A rigorous benchmarking of methods for SARS-CoV-2 lineage detection in wastewater.

Applicant: Bohdan Tyshchenko
jityshchenko@gmail.com

Mentor: Serghei Mangul

1 Introduction

In response to the devastating COVID-19 pandemic, new bioinformatics methods have been swiftly developed and as a result adequate benchmarking is required to be provided for measuring performance of novel methods. One of the most promising fields is analyzing COVID-19 viral genomes in wastewater studies because of wastewater monitoring advantages.

The most convenient and reliable point of current and future monitoring for circulating viruses is wastewater facilities. It is already proven that they are capable of monitoring population viral prevalence with equal quality to the quality of clinical monitoring. The qPSR analysis already serves for monitoring overall viral COVID-19 prevalence in the communities. We propose to extend this successful experience by adding to the current methodology efficient sequencing and computational analytics. That will allow not only monitor the prevalence of the virus, but also monitor the prevalence of all novel and appearing strains. Furthermore, this experience can be extended to other viral diseases, and even more coupled with metagenomic analysis, it can help monitoring of all genomic data from all viruses and pathogens located in each wastewater sample. [1][2][3]

A number of strain quantifying methods have recently been developed. [4][5] To test these tools we will create a novel benchmark based on modern genome engineering technologies, that will allow us to create samples that will be identical to real samples but with known ground truth.

2 Objectives

2.1 Plan

1) Take 5 different viral strains from Gisaid database, for example, Omicron, Alpha, Beta, Gamma, Delta.
2) Create an in-silico mixture of equal abundance.
3) Generate in-silico Illumina reads using SimSeq.
4) Run CliqueSNV, PredictHaplo, aBayesQR tools.
5) Measure precision and recall of predictions.

2.2 (Additionally) Plan maximum

Run Savage, Shiver, Vicuna, RegressHaplo, PEHaplo tools.
Generate in-silico nanopore reads using NanoSeq; run all tools on this benchmark and measure their performance.

3 Timeline

Weeks 1-2

Getting used to working on the USC cluster. Write Bash scripts to create a benchmark dataset using sequencing read simulation software.

Weeks 3-8

Figuring out how to run every tool on the read data. Here, we will implement snakemake-based pipeline for tools benchmarking on a cluster.

Weeks 9-10

Adding tool performance evaluation and making visualizations.

4 Deliverables

- Bash scripts for creating the Sars-Cov-2 benchmarking dataset.
- Bash scripts that run tools on read benchmarks measuring precision and recall.
- Visualization tables and graphs of tools' performance and Python jupyter notebooks to create them.

5 References

1. Jasmin, A.B., Alessandro, Z., Isabel, O.Z., et al. Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-seq quantification. (2021). https://doi.org/10.1101/2021.08.31.21262938

2. Karthikeyan, Smruthi Levy, Joshua Hoff. et al. Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant transmission. (2021). https://doi.org/10.1101/2021.12.21.21268143.

3. Smyth, D.S., Trujillo, M., Gregory, D.A. et al. Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater. Nat Commun 13, 635 (2022). https://doi.org/10.1038/s41467-022-28246-3

4. Knyazev, S., Hughes, L., Skums, P., Zelikovsky, A. (2020). Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. Briefings in Bioinformatics. https://doi.org/10.1093/bib/bbaa101

5. Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction Nucleic Acids Research (2021), Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction, https://doi.org/10.1093/nar/gkab576