Kallisto Repurposed

Using sequencing reads from the spike, nucleocapsid, and a middle region of nsp3 in the kallisto pipeline to better predict SARS-CoV-2 variants in wastewater

by

Matei Anton

Supervisor: Jasmijn A. Baaijens

A Dissertation

Submitted to EEMCS faculty

Delft University of Technology,

In Partial Fulfilment of the Requirements

For the Bachelor of Computer Science and Engineering

January 23, 2022

Kallisto Repurposed

Using sequencing reads from the spike, nucleocapsid, and a middle region of nsp3 in the kallisto pipeline to better predict SARS-CoV-2 variants in wastewater

Matei Anton[*]
Supervisor: Jasmijn A. Baaijens[†]
EEMCS, Delft University of Technology, The Netherlands

January 23, 2022

Abstract

During a viral infection, we expel remnants of the virus. This makes it possible to conduct wastewater analysis which aid in the efforts to track the evolution of the current Covid-19 pandemic. It has been shown that by repurposing the kallisto algorithm, the abundance of SARS-CoV-2 variants in wastewater samples can be estimated. Since this is a novel method for this scope, its precision could probably be improved by adjusting certain aspects. In this work, I look at one of those aspects: sequencing particular genomic regions of the virus rather than the entire genome. I have indeed found that the regions that code for the spike (S) and nucleocapsid (N) regions and a section around the region coding for the non-structural protein 3 (nsp3) give particularly accurate results when sequenced on their own. In addition, in at least one case, combining two well-performing regions further improves accuracy at lower simulated abundances of variants. This suggests that sequencing depth is preferred over sequencing breath as long as the region being sequenced contains enough information to distinguish between variants. These findings are important as they can aid in the improvement of this method of variant quantification. Moreover, they can also help in improving other algorithms applied to the SARS-CoV-2 genome by highlighting the genomic sections containing the most differentiating information between variants.

1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the virus responsible for the Covid-19 disease which was first detected in Wuhan, China in December 2019. Since then, it was declared a Public Health Emergency of International Concern on 30 January 2020 by the World Health Organization and a pandemic since 11 March 2020 which is ongoing at the time of writing.

In the effort to counter this pandemic, effective monitoring is essential. To this end, the European Commission has put forth a number of recommendations [1]. Despite this, the recommendation to sequence 5-10% of the SARS-CoV-2 positive patient cases was not met

[*] M.Anton@student.tudelft.nl
[†] J.A.Baaijens@tudelft.nl@tudelft.nl

by most Member States [2]. As such, alternative ways of monitoring the virus is essential in tracking the spread and evolution of the pandemic.

Wastewater based epidemiology (WBE) can aid in this as it can be done when clinical sequencing, a laboratory method for determining the genetic makeup of an organism, proves unavailable or impractical [3]. This is possible as genomic fragments of SARS-CoV-2 are expelled from the human body through urine and faeces during an infection. WBE has the added benefits of not being subject to patient sequencing biases and of revealing the spatial spread of mutations at community and city level [4].

In [3], a method for quantifying the abundance of SARS-CoV-2 variants from wastewater samples has been proposed. For this, kallisto [5] is employed, a tool initially used for RNA sequencing (RNA-Seq) quantification. Since the genome is divided into sections that code for different proteins, the building blocks of life, they also thought to only sequence the genomic region coding for the spike protein, the one the virus uses to attach to our cells. In this analysis, they observed greater prediction accuracy. This last fact also motivates the main question of the work presented in this paper: Which genomic regions should be sequenced in order to maximize the prediction accuracy for the method proposed in [3]? To answer this, I first look at how genomic regions coding for specific proteins compare. Since those regions vary in length, equal length regions within the genome are also analyzed. Finally, I check if prediction accuracy is further improved if combinations of the best-performing regions from previous experiments are used in the pipeline.

Adaptive evolution of SARS-CoV-2 mainly affects sites on the already discussed spike protein and the nucleocapsid protein regions [6]. The role of the nucleocapsid protein is to bind to the RNA sequence of the virus, which holds the genomic information, and to enter the host cell. As such, the expectation before performing the experiments was that regions which include those regions are also especially fit[1] for distinguishing between virus variants when used in the algorithm. This is also partially confirmed in [3] as they found that the region coding for the spike protein is more fit in the algorithm compared to sequencing the entire genome.

In my analysis, I also find the spike region to be particularly good at predicting the correct abundances of variants. In addition, this was also found for the nucleocapsid region and an area around the middle of the nsp3 region. When good performing regions are combined, the results seem to indicate that combinations of two confer a slightly better prediction accuracy at lower simulated abundances, but this does not appear to be the case for combinations of 3.

Those results should inform future studies attempting to perform wastewater based epidemiology using not only kallisto, but also other methods for quantifying variants of the SARS-CoV-2 virus. Since this is a novel method, the hope is that this work will aid in further efforts to improve the accuracy of kallisto. For other techniques, focusing on those well-performing regions, if that is possible in their pipeline, could also be attempted.

2 Experimental Setup

The SARS-CoV-2 virus presents a set of proteins: structural, non-structural, and accessory as seen in Figure 1. This presents a good starting point in trying to optimize the algorithm based on the choice of genomic region, as it enables us to compare functional sections of the virus. I don't analyze all possible sections as some of them are so small they approach the

[1] Fitness defined as giving the most accurate results in the kallisto algorithm

length of a sequencing read (concept explained later in this section). The threshold which I chose for a section to be included in the analysis is 300, double of 150, the length of a sequencing read.

Since the spike protein is of much interest, I also look at its composite parts, namely the subunit 1 (S1) and subunit 2 (S2) regions as seen in the figure. Further, in [7] wastewater from New York City was sampled and unique, never seen before, mutations on the spike region were observed by sequencing amino acid residues 412 to 579 for the MiSeq sequencer. For this reason, I am also interested in how this specific region performs in the analysis when given to kallisto. Note that in the last-mentioned paper, amino acid residues 434 to 505 were also analyzed for the iSeq sequencer. This was analyzed as it does not pass my set length threshold.

