

The impact of sequencing errors and contaminating viruses on SARS-CoV-2 variant detection in wastewater

Mart van der Lugt - m.j.vanderlugt@student.tudelft.nl — SUPERVISOR: J. A. Baaijens - j.a.baaijens@tudelft.nl

Substitution errors

Deletion errors

Insertion errors

Fig. 1 - Relative prediction error plotted against the induced error frequency for substitution, deletion and insertion errors. Plotted at a VoC frequency of 10.8%.

Research

How does the frequency of **sequencing errors** affect **prediction accuracy** of the pipeline[1]?

- Datasets with different error frequencies.
 - Substitution errors
 - Deletion errors
 - Insertion errors
 - Chimeric reads

How does the amount of **contaminating viruses** affect **prediction accuracy** of the pipeline[1]?

- Datasets with different levels of contaminants.
 - SARS-CoV-1
 - MERS-CoV
 - 4 common human coronaviruses
 - 15 other common wastewater-borne viruses [2]

Results

- Substitution errors > insertion and deletion errors. (Fig. 1)
- **Correlation** between impact of sequencing errors and prevalence of corresponding mutations. [3]
- Only **SARS-CoV-1** and **hCoV-HKU1** as contaminants have impact. (Fig. 3a)
- Adding contaminants to **reference set** removes effect. (Fig. 3c)
- Great performance of B.1.1.7. (Fig. 3b)

Discussion

- Impact is **acceptable** for error rates of most modern sequencing machines.
- There could be a link between the **mutational spectrum** of a virus and the impact of various sequencing errors.
- Only a few viruses impact performance. These viruses are **closely related** to SARS-CoV-2.
- Adding these viruses to the **reference set** removes any effect.
- Great performance of B.1.1.7 is due to under- and overestimation in different datasets. (Fig. 2)

SARS-CoV-1

hCoV-HKU1

Fig. 2 - Estimated VoC frequency plotted against SARS-CoV-2 frequency in the entire dataset. Plotted at a VoC frequency of 10%.

(a) Seperate contaminants

(b) All hCoV-contaminants

(c) Contaminants in reference set

Fig. 3 - Relative prediction error plotted against SARS-CoV-2 frequency in the entire dataset. The rest of the dataset is filled with contaminants. Plotted at a VoC frequency of 10%.

[1] Jasmijn A Baaijens et al. Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-Seq quantification. medRxiv, page 2021.08.31.21262938, 12021.
[2] Mary Vermi Aizza Corpuz et al. Viruses in wastewater: occurrence, abundance and detection methods. Science ofThe Total Environment, 745:140910, 11 2020.
[3] Kijong Yi et al. Mutational spectrum of SARS-CoV-2 during the global pandemic. Experimental & Molecular Medicine, 53(8):1229–1237, 8 2021.