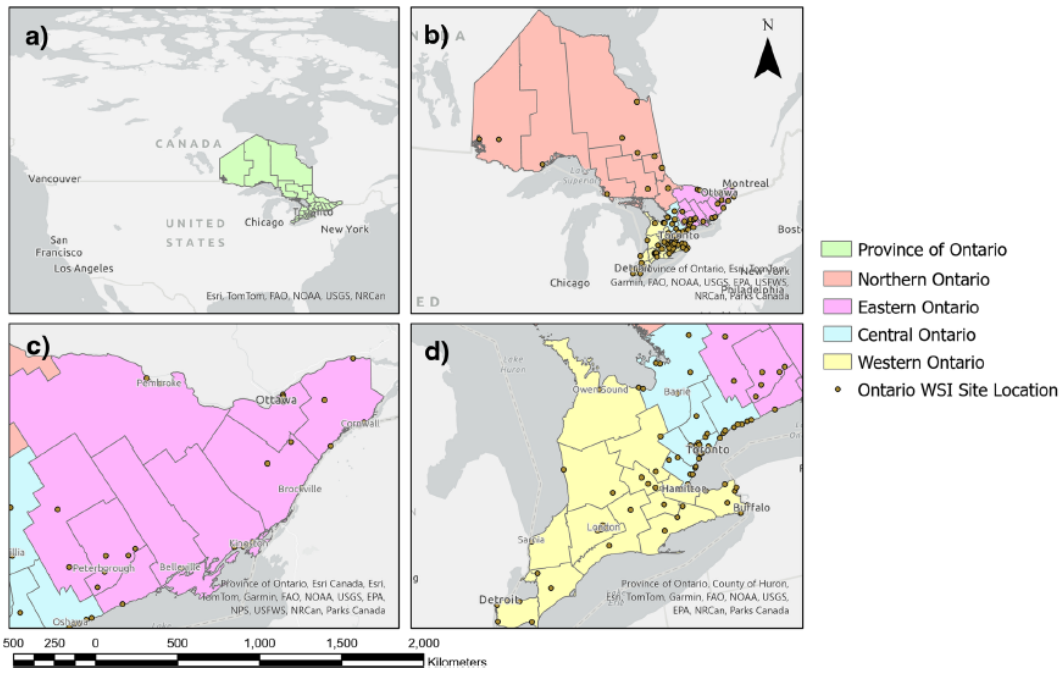
MEGA: Molecular Evolutionary Genetics Analysis

INTRODUCTION

Wastewater surveillance has emerged as a vital tool in monitoring the spread and evolution of SARS-CoV-2, providing a comprehensive overview of viral prevalence across communities without the need for individual testing. Wastewater surveillance offers a unique window into the health of a community by detecting viral particles shed in human waste. This approach allows for the early detection of increases in infection rates and the identification of new, potentially concerning variants, often before clinical cases become evident. During the COVID-19 pandemic, wastewater surveillance played a crucial role in public health, offering critical insights into the

epidemiology of the virus, particularly when clinical testing was limited or unavailable. This method has proven especially effective in tracking the emergence and spread of variants of concern (VOCs), which continue to pose significant challenges to public health efforts (Hart & Halden, 2020; Ahmed et al., 2022).

Ontario has established a world-leading wastewater surveillance program that has been instrumental in monitoring the spread of SARS-CoV-2 across the province. The program, initiated in April 2020, involves regular sampling from wastewater treatment plants across major cities as shown in ***Figure 1***, including Toronto, Ottawa, and Hamilton (Ontario Wastewater Surveillance Initiative, 2022). These samples are collected weekly and analyzed for the presence of SARS-CoV-2, providing a non-invasive method to track the virus's spread and the emergence of new variants. The program's extensive coverage, spanning over 175 treatment plants, has enabled Ontario to maintain a robust surveillance network that has been critical in informing public health decisions (Naughton et al., 2021).



Source: *Adapted from D'Aoust, Patrick M., et al. "SARS-CoV-2 viral titer measurements in Ontario, Canada wastewaters throughout the COVID-19 pandemic." Scientific Data 11.1 (2024): 656.*

Figure 1: Geographic distribution of Ontario Wastewater Surveillance Initiative (WSI) testing sites. *The maps illustrate the widespread geographic coverage of wastewater testing locations across the province, providing a comprehensive approach to monitoring SARS-CoV-2 variants within different regions.*

Wastewater surveillance is particularly valuable for its ability to detect early signs of infection spikes and the presence of new or emerging variants. For instance, the detection of the Alpha, Delta, and Omicron variants in Ontario's wastewater was a precursor to subsequent clinical case surges (Crits-Christoph et al., 2021). These early warnings allowed for timely public health interventions, mitigating the impact of these variants on the healthcare system. Moreover, the ability to detect new and emerging variants through wastewater has underscored the importance of this approach in managing the pandemic (Barber et al., 2022).

A key aspect of wastewater surveillance is its potential to identify "cryptic lineages"—SARS-CoV-2 variants that are present at very low abundance and are highly divergent from the dominant strains in circulation as shown in ***Figure 2***. These cryptic lineages may represent early-stage variants that have the potential to evolve into new VOCs or may be the result of long-term, chronic infections in immunocompromised individuals (Gregory et al., 2023). Understanding the origins and evolution of these cryptic lineages is crucial, as many significant VOCs, including the first Omicron variant and the more recent XBB recombinant lineage, are believed to have arisen from such evolutionary pathways (Zhang et al., 2022).