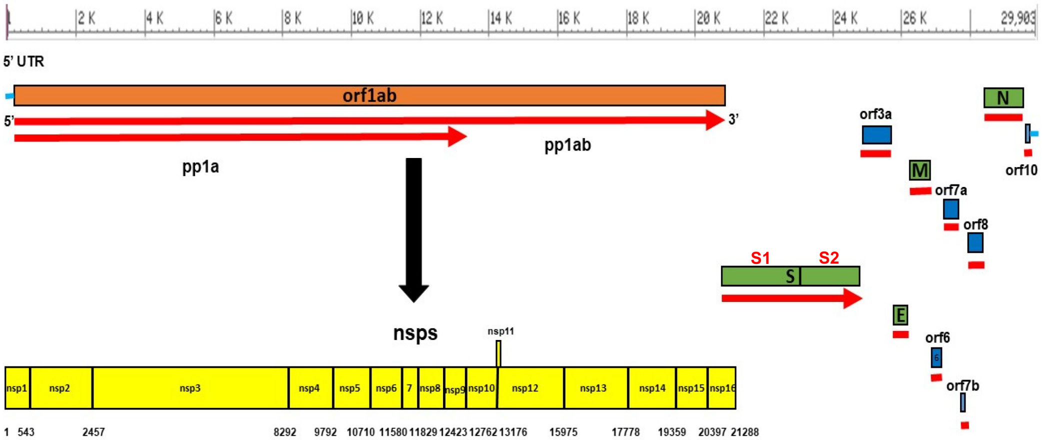

Figure 1: Genome organization of SARS-CoV-2 with its encoded proteins retrieved from [8] and modified to also include the S1 and S2 components of the spike protein

This leaves the following regions to be analyzed: the whole genome, all nsps with the exception of nsp11, orf1a, orf1ab, accessory proteins orf3a, orf7a, and orf8, structural proteins S, M, and N, and S1, S2, and amino acid residues 412 to 579 from MiSeq sequencing from the spike region. This first set of experiments already give interesting results (see section 3), but they present one big drawback: those regions are of unequal length, hence those comparisons are biased in this sense. Therefore, this motivates the next set of experiments, which is to look at equal length regions throughout the genome and compare them.

Lastly, I look at what happens if the best-performing regions from the previous experiments are combined. I am especially interested in seeing if I am able to improve prediction accuracy even in situations where the percentage of a variant in a wastewater sample is low.

For all those experiments, I look at the variants of concern (VOCs) alpha, beta, gamma, delta, and omicron.

Much of the setup is taken from [3] and adapted in order to answer the specific questions this study is attempting to answer. First, a reference set has to be constructed which kallisto would use to build an index, followed by simulating different wastewater samples, both of those being then used in the actual kallisto algorithm, and finally the results are evaluated using certain plots. Following, is a detailed description of each of those components.

In order to run kallisto, a reference set is first needed. For this, I choose to use data from

Connecticut from the entire year of 2021, retrieved from the GISAID EpiCov database [9]. Before using this, I further filter the data using the existing pipeline such that all mutations frequency of at least 50% are captured at least once. For that, we call each variant compared to the original SARS-CoV-2 reference (NC_045512.2) and compute allele frequencies per lineage. To achieve this, VCFtools [10], BCFtools [11], and minimap2 [12] from Bioconda [13] are used. All the exact epi accession numbers for the strains used in the reference set can be found in appendix A. After filtering, the indexing routine of kallisto can already be run.

Initially, I constructed the reference set using data from Connecticut between October 2020 and September 2021. The reason for this is that in the meantime the omicron variant started gaining ground as a variant of concern, leading to the decision to also include it in my analysis. Before this, I had already managed to run the first two sets of experiments based on the first reference set.

When a sequencer is used to analyze a sample containing genomic information, it does not read genomes in their entirety, but rather it takes multiple smaller overlapping regions, named sequencing reads, that can be used to reconstruct the actual genomes. The wastewater samples are simulated using ART [14], based on Illumina sequencers, with a sequencing depth of 100, meaning that each base is read an average of 100 times. In order to make the simulations better approximate real wastewater samples, each of the benchmarks also contains background virus variants, meaning that they are not from variants of concern amongst the variant which is being simulated. For those, I use sequences sampled on the 1st of March in Connecticut, while for the VOCs, B.1.1.7 (EPI_ISL_4371224), B.1.351 (EPI_ISL_4372166), B.1.617.2 (EPI_ISL_2035068), P.1 (EPI_ISL_3104694), and BA.1 (EPI_ISL_8185077) samples are used as representative of alpha, beta, gamma, delta, and gamma respectively, all retrieved from the GISAID EpiCov database. The sublineage BA.1 was used in the case of omicron instead of the original B.1.1.529 lineage as there were no samples of it at the selected time and location. For each VOC and for each region I test, I have built different benchmarks based on the simulated abundance of the VOCs being analyzed: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 per cent of the wastewater sample.

Now having a reference set, the kallisto index, and simulated wastewater samples, the quant routine of kallisto can be run, the results of which are turned in abundance estimations of each VOC. Finally, having those results, the relative predictions error can be plotted against the simulated VOC frequency or the region of the genome, depending on the experiment being performed. For plotting, matplotlib is used [15]. The relative prediction error is calculated as follows:

$$relative\_prediction\_error = estimated\_abundance/true\_abundance * 100$$

Of note is that pandas [16] is extensively used throughout the descried pipeline.

For the first set of experiments, where I look at functional regions of the genome, the actual simulated region is larger by a buffer of 150 nucleotides on each side. That is for two reasons, the first one being that the regions don't start at the exact position for each genomic sample as each one can present insertion or deletion mutations at different locations. Secondly, because of how sequencing works, the density of reads is less at the very ends of a region and therefore a buffer region is able to account for that. The length of 150 was specifically chosen because this is also the length of the sequencing reads when simulating with ART.

Equal length regions are analyzed in the second set of experiments. The nucleotide lengths chosen for analysis are as follows: 250, 450, 750, 1000, 1500, 2000, 3000, 4000, 5000, 6000, and 7500. Lengths 250 and 450 are used as they reflect standard sequencing protocols (Illumina ISeq and Illumina MiSeq respectively). Since analyzing every possible region for each of those lengths quickly becomes computationally infeasible, I opted for choosing them so that the starting positions of consecutive regions are at a distance of a fifth of the region length from each other (ex. for the 1000 nucleotide length, the analyzed regions are: 1 to 1001, 201 to 1201, 401 to 1401 etc.).

