Kallisto repurposed

Focusing on genomic regions of SARS-CoV-2 to better predict variant abundances in wastewater

Anton Matei
E-mail: M.Anton@student.tudelft.nl
Supervisor: Jasmijn Baaijens

Background

- We expel remnants of the virus when we get infected
- This means we can detect genomic fragments when we analyze wastewater samples
- Kallisto can be used to estimate variant abundances from wastewater samples
- **My project:** Is it better to look at specific regions of the genome rather than sequencing it in its entirety in order to improve prediction accuracy of kallisto? If so, which ones and in what combinations?

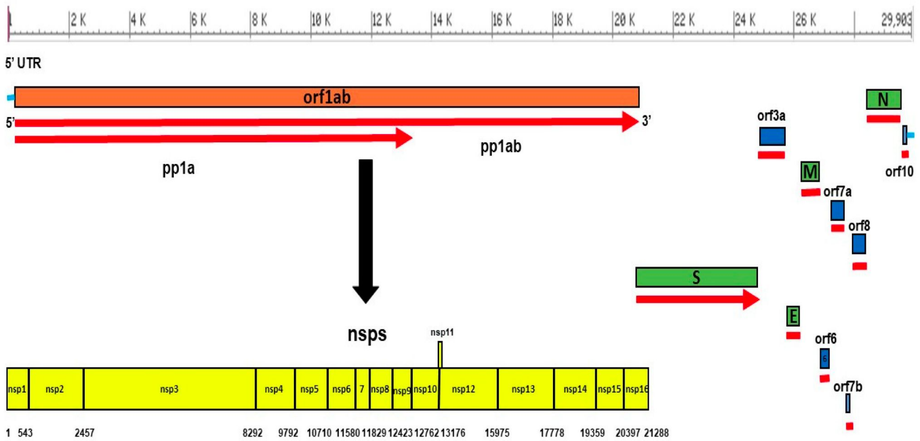

Fig. 1 *genome organization of Sars-COV-2. Each region codes for a different protein that goes into the structure of the virus* ***Source****: (Yosra et. al, 2020)*

Results

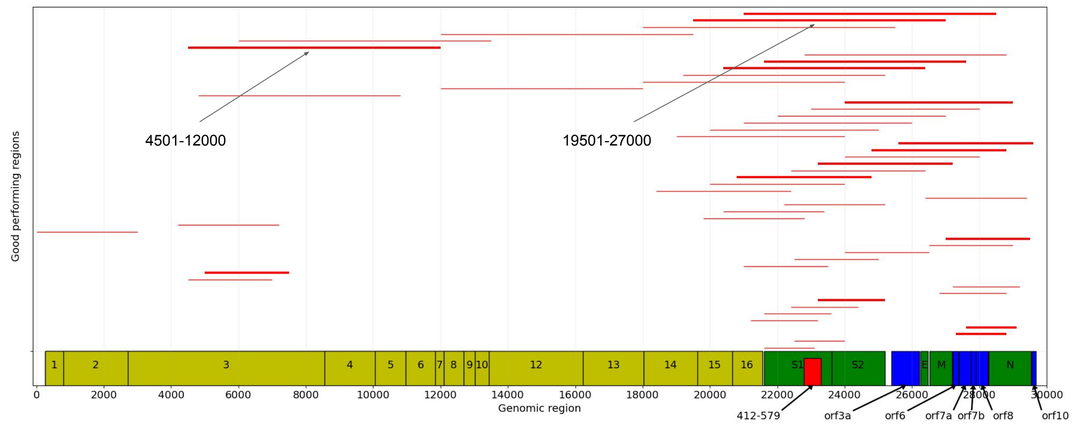

Fig. 2 *Best performing regions from sequencing and analyzing various length regions throughout the genome. These are those that have 20 or under relative prediction error for 20 and more simulated abundances. The regions shown in bold are even better as they have this prediction error also for simulated abundance of 10*

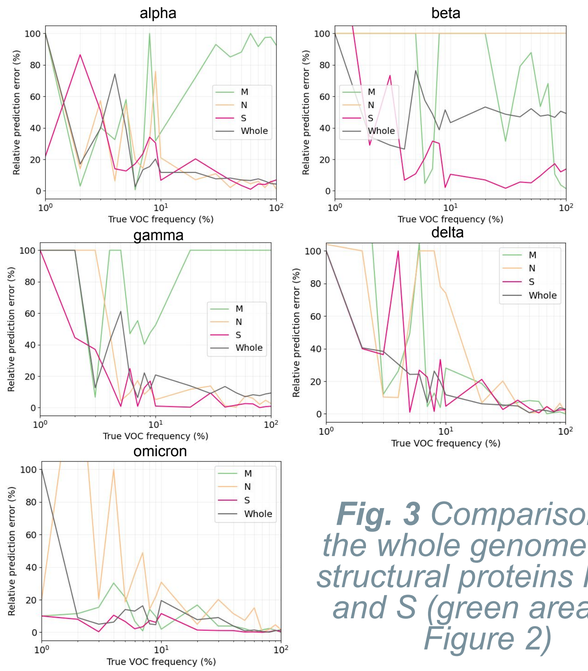

Fig. 3 *Comparison of the whole genome with structural proteins M, N, and S (green areas in Figure 2)*

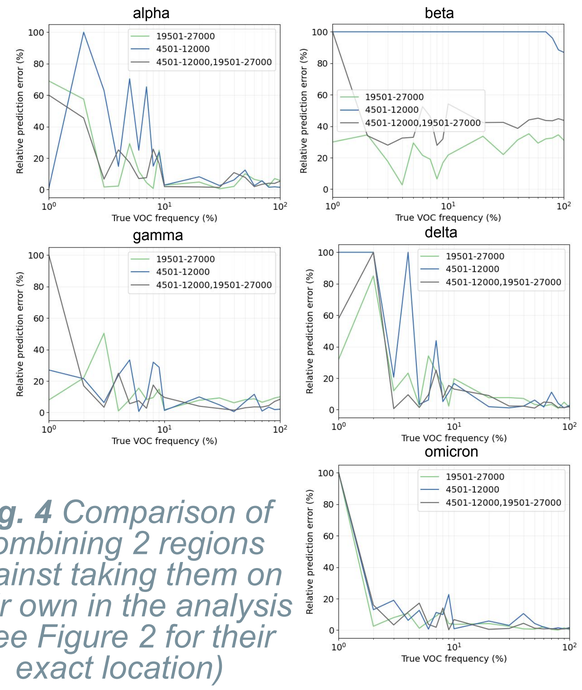

Fig. 4 *Comparison of combining 2 regions against taking them on their own in the analysis (see Figure 2 for their exact location)*

Main Takeaways

- Good regions for good predictions: N, S, and an area around the middle of nsp3
- Combining 2 good performing regions seems to confer a slight advantage at lower abundances
- Beta variant is uniquely hard to predict accurately