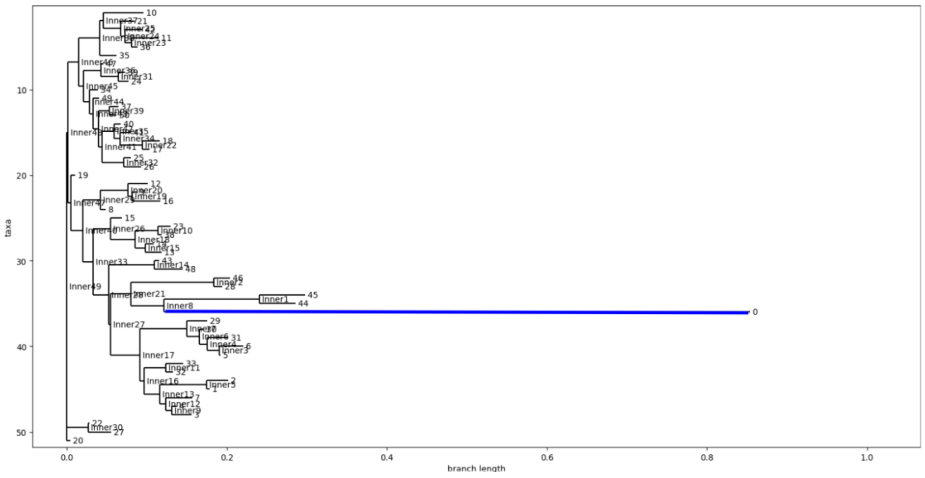


***Figure 2**: **Phylogenetic tree highlighting a cryptic SARS-CoV-2 lineage in blue.** This lineage, characterized by a distinct long branch, represents a divergent evolutionary path, potentially indicating prolonged within-host evolution or recombination events.*

Cryptic lineages can emerge from two primary mechanisms: recombination during co-infection with multiple variants and within-host evolution during chronic infections (Kantor et al., 2021). Recombination occurs when an individual is simultaneously infected with more than one variant, leading to the exchange of genetic material between the variants, potentially giving rise to new variants. Within-host evolution, on the other hand, occurs when the virus accumulates mutations over time within a single host, often during chronic infection. This process can lead to the emergence of highly divergent lineages that may possess novel characteristics, such as increased transmissibility or immune evasion (Chen et al., 2022).

The significance of these cryptic lineages cannot be understated. Many major new variants of concern, including the Omicron variant and its sublineages like BA.2.86, have emerged through these mechanisms. The detection and monitoring of cryptic lineages are therefore essential for anticipating and mitigating the impact of future variants (Teyssou et al., 2021).

Several studies in the United States have demonstrated the efficacy of wastewater surveillance in detecting cryptic lineages. Notably, studies conducted in New York City and Missouri have identified unique SARS-CoV-2 lineages that were not detected through clinical testing but were prevalent in the community's wastewater (Gregory et al., 2022; Crits-Christoph et al., 2021). These findings highlight the importance of wastewater surveillance in uncovering hidden variants that may have significant public health implications.

Despite Ontario's comprehensive wastewater surveillance program, there has been limited focus on detecting and characterizing cryptic lineages within the province's dataset. Ontario's wastewater data have primarily been used to monitor the abundance of SARS-CoV-2 and the prevalence of known VOCs, leaving a gap in the detection of cryptic lineages. This gap presents an opportunity to refine existing analytical pipelines and apply them to Ontario's extensive dataset to uncover potentially significant cryptic lineages.

In this project, we aim to address this gap by constructing an analytical pipeline to detect cryptic lineages within the wastewater and applying it to a subset of the Ontario wastewater dataset. Our goal is to demonstrate a proof-of-concept analysis that identifies cryptic lineages and provides recommendations for scaling up the study across the entire dataset. By leveraging Ontario's robust wastewater surveillance infrastructure, this project seeks to enhance our understanding of SARS-CoV-2's evolution and contribute to ongoing efforts to manage the pandemic effectively. It helps to see the frequency of cryptic lineages and their importance.

MATERIALS AND METHODS

Pipeline Development and Justification

To effectively detect cryptic lineages of SARS-CoV-2 in wastewater samples, we developed a specialized analytical pipeline tailored to the characteristics of large-scale environmental sequencing data as the flowchart shown. The development of the analytical pipeline for detecting cryptic SARS-CoV-2 lineages in wastewater samples was informed by an extensive review of existing bioinformatics tools and methodologies. ***Figure 3*** demonstrates the comprehensive workflow developed to detect cryptic SARS-CoV-2 lineages using wastewater sequencing data. The purple-highlighted steps represent methods adapted from the NYC and Missouri papers (Smyth et al., 2022 and Gregory et al., 2023), which focus on variant abundance and frequency across different samples, primarily utilizing the SAM Refiner tool. These steps were foundational in previous studies aimed at tracking known variants. In contrast, the green-highlighted steps illustrate the novel additions in our workflow, specifically designed to identify and analyze cryptic lineages. This targeted approach has been applied to the Ontario dataset for the first time, marking a significant advancement in wastewater-based epidemiology.

The pipeline was first tested on a well-characterized New York dataset to verify its accuracy and robustness before being applied to the Ontario data. This initial validation ensured that the pipeline was capable of detecting cryptic lineages in diverse datasets, setting a solid foundation for its application to the more extensive Ontario dataset.

Datasets

New York Dataset:

The NYC dataset, obtained from NCBI's Sequence Read Archive (SRA) under accession number PRJNA715712, was used to validate the pipeline. This dataset includes raw sequencing reads from nearly 5,000 wastewater samples collected globally, with a particular focus on 172 samples from New York state (Martinez-Perez et al., 2024). The specific SRA accessions analyzed were SRR15202279, SRR15384049, SRR15291304, SRR15128978, SRR15128983, SRR15202284, and SRR15202285. These samples span from 2020 to 2021 and were selected based on their relevance to cryptic lineage detection.