Finally, depending on the results of the second experiment, those best performing regions are combined in order to determine if this would lead to even better accuracy compared to sequencing them alone. The way in which those regions were selected can be found at the start of section 3.3, while the results that motivated it is present in section 3.2. All non-overlapping combinations of best-performing regions were analyzed. This was done by concatenating the fastq files representing the simulated sequencing reads of a genome and then applying the quant routine of kallisto on them.

3 Results

Throughout this section, the results which I considered to be most significant are presented. Their interpretation can be found in section 5. Throughout this section, I keep mentioning the new and old reference sets/ results. Those refer to the experiments done with the data from the entire year of 2021 and from October 2020 to September 2021 respectively. The full set of results can be found on data.4tu.nl at DOI 10.4121/18532973.

At the end of this section, I also reflect on possible biases that might have influenced the results due to the choice of reference and simulation data.

3.1 Specific regions

As discussed in section 2, the first set of experiments consists of looking at regions associated with actual proteins of the genome.

In Figure 2 the performance of the spike region, together with its subregions is shown for the new set of experiments which include the omicron variant. Here, it does not seem like looking at subregions gives any significant advantage over taking the entire spike region, which is at odds with the results from the experiments made on the old reference set. The results of those for the spike region can be seen in Figure 3. Only having the latter, one could have assumed that region S1 of the spike region would be advantageous over the others.

In Figure 4 I show the comparison between sequencing the entire genome and sequencing the S, M, and N regions for the new set. With the exception of the beta variant, the N region, even though not as much as the spike region or the whole genome, seems to also be good at predicting the correct variant abundance.

3.2 Equal length regions

For the second set of experiments, I look at equal length regions. Namely, I analyze regions of 250, 450, 750, 1000, 1500, 2000, 3000, 4000, 5000, 6000, and 7500 nucleotides.

Some of the results of the second set of experiments (those for regions of length 750, 1500, and 3000 with VOCs simulated at 80% abundance) can be observed in figures 5 and 6 for the new and old reference sets respectively. The rest of the graphs relating to different simulated abundances can be found on data.4tu.nl at DOI 10.4121/18532973.

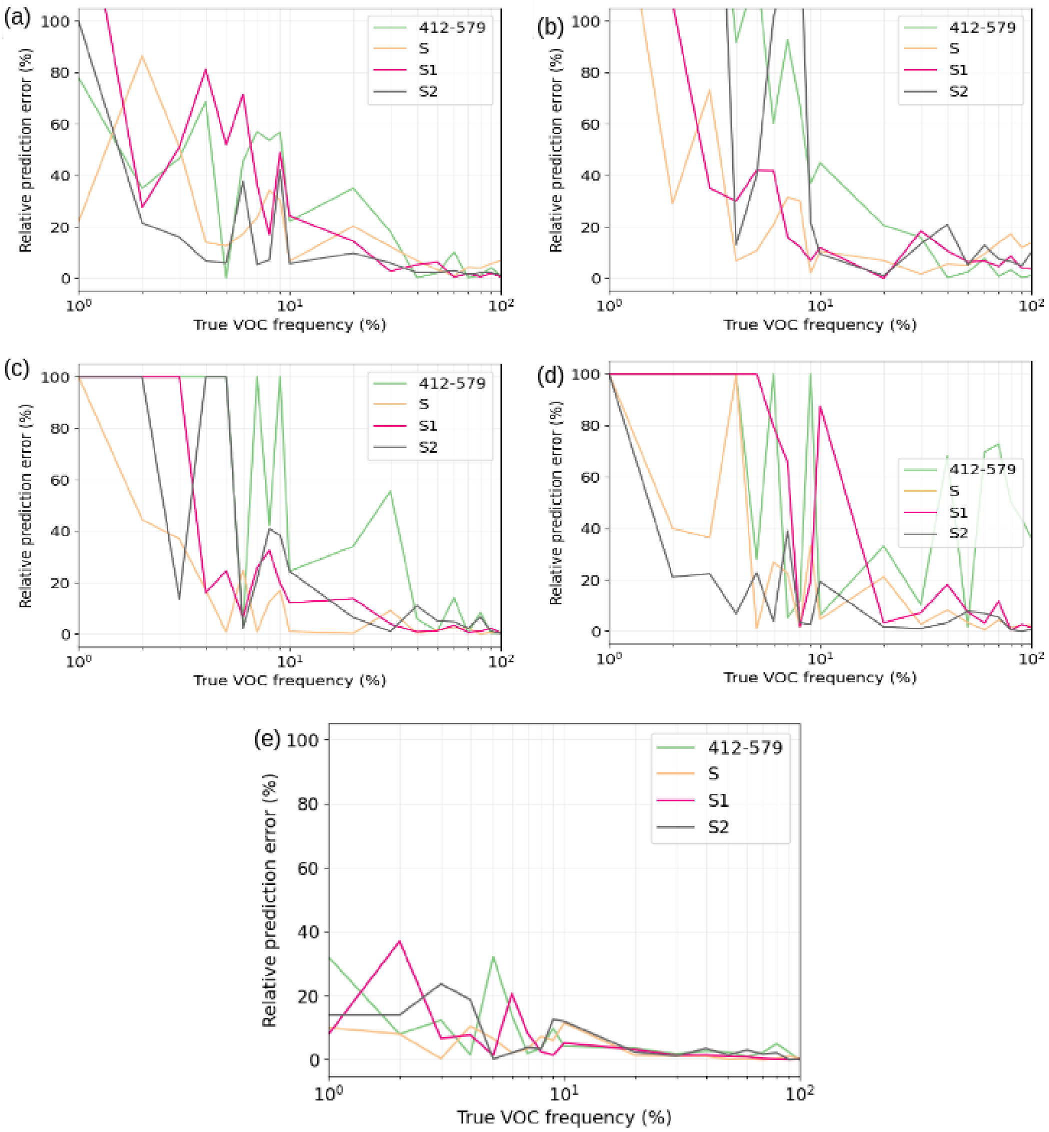

Figure 2: Graphs showing the new results from sequencing the region coding for the spike protein (S) together with sub-regions of it: S1, S2, and 412-579 from the New York City wastewater [7]. Each of the (a), (b), (c), (d), and (e) are associated with the variants of concern alpha, beta, gamma, delta, and omicron respectively

