Blah blah blah title

Leah Lariscy

# Abstract

SARS-CoV-2 wastewater-based detection methods are used as indicators of COVID-19 prevalence and can serve as a tool in understanding transmission dynamics as infections are less often reported now than initially during the pandemic. While wastewater-based surveillance can be less resource demanding and time consuming than clinical surveillance, more work could be done to reduce reporting delays. Previous work has been done to explore the most robust and accurate wastewater-based models for predicting incidence of COVID-19, but work has not yet been done to understand how simplified processing methods could impact model performance. This study aims to utilize a machine learning framework to compare RT-qPCR-based wastewater metrics, Viral Load and Detection Frequency, for their power to predict COVID-19 incidence, as well as evaluate the impact of simplified processing methods on model performance. Composite samples were collected between June 2020 and December 2022 from three treatment facilities in Athens-Clarke County, Georgia (USA). Six replicate nucleic acid extractions were performed and SARS-CoV-2 specific RT-qPCR assays were used to amplify targets on the N-gene (N1 and N2). Quantification cycle (Cq) values and wastewater flow data were used to estimate daily viral loads. Assay detection frequency was determined by totaling the positive reactions for each date. COVID-19 clinical data were obtained from the Georgia DPH. Using a Random Forest framework, viral load and detection frequency were compared for their ability to accurately predict clinical test positivity rates from June 2020 to January 2021. Subsequently, extraction replicates were removed one at a time and model performances were assessed.

(Insert results and conclusion)

# Introduction

## Background

Early in the pandemic, wastewater-based epidemiology (WBE) was identified as a promising method for detection and quantification of SARS-CoV-2 at population levels. SARS-CoV-2 viral particles can shed from a person during and after infection (Wu et al., 2020) and can enter wastewater streams when an infected individual uses the toilet, showers, or washes their hands. Viral particles in wastewater are able to be detected via Reverse Transcriptase Quantitative Polymerase Chain Reaction (RT-qPCR) by amplification of universally-conserved gene targets (Ahmed et al., 2020). Previous studies have shown that viral load quantification and detection frequency in wastewater samples correlate with COVID-19 clinical case data, especially when reporting efforts are strong (cite). As COVID-19 shifts from pandemic to endemic and reporting is no longer robust, WBE methods have the potential to fill gaps left by clinical data.

Wastewater-based surveillance for public health offers a number of advantages relative to clinical surveillance. To get clinical data, an infected individual must have access to and decide to be tested in a clinical setting and data sharing must be approved by an institutional review board. Alternatively, WBE is unbiased by healthcare seeking behavior and is anonymous by nature, allowing for more infections to be captured with less concerns over privacy. WBE can also be less resource demanding as one wastewater sample can potentially capture numerous infections at once, whereas a greater magnitude of clinical sampling is needed to capture comparable information. Limited sampling necessity with WBE combined with real-time detection methods could make informing public health faster and more accessible.

While WBE does offer considerable advantages to traditional surveillance, it is important to look for ways to improve access and turnaround time by evaluating trade offs between data robustness and methodology simplicity. RT-qPCR is the most common method used to detect and quantify SARS-CoV-2 particles in wastewater, but this requires access to specialized equipment that may not be available to all research groups. Meanwhile, conventional PCR methods may be more widely accessible, but data is limited to observance of presence/absence because quantification is not possible. While conventional methods may lack the robustness that RT-qPCR offers, little work has actually been done to examine the relationship between presence/absence data and infection incidence (cite). Most groups conducting wastewater surveillance also utilize multiple replicates per sample to obtain more accurate results, however, this can drastically increase the processing time. Previous work suggests that replicate samples can minimize errors associated with viral wastewater detection (cite), but work has yet to examine how reducing replicates impacts case incidence predictions. Understanding these trade offs will be necessary for expanding global engagement with wastewater-based surveillance methods, not only for COVID-19 but other concerning pathogens as well.

## Aims

This work aims to understand the trade off between wastewater surveillance robustness and predictive power through a series of feature selections, evaluated within a Random Forest (RF) framework. To do this, RF models will be trained on either a) viral load quantification or b) detection frequency of SARS-CoV-2 particles in wastewater. Subsequently for each feature, replicates will be dropped one at a time at random (n = 6..1). Each feature selection will be used in a RF model to predict clinical case positivity rates (positive tests/total test administered) and model performance metrics will be compared.

# Materials and Methods

# Results

# Discussion

# Conclusion

## Sources cited