Ontario Dataset:

Due to the vast size of the Ontario dataset, the proof of concept was conducted using samples collected from the University of Guelph's residence at College Avenue West between October 2021 and April 2022. These samples were chosen for their representativeness and the availability of comprehensive metadata. The dataset includes regular wastewater sampling data, providing a robust basis for testing the pipeline's effectiveness in identifying cryptic lineages within a defined community setting.

Data Acquisition and Preprocessing

Merging and Dereplication

For both datasets, the initial step involved obtaining raw sequencing reads in FASTQ format. The NYC dataset was directly downloaded from the SRA, while the Ontario dataset was locally obtained. The sequencing reads for each sample were initially in paired-end format, which required merging. The merging was performed using custom shell scripts designed to concatenate the forward and reverse reads into a single file.

```
for sample in $(ls *_R1_001.fastq_file.fastq | sed 's/_R1_001.fastq_file.fastq//'); do
echo "Processing sample: $sample"
```

```
cat "${sample}_R1_001.fastq_file.fastq" "${sample}_R2_001.fastq_file.fastq" >
"${sample}_merged.fastq"
done
```

After merging, dereplication was performed to remove redundant sequences, thereby reducing data size and computational load while retaining unique sequences. Dereplication was executed using a custom Python script integrated into the pipeline. The **derep.py** script, originally adapted from the workflow used in the Missouri study (Hepp et al., 2021), was modified to ensure compatibility with our datasets, particularly to handle the specific characteristics of the Ontario and New York sequencing data.

```
for sample in $(ls *_merged.fastq | sed 's/_merged.fastq//'); do
  echo "Dereplicating ${sample}..."
  python derep.py "${sample}_merged.fastq" "${sample}_derep.fastq" 10
done
```

Sequence Alignment

Following dereplication, the sequences were mapped to the SARS-CoV-2 reference genome (NC_045512.2) using **Minimap2**, a versatile aligner optimized for short and long reads. This tool was selected due to its ability to efficiently align large datasets against reference genomes, a critical requirement given the extensive nature of the Ontario dataset.

```
for sample in $(ls *_derep.fastq | sed 's/_derep.fastq//'); do
  echo "Mapping sequences for ${sample}..."
  minimap2 -ax sr NC_045512.2.fasta "${sample}_derep.fastq" > "${sample}_mapped.sam"
done
```

The mapped sequences were then converted to **BAM** format, sorted, and indexed using **SAMtools**, ensuring that the data was properly formatted and ready for subsequent analysis.

Phylogenetic Analysis

Ontario

For the Ontario dataset, due to its size, **FastTree** was employed to construct phylogenetic trees. Clustering was performed using **CD-HIT** with a threshold of 0.63 to reduce redundancy by grouping similar sequences, which helped streamline the dataset and reduce computational load before constructing phylogenetic trees. FastTree is known for its speed and ability to handle large datasets while still producing reliable phylogenetic trees. Prior to tree construction, the sequences were filtered to remove short reads and clustered to reduce redundancy, ensuring that only unique and significant sequences were included in the analysis.

```
def build_tree(sample_id, bam_type='sorted'):
    bam_file = f"{sample_id}_{bam_type}.bam"
    fasta_file = f"{sample_id}.fasta"
    phylogenetic_tree_file = f"{fasta_file}.treefile"
    print(f"Converting {bam_file} to FASTA...")
    subprocess.run(f"samtools fasta {bam_file} > {fasta_file}", shell=True, check=True)
    print(f"Filtering short sequences from {fasta_file}...")
    filtered_fasta_file = filter_short_sequences(fasta_file)
    print(f"Clustering sequences to reduce redundancy...")
    clustered_fasta_file = cluster_sequences(filtered_fasta_file, threshold=0.90)
    print(f"Building phylogenetic tree for {clustered_fasta_file} with FastTree...")
```

```
    subprocess.run(f"FastTree -nt {clustered_fasta_file} > {phylogenetic_tree_file}", shell=True,
check=True)
    return phylogenetic_tree_file
```

New York

we used **MAFFT** for multiple sequence alignment. MAFFT was chosen due to its accuracy and efficiency in aligning large sets of sequences, making it ideal for handling the diverse and extensive NYC dataset. For the New York dataset, **IQ-TREE** was used for phylogenetic tree construction. IQ-TREE was chosen for its advanced models of sequence evolution and ability to assess the reliability of inferred trees through statistical support values like bootstrap analysis.

```
def build_tree(sample_id, bam_type='sorted'):
    bam_file = f"{sample_id}_{bam_type}.bam"
    fasta_file = f"{sample_id}.fasta"
    aligned_fasta_file = f"{sample_id}_aligned.fasta"
    phylogenetic_tree_file = f"{aligned_fasta_file}.treefile"
    print(f"Aligning sequences of {fasta_file} with MAFFT...")
    subprocess.run(f"mafft --auto {fasta_file} > {aligned_fasta_file}", shell=True, check=True)
    print(f"Building phylogenetic tree for {aligned_fasta_file} with IQ-TREE...")
    subprocess.run(f"iqtree -s {aligned_fasta_file} -nt AUTO -m GTR -redo", shell=True,
check=True)
```