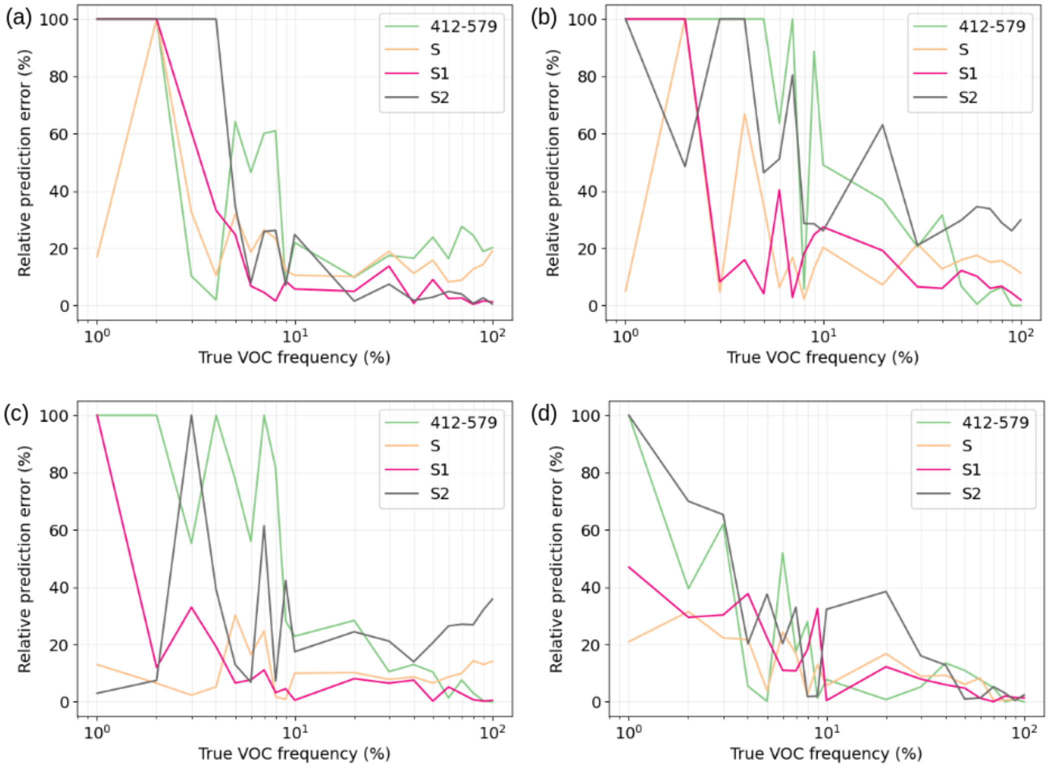

Figure 3: Graphs showing the old results from sequencing the region coding for the spike protein (S) together with sub-regions of it: S1, S2, and 412-579 from the New York City wastewater [7]. Each of the (a), (b), (c), and (d) are associated with the variants of concern alpha, beta, gamma, and delta respectively

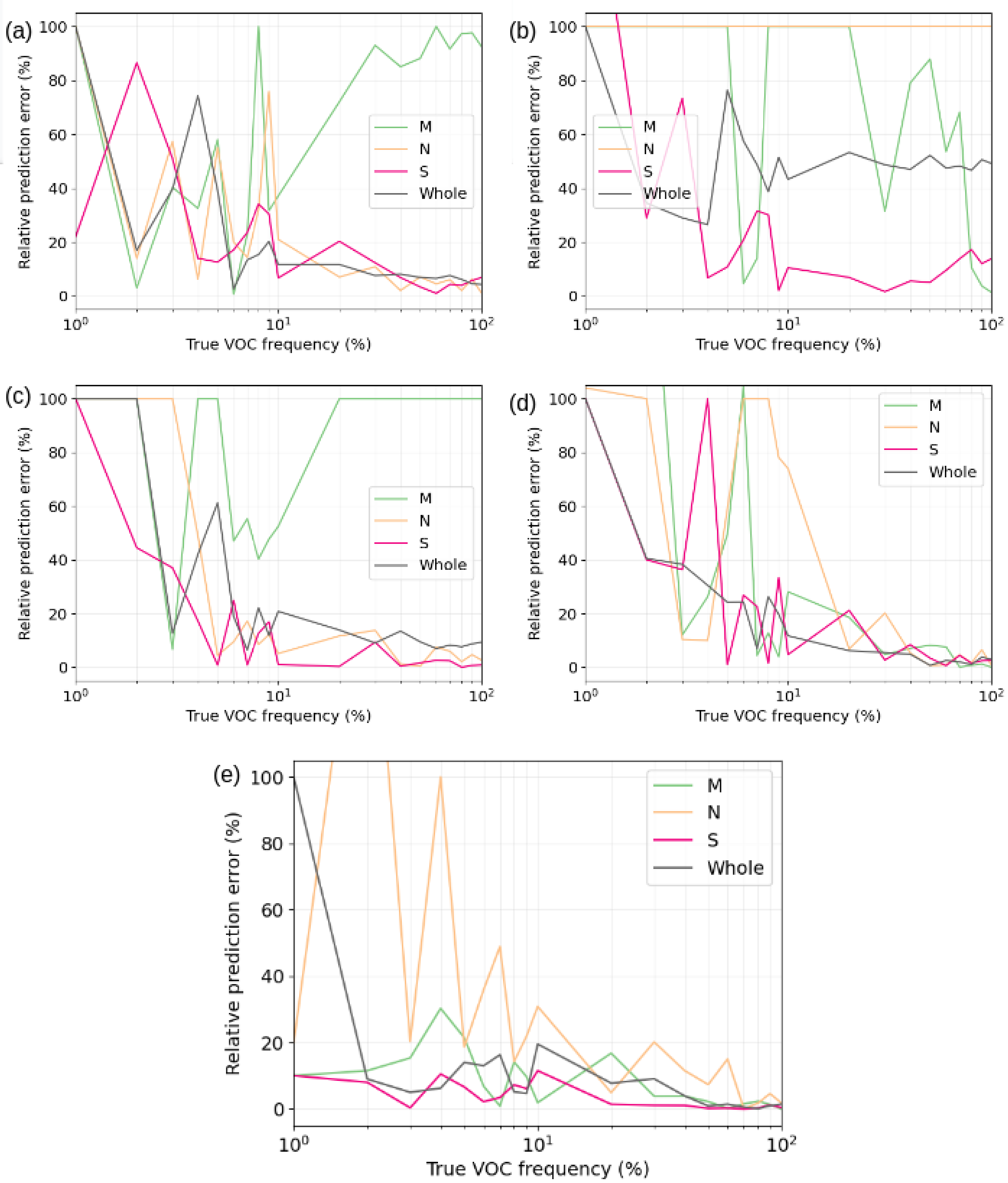

Figure 4: Graphs showing the results from sequencing the whole genome compared with the regions coding for the membrane (M), nucleocapsid (N), and spike (S) structural proteins. Each of the (a), (b), (c),(d), and (e) are associated with the variants of concern alpha, beta, gamma, delta, and omicron respectively

Those support the findings of the previous experiments as the areas appertaining to the S and N regions seem to be particularly good at predicting abundances for all variants, with the exception, again, of the beta variant in the case of the new reference set. The beta variant proves to be particularly hard to predict, especially for the new reference set. Analyzing the results, it turns out that kallisto misassigns reads from the beta variant to other background variants from the simulated wastewater samples which leads to underestimation of the true abundance.

In addition to S and N, it appears that a section around the middle of the region coding for the non-structural protein 3 (nsp3) is also particularly good at predicting the correct abundances in both the new and old results.

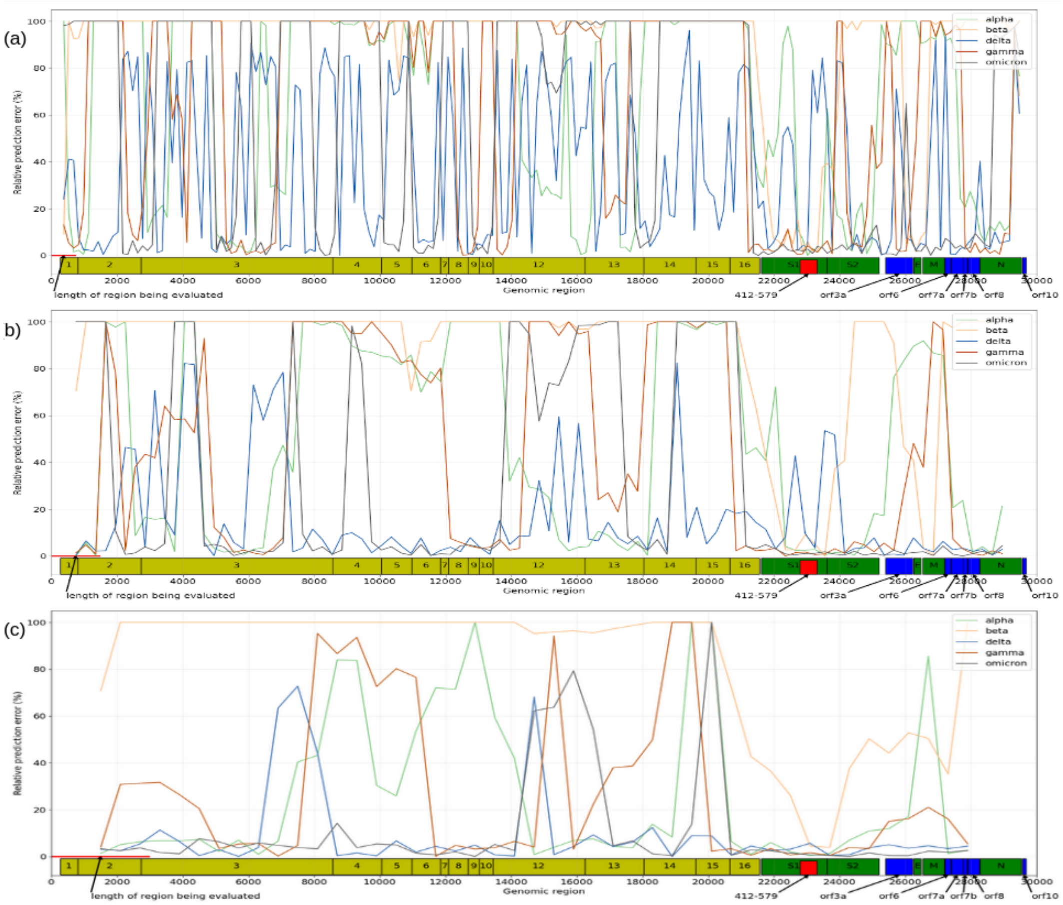

Figure 5: Graphs showing the new results from sequencing equal length region at a simulated abundance of 80% for the new reference set. The x-axis is a mapping to the actual position in the genome, while the y-axis represents the relative prediction error. A data point represents the middle of an analyzed region. The shown graphs are for the regions of length 750 (a), 1500 (b), and 3000 (c). At the bottom left of each graph, there is a red line that shows the nucleotide length being analyzed relative to the x-axis. Just above the horizontal axis, there is also a bar showing how each position maps to a specific functional region which helps in interpreting the results.

To better visualize the good performing regions, I plot those that have a relative prediction error of 20 or under, for true simulated abundances of 20 and higher for all variants except beta with the new reference set (see Figure 7). The choice to exclude beta was made as its inclusion would have resulted in way less of those regions since this variant is particularly hard to predict accurately. The highlighted regions are even better as they also have 20 or smaller relative prediction error for simulated abundance 10, again, with the exception of the beta variant. This goes to confirm my initial observation regarding nsp3, S, and N regions. In addition, those best performing regions are also the ones that are used in the last set of experiments where combinations of regions are analyzed.

3.3 Combining regions

The final set of experiments consists in combining those regions I found to be best performing from the previous experiments. This was done only for the new reference set.