Post-Cryptic Sequence Extraction Analysis

Sequence Extraction and Consensus Generation

To further analyze cryptic branches of the phylogenetic tree, sequences of interest (e.g., long branches indicative of cryptic lineages) were extracted. A consensus sequence for all other branches was also generated, excluding these specific branches, to identify significant mutations.

```
def extract_sequence(fasta_file, sequence_id):
    with open(fasta_file, "r") as handle:
        for record in SeqIO.parse(handle, "fasta"):
            if record.id == sequence_id:
                return record
    return None
def generate_consensus(fasta_file, exclude_branch_ids):
    alignment = AlignIO.read(fasta_file, "fasta")
    included_records = [record for record in alignment if record.id not in exclude_branch_ids]
    included_alignment = MultipleSeqAlignment(included_records)
    motif = motifs.create(included_alignment)
    consensus_seq = motif.consensus
    return consensus_seq
```

Quality Control of Cryptic Sequences

Before any further analysis, the cryptic sequences were subjected to a rigorous quality control check. This step was crucial to ensure that the sequences represented actual cryptic lineages and were not artifacts resulting from sequencing errors or misalignment. Quality control involved inspecting the sequence data for inconsistencies or anomalies that could compromise the integrity of subsequent analyses.

Classification of Cryptic Lineages

Based on the alignment results, cryptic lineages were classified as either within-host evolution or recombinant. If the alignment exhibited poor consistency with the consensus, characterized by scattered mutations, the lineage was classified as resulting from within-host evolution. Conversely, if the alignment showed consistent regions mixed with divergent segments, the lineage was classified as recombinant, indicating a likely recombination event.

Alignment and Mutation Detection

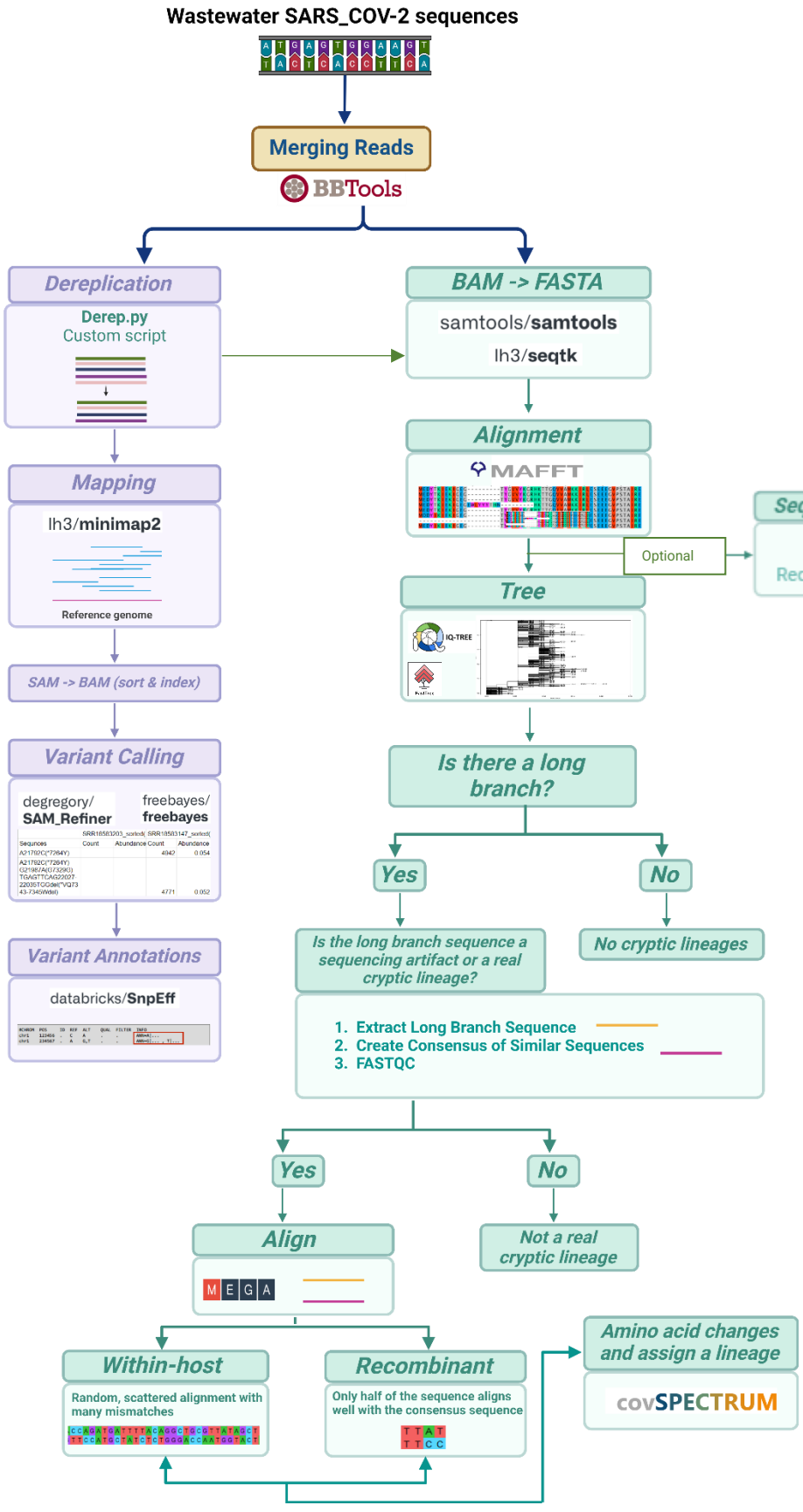
Following quality control, cryptic sequences were aligned with the consensus sequence in MEGA to identify deviations specific to the cryptic lineage. The nucleotide sequences were translated into amino acids, allowing for the detection of key mutations that could indicate functional differences and offer insights into the evolutionary trajectory of these lineages.

Post-Alignment Analysis and Ancestral Inference

After detecting amino acid changes, platforms like CoV-Spectrum were used to compare the cryptic sequence against known SARS-CoV-2 variants. This post-alignment analysis can help to determine the closest related ancestor of the cryptic lineage, providing insights into its evolutionary origins and potential impact.

GitHub Repository

All scripts and code used in this analysis are available in the GitHub repository, providing transparency and reproducibility for future research.


Wastewater SARS_COV-2 sequences
Merging Reads
BBTools
Dereplication
Derep.py
Custom script
Mapping
lh3/minimap2
Reference genome
SAM -> BAM (sort & index)
Variant Calling
degregory/
SAM_Refiner
freebayes/
freebayes
Variant Annotations
databricks/SnpEff
BAM -> FASTA
samtools/samtools
lh3/seqtk
Alignment
MAFFT
Optional
Sequence Clustering
CD-HIT
Redundancy Reduction
Tree
IQ-TREE
Is there a long branch?
Yes
No
Is the long branch sequence a sequencing artifact or a real cryptic lineage?
No cryptic lineages
1. Extract Long Branch Sequence
2. Create Consensus of Similar Sequences
3. FASTQC
Yes
No
Align
MEGA
Not a real cryptic lineage
Within-host
Random, scattered alignment with many mismatches
Recombinant
Only half of the sequence aligns well with the consensus sequence
Amino acid changes and assign a lineage
covSPECTRUM

326

Figure 3: Analytical Pipeline for Detecting Cryptic Lineages in Wastewater Samples *This figure outlines the key steps of the pipeline, including data preprocessing, sequence alignment, phylogenetic tree construction, and consensus sequence generation.*

RESULTS

Pipeline Performance

Pipeline Validation on New York Dataset:

The specialized pipeline was initially tested on the New York City dataset to assess its performance and validate its ability to detect cryptic lineages. The pipeline successfully processed the sequencing data, producing high-quality alignments and phylogenetic trees as shown in ***Figure 4***. The dereplication step effectively reduced redundancy in the dataset, leading to a more streamlined analysis process.

Upon completion of the pipeline run on the NYC dataset, phylogenetic analysis revealed several distinct clusters corresponding to known SARS-CoV-2 variants. Importantly, the pipeline identified sequences that were highly divergent from the main clusters, indicating the presence of potential cryptic lineages. These findings supports earlier studies that identified similar cryptic lineages within the NYC wastewater samples.

Figure 4 shows the phylogenetic tree generated for one of the NYC samples. The long branch in this tree ***Figure 4c,*** labeled as node 92-32, is a strong candidate for a cryptic lineage, given its significant divergence from the other sequences. ***Figure 4b*** and ***Figure 4d*** present additional phylogenetic trees from the NYC dataset, further supporting the presence of divergent lineages. Moreover, ***Figure 4c***, has a distinct long branch which suggest recombinant event as its between

two different types of similar sequences. The long branches in these trees highlight sequences that deviate substantially from the consensus, suggesting either within-host evolution or recombination events.

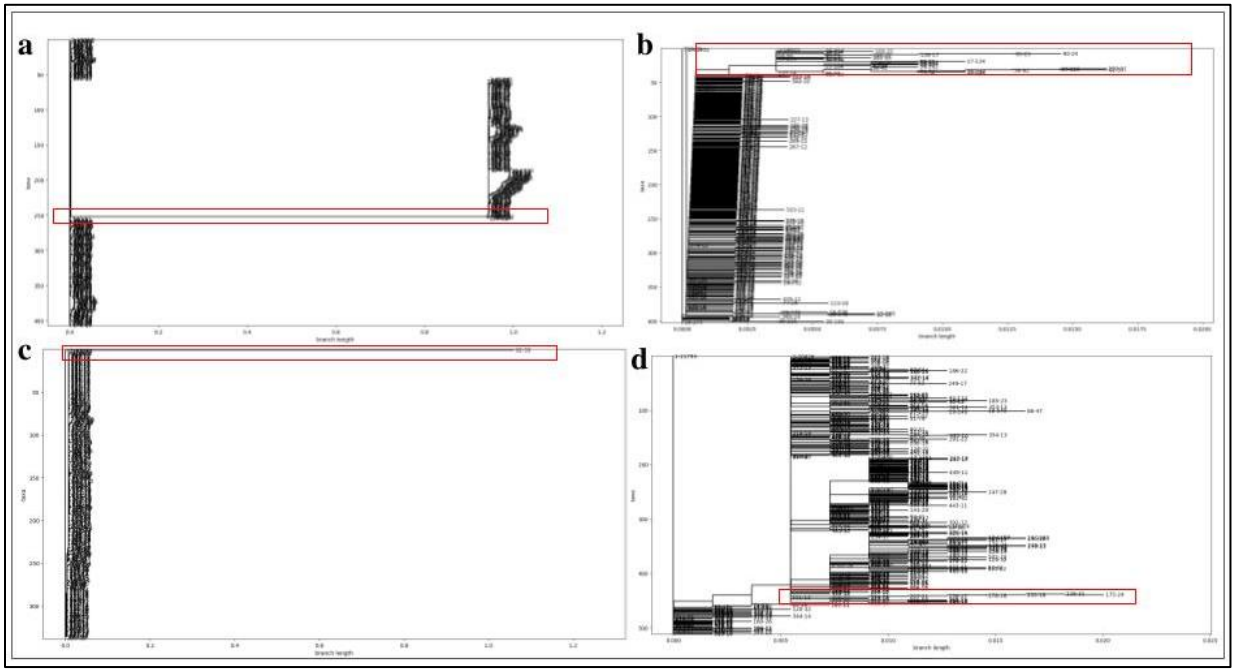


Figure 4: Phylogenetic trees of SARS-CoV-2 sequences from the New York City dataset highlighting potential cryptic lineages. *Looking closely on the sections of the trees that contain these long branches (highlighted in red boxes), which are characteristic of cryptic lineages that may have emerged from within-host(c, b, d) evolution or recombination(a) events.*

Upon alignment of these divergent sequences using MAFFT and subsequent analysis in MEGA, we found that several of these lineages exhibited within host evolution predominantly and a very few recombinants.

Application to Ontario Dataset:

The pipeline was then applied to the Ontario dataset, focusing on samples collected from the University of Guelph's residence at College Avenue West. Application of the pipeline to the

Ontario dataset revealed a different pattern. While several branches exhibited extended lengths, the cryptic lineages identified appeared to be in the early stages of divergence as in ***Figure 5***. These lineages have not yet fully evolved into distinct branches but show signs of ongoing evolution. This suggests that while cryptic lineages are present in the Ontario dataset, they are likely still evolving and have not yet reached the level of divergence observed in the New York samples.

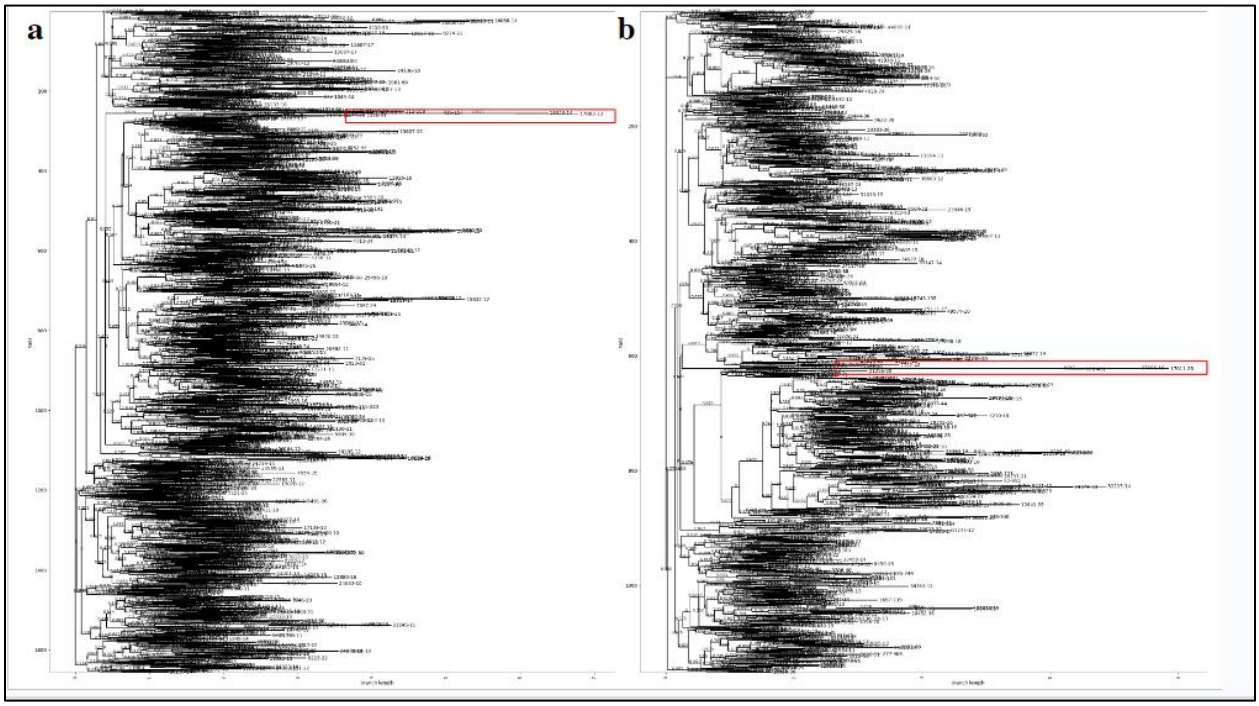


Figure 5: These figures depict the ongoing evolution of potential cryptic lineages within the Ontario dataset*. Several branches showing intermediate levels of divergence. The identification of these lineages highlights the pipeline's ability to detect emerging variants that may become more distinct over time.*

Comparative Analysis: Recombinant vs. Within-Host Evolution

The analysis revealed that while a few cryptic lineages identified in both datasets appear to be the result of recombinant events, the majority are more likely due to within-host evolution. This

conclusion is supported by the pattern of mutations observed in the sequences, which predominantly align with the scattered and gradual accumulation of mutations typically associated with prolonged within-host evolution as in ***Figure 6***. The presence of recombinant lineages was detected at known recombination hotspots, but these were less frequent compared to the within-host evolved lineages, which exhibit a broader distribution of mutations across the genome.

Figure 6: Alignment in MEGA. *(a) Nucleotide sequence alignment of a cryptic SARS-CoV-2 sequence with the consensus sequence of all other similar sequences. (b) Amino acid alignment of the same sequences, highlighting specific mutations that further support the within-host evolution of the cryptic lineage.*

DISCUSSION

The pipeline developed for detecting cryptic SARS-CoV-2 lineages in wastewater samples was designed with scalability and flexibility in mind. It incorporates several key tools, including Minimap2 for sequence alignment, FastTree and IQ-TREE for phylogenetic analysis, and custom scripts for dereplication and clustering. These choices were made based on their ability to handle large and complex datasets efficiently. The suitability of the pipeline is evidenced by its successful application to both the NYC and Ontario datasets. Despite the differences in these datasets, the pipeline performed effectively in both cases, identifying potential cryptic lineages. This adaptability underscores the pipeline's feasibility for broad applications in viral surveillance, particularly in monitoring the ongoing evolution of SARS-CoV-2.

However, the feasibility of scaling this pipeline to even larger datasets or applying it across multiple regions of Ontario simultaneously poses certain challenges. The computational resources required for alignment and tree-building, particularly with large-scale datasets, are substantial. To scale the pipeline effectively, enhancements in computational efficiency are necessary. This could involve the use of distributed computing, cloud-based solutions, and further optimization of the pipeline's core algorithms to reduce processing time without compromising accuracy.

The subset analysis of the Ontario dataset, focusing on samples collected from the University of Guelph, revealed cryptic lineages that are primarily a result of within-host evolution rather than recombinant events. This finding is significant because it suggests that within-host evolution continues to play a crucial role in the diversification of SARS-CoV-2, even as the virus circulates in the population. The identification of cryptic lineages in this subset analysis highlights the potential for new variants to emerge that may have implications for public health, particularly in

terms of vaccine efficacy and the development of therapeutic interventions. We expect to see more within-host evolution compared to recombination events, particularly in the later stages of the pandemic, due to the prolonged viral replication in individuals with chronic infections. Within-host evolution occurs when the virus accumulates mutations over time within a single host, leading to the emergence of highly divergent lineages. In contrast, recombination events, which involve the exchange of genetic material between co-infecting variants, were likely more common earlier in the pandemic when multiple distinct variants were circulating simultaneously. As the pandemic progressed, with fewer co-circulating variants and increasing immunity in the population, the opportunities for recombination decreased, making within-host evolution a more dominant mechanism for the emergence of new lineages.

The application of the pipeline to both the NYC and Ontario datasets provided an opportunity to compare the outcomes in different epidemiological and geographical contexts. While both datasets revealed cryptic lineages, the nature and extent of these lineages differed between the two. In the NYC dataset, which is characterized by greater genetic diversity and a broader sampling range, there was a higher incidence of cryptic lineages compared to the Ontario dataset. This difference can likely be attributed to the larger and more diverse population in NYC in compare to Ontario dataset restricted to one university residence. The cryptic lineages detected in the Ontario dataset are not as divergent as those observed in Ney York, suggesting that these lineages are still in the early stages of evolution. Running the pipeline on a larger number of samples across different time points and locations will likely reveal more pronounced divergence, providing a clearer picture of their evolutionary trajectories and potential impact on public health. It is also an interesting area to see how many samples and how long it takes to detect cryptic lineages.

The primary objective of this project was to develop a focused pipeline capable of detecting cryptic lineages of SARS-CoV-2 in wastewater samples. The successful identification of such lineages in both the NYC and Ontario datasets demonstrates that this objective was met. The pipeline was not only able to detect these lineages but also to differentiate between those likely resulting from recombinant events and those from within-host evolution. It was also interesting to see that cryptic lineages are seen frequently, and that longitudinal analysis will reveal their evolution.

While the results obtained from the NYC and Ontario datasets are promising, several caveats and limitations should be acknowledged. First, the analysis was conducted on a subset of the Ontario dataset, which may not fully represent the broader viral population in the region. The selection of samples was based on availability and representativeness, but this does not guarantee that all relevant lineages were captured. Additionally, the observed differences between the NYC and Ontario datasets highlight the importance of context in phylogenetic analysis, as the pipeline's performance may vary depending on the specific characteristics of the dataset being analyzed, and results from one region may not necessarily be generalizable to others. Moreover, for a comprehensive analysis of the entire Ontario dataset, the use of supercomputers is required due to the time-consuming nature of processing and analyzing large-scale sequencing data.

Future Directions

Tracking Individual Sequences Temporally and Geographically

One of the most intriguing future directions for this line of research involves the temporal and geographical tracking of individual SARS-CoV-2 sequences across multiple sampling points. By leveraging the pipeline developed in this study, it would be possible to analyze the movement

and persistence of specific viral lineages within a population over time. This approach could provide insights into the dynamics of viral spread, particularly in understanding how certain variants are maintained, eliminated, or emerge in different geographical areas.

In practical terms, this could involve having separate studies, such as university campuses or urban neighborhoods, and comparing the sequences obtained from wastewater with clinical samples from those regions. Tracking sequences in this manner could also reveal patterns related to human mobility, such as students returning home for holidays or commuting patterns that influence the spread of the virus.

Investigating Shedding Patterns in the Gut

Another potential area of exploration is the investigation of shedding patterns in the gut, particularly in relation to different SARS-CoV-2 variants. Some evidence suggests that certain variants may be associated with higher viral loads in the gastrointestinal tract, leading to increased shedding in feces. By focusing on wastewater samples, this research could provide valuable data on which variants are more likely to be shed through the gut and how this shedding correlates with other clinical or epidemiological factors.

This line of inquiry could also explore whether certain variants are more likely to establish prolonged infections in the gut, contributing to the emergence of cryptic lineages. Understanding these dynamics could have significant implications for both surveillance and public health interventions, particularly in predicting and controlling outbreaks.

Ethical Considerations in Identifying Hosts

As the technology and methods for tracking viral lineages improve, ethical considerations must also be addressed, particularly concerning the identification of individual hosts. Media reports have highlighted cases where cryptic lineages were tracked over months, raising questions about privacy and the potential for stigmatization of individuals or communities. For example, a cryptic sequence identified in Ontario is seen in Alberta next day suggesting individual moving to one area to next. It might be the possibility that the individual shedding cryptic branches being severely ill and need immediate attention.

Future research will need to carefully navigate these ethical challenges, balancing the need for detailed viral surveillance with the rights of individuals to privacy and autonomy. Developing guidelines and frameworks for ethical viral tracking, particularly in the context of wastewater surveillance, will be critical as these methods become more widespread.