The specific regions which were chosen are the ones that had 20 or under relative prediction error for simulated abundances of 10 or higher for all variants except beta. The beta variant was excluded as it is uniquely hard to predict compared to the others. Note that those regions are also the ones highlighted in Figure 7. I analyze all the possible non-overlapping combinations of those regions. Thus, the maximum number of regions in a combination is three, as can also be intuitively deduced by looking at the last-mentioned figure.

Here I only discuss 2 sets of combinations, as seen in Figure 8, because of space limitations. The full results can be found on data.4tu.nl at DOI 10.4121/18532973.

In Figure 8.I, it seems that combining two regions gives a slight advantage over only applying the algorithm to single regions. This advantage does not seem to increase with the number of regions in a combination, since, as seen in Figure 8.II, the three-region combination consisting of regions 5001-7500, 20801-24800, and 27001-29500 only seems to have a clear advantage over the two-region combination consisting of regions 5001-7500 and 27001-29500.

3.4 Possible biases

Lastly, I would like to draw attention to the fact that the genomic strains used for delta (EPI_ISL_2035068) and omicron (EPI_ISL_8185077) variants for simulating wastewater samples are also present in the new reference set. This is not the case for the other variants which might confer the delta and omicron variants an unfair advantage when their results are compared with the others' results.

More than that, since omicron is a new variant, at the time the data was retrieved for the year 2021 for Connecticut there was only this one omicron variant left to be used in the reference set, which could have clearly given it a big advantage in my analysis. It can indeed be observed that omicron is usually more easily predicted than the other variants. The question stands if omicron is indeed easier to predict overall or if this is just a result of my experimental setup.

4 Responsible Research

All scripts that I created and used to achieve my results are available on data.4tu.nl at DOI 10.4121/18532973. I am not, however able to share the data directly share this data because

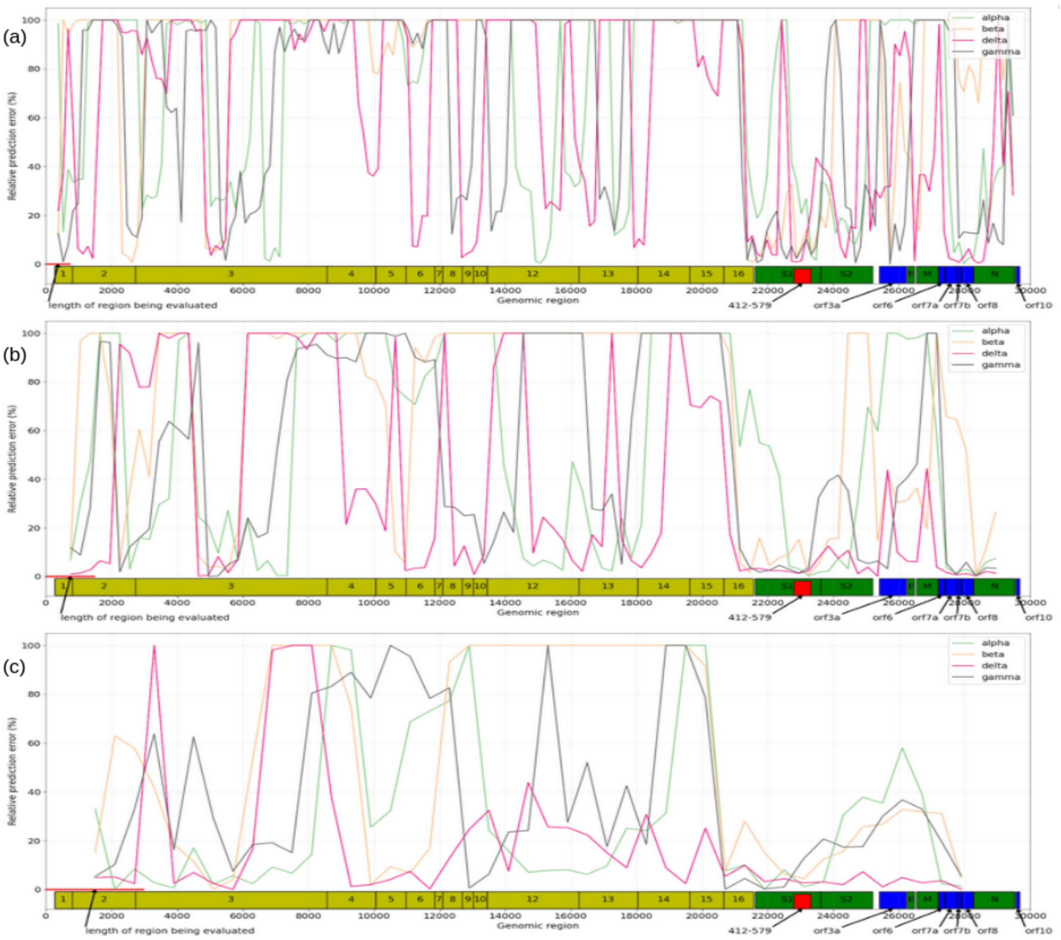

Figure 6: Graphs showing the new results from sequencing equal length region at a simulated abundance of 80% for the old reference set. The x-axis is a mapping to the actual position in the genome, while the y-axis represents the relative prediction error. A data point represents the middle of an analyzed region. The shown graphs are for the regions of length 750 (a), 1500 (b), and 3000 (c). At the bottom left of each graph, there is a red line that shows the nucleotide length being analyzed relative to the x-axis. Just above the horizontal axis, there is also a bar showing how each position maps to a specific functional region which helps in interpreting the results.

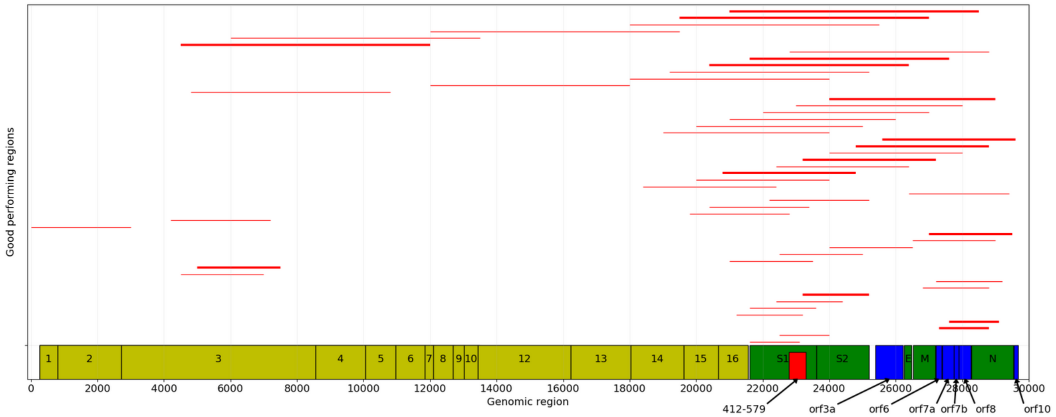

Figure 7: Graph showing those regions that have a relative prediction error of 20 or under, for true simulated abundances of 20 and higher for all variants except beta with the new reference set. Additionally, the regions represented by a thicker line also have 20 or smaller relative prediction error for simulated abundance 10, with the exception of the beta variant. They are also the ones used in the last set of experiments where good performing regions are combined.

this would be against the terms of use that I signed before gaining access to the GISAID EpiCov database. Nevertheless, it is generally easy for someone interested to gain access to this database themselves.

This can be done by registering on their platform at www.gisaid.org. To do this, it is necessary to provide your identity and accept the terms of use. It is mentioned on their FAQ page that *"This requirement is not only essential to help uphold the integrity of the GISAID user community, but necessary to enforce the GISAID sharing mechanism that assures reciprocity of the data for future generations"*.

All epi accession numbers for the strains used in the new reference set, for creating the background virus variants and for representing each variant of concern can be found in appendix A. Having those, someone interested in reproducing the experiments could find and retrieve the SARS-CoV-2 genomes used here. Unfortunately, I did not think to also keep the epi accession numbers for the old reference set as well and, consequently, I cannot provide those. Despite this, in section 2 I provide the process of retrieving and filtering the data from GISAID to create this reference set. Therefore, someone interested in also redoing my old experiments could in theory recreate the old reference set.

Also relevant to the responsible research section, is that I lay out my concerns of how the way in which I set up my experiments might give advantages to certain variants in the final analysis. Those can be found in section 3.4.

5 Discussion

In [3], the paper that first looked at kallisto as a means to estimate variant abundance for the Sars-Cov-2 virus in wastewater samples, Baaijens et al. have already found that their results are more accurate when simulating reads only for the genomic region coding for the spike protein (S). As they pointed out, it suggests that the algorithm prefers sequencing

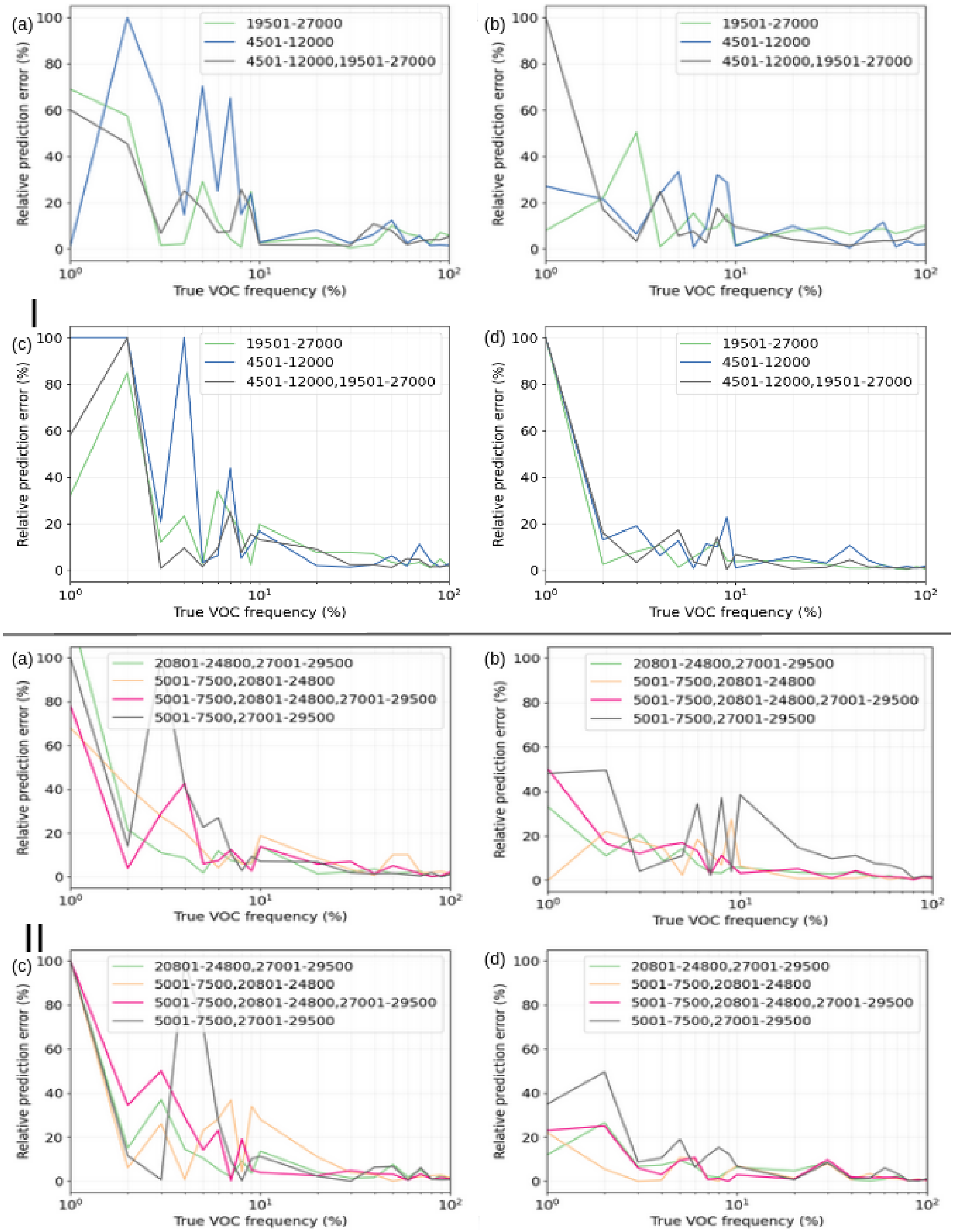

Figure 8: Graphs showing the results from combining sequencing data from regions that perform very well on their own. Those on the top, denoted with (I), show a comparison between combining 2 regions (namely 4501-12000 and 19501-27000) and analyzing them alone, while those on the bottom, denoted with (II), compare combinations of 3 (namely 5001-7500, 20801-24800, and 27001-29500) with combinations of 2. Each of the (a), (b), (c), and (d) are associated with the variants of concern alpha, gamma, delta, and omicron respectively

death over sequencing breath, but the selected region also needs to uniquely identify the virus variants.