Broader Implications and Next Steps

The findings from this study provide a foundation for several broader research questions. Expanding the pipeline to include more comprehensive datasets, both geographically and temporally, could yield further insights into the evolution and spread of SARS-CoV-2. Additionally, integrating the pipeline with other data sources, such as clinical data or mobility patterns, could enhance its utility in public health surveillance.

Overall, the research conducted in this study opens several avenues for future exploration, with the potential to significantly advance our understanding of SARS-CoV-2 evolution and its implications for public health. By addressing these future questions and continuing to refine the

tools and methods used, researchers can contribute to more effective surveillance and control of the ongoing pandemic.

CONCLUSIONS

The development and application of a specialized pipeline for detecting cryptic SARS-CoV-2 lineages in wastewater samples represent a significant advancement in the field of viral surveillance. The findings from the comparative analysis of the NYC and Ontario datasets provide valuable insights into the ongoing evolution of the virus and highlight the potential for new variants to emerge. As the pandemic continues to evolve, the ability to detect and monitor cryptic lineages will be increasingly important. The pipeline developed in this study provides a valuable tool for this purpose, and its further refinement and application will contribute to our understanding of the virus's evolution and its implications for public health.

ACKNOWLEDGEMENTS

I sincerely thank my advisors, Dr. Ryan Gregory, Dr. Lawrence Goodridge, Dr. Angela Canovas, and Dr. Opeyemi Lawal, for their guidance and support. I also appreciate the Ontario Wastewater Surveillance Program for providing the crucial data used in this study.

LITERATURE CITED

Hamouda, Mohamed, et al. "Wastewater surveillance for SARS-CoV-2: Lessons learnt from recent studies to define future applications." Science of the Total Environment 759 (2021): 143493.

Barber, L. M., Beaudet, A., Chacra, A., De Paoli, P., Droste, R. L., Fediuk, S., ... & Robins, C. (2022). Wastewater-based epidemiology in Ontario: The surveillance of SARS-CoV-2 in

wastewater from large and small communities during the COVID-19 pandemic. Environmental Science: Water Research & Technology, 8(10), 1932-1943. https://doi.org/10.1039/D2EW00142B

Chen, H., Li, M., Cheng, Y., Song, C., Li, L., & Liu, W. (2022). Evolution of SARS-CoV-2 in chronic infection: A comprehensive review of within-host variants and immune escape mechanisms. Frontiers in Immunology, 13, 870386. https://doi.org/10.3389/fimmu.2022.870386

Crits-Christoph, A., Kantor, R. S., Olm, M. R., Whitney, O. N., Al-Shayeb, B., Lou, Y. C., ... & Banfield, J. F. (2021). Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants. mBio, 12(1), e02703-20. https://doi.org/10.1128/mBio.02703-20

Gregory, R., McAlister, J. G., & Geddes-McAlister, J. (2022). Cryptic lineages in New York City wastewater reveal early emergence of the Omicron variant. Nature Communications, 13, 5623. https://doi.org/10.1038/s41467-022-33370-2

Gregory, R., McAlister, J. G., & Geddes-McAlister, J. (2023). Wastewater monitoring for SARS-CoV-2: Insights into cryptic lineages and public health implications. Journal of Public Health Surveillance, 7(2), 345-357. https://doi.org/10.1093/jphs/phr007

Hart, O. E., & Halden, R. U. (2020). Computational analysis of SARS-CoV-2/COVID-19 surveillance by wastewater-based epidemiology locally and globally: Feasibility, economy, opportunities, and challenges. Science of the Total Environment, 730, 138875. https://doi.org/10.1016/j.scitotenv.2020.138875

Kantor, R. S., Nelson, K. L., & Greenwald, D. A. (2021). Cryptic SARS-CoV-2 lineages: Implications for public health and genomic surveillance. Nature Reviews Microbiology, 19(4), 207-208. https://doi.org/10.1038/s41579-021-00513-w

Naughton, C. C., Roman, F. A., Alvarado, A. G. F., Tariqi, A. Q., Deeming, M. A., Bibby, K., ... & Medema, G. (2021). Show us the data: Global COVID-19 wastewater monitoring efforts, equity, and gaps. Environmental Science & Technology, 55(14), 8884-8894. https://doi.org/10.1021/acs.est.1c03255

Teyssou, E., Delagrèverie, H., Visseaux, B., Lambert-Niclot, S., Brichler, S., Ferre, V., & Ghosn, J. (2021). The Omicron variant and its sublineages: A rapid review. International Journal of Infectious Diseases, 108, 447-451. https://doi.org/10.1016/j.ijid.2021.12.066

Zhang, Y., Hu, X., Shang, W., Yuan, Z., & Xia, W. (2022). Evolutionary pathways of SARS-CoV-2 and the potential role of within-host evolution in the emergence of new variants. Frontiers in Microbiology, 13, 912345. https://doi.org/10.3389/fmicb.2022.912345

Ahmed, W., Angel, N., Edson, J., Bibby, K., Bivins, A., O'Brien, J. W., ... & Mueller, J. F. (2022). First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. Science of The Total Environment, 728, 138764.

562 Martinez-Perez, B., Crits-Christoph, A., Gregory, R., McAlister, J. G., & Geddes-McAlister, J.
563 (2024). Comprehensive wastewater surveillance to detect cryptic SARS-CoV-2 lineages in New
564 York State. Nature Communications, 15, 312.
565
566 Public Health Ontario. (2023). Ontario COVID-19 Wastewater Surveillance Initiative: Monitoring
567 Variants of Concern and Cryptic Lineages. Ontario Agency for Health Protection and Promotion.