In Figure 9, I plotted the allele frequency of the variants that I analyzed. An allele is a specific variant of a gene, while the allele frequency refers to the number of times this variant of the gene appears in a specific population compared to the number of individuals in that population. In order to count alleles, there also needs to be an original strain from which to compare the others and for that, I used the NC_045512.2 strain. The populations I used for plotting this are made up of the strains that are used in the new reference set, with the exception of the omicron variant which also contains three additional strains that got uploaded for Connecticut in the year 2021 after the reference set was built.

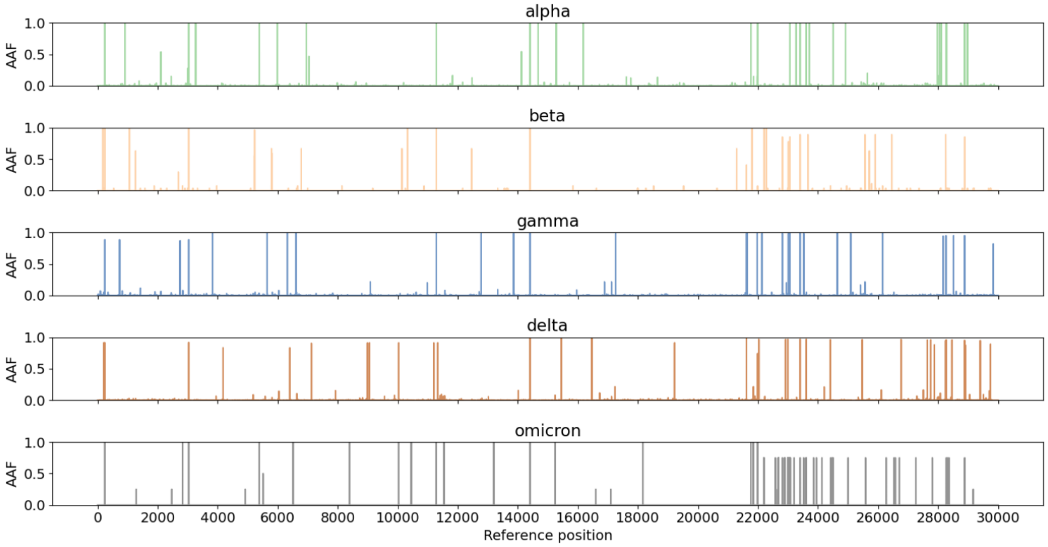

Figure 9: Plot showing the allele frequency for each of the variants of concern compared to the original SARS-CoV-2 reference (NC_045512.2)

What is obvious from looking at this plot is that all variants are mostly defined by changes present at the last quarter of the genome, especially in the spike area. Except for beta, they also have many defining alleles at the end of the genome, around where the N region starts. Beta seems to have the least amount of defining alleles which would in part explain the difficulty kallisto had in predicting its correct abundance.

In [6] it was shown that the evolution of SARS-CoV-2 is mostly characterized by purifying selections (removal of deleterious alleles), but there are also some positive selections (selection for fit alleles) on certain sites. The spike region, specifically the receptor-binding domain, and the nucleocapsid (N) region, specifically the area associated with nuclear localization signals, were shown to be particularly enriched with those positive selections. This, together with the graph where I plotted the allele frequencies for the analyzed variants, supports my finding that the genomic regions coding for the spike and nucleocapsid proteins are particularly good at accurately predicting virus variant abundances.

The most surprising result is that a section around the middle of the nsp3 region is particularly fit in terms of prediction accuracy for this method of abundance quantification. According to [17], the nsp3 is a *'papain-like proteinase'* which *'functions as a protease to* *separate the translated polyprotein into its distinct proteins'.*

Lastly, it seems like combining two good performing regions confers a slight advantage over using one on those alone. This is not usually the case for combining three of those regions as there are cases in which two-region combinations are found to be better. This is perhaps because by increasing sequencing breath, relative sequencing depth is lost and, since kallisto seems to prefer sequencing depth over breath, accuracy is lessened.

6 Conclusions and Future Work

Using kallisto, it is possible to quantify different variants of SARS-CoV-2 in wastewater samples. In this research, I showed that it is possible to further improve this technique by only sequencing for specific segments of the genome, as opposed to looking at the whole. Specifically, simulating reads from the spike protein, the region coding for the nucleocapsid protein, and a section around the middle of the non-structural protein 3 give the most accurate results in this method of quantifying abundances. Further, I have shown that combining sequencing reads from two good performing regions appears to slightly improve the results.

Those results should inform those actually conducting wastewater analysis as they could focus on those regions which I found to be the best performing. Furthermore, they can also use combinations (of two) of those to further increase prediction accuracy at lower abundances.

Future work should be aimed at further improving the prediction accuracy of this technique. A logical follow up would be to address the concerns expressed in section 3.4 by making sure all strains used to simulate the wastewater samples don't also appear in the reference set which should also contain more strains from the omicron variant.

The area around the nsp3 middle section giving good results in distinguishing between variants suggests that this region, in particular, carries certain important information which is able to distinguish between variants. Hypothesising about the role of this information in the SARS-CoV-2 virus could be of much interest to someone more well versed in the biology part of bioinformatics.

References

[1] The European Commission, "Commission recommendation (eu) 2021/472 of 17 march 2021 on a common approach to establish a systematic surveillance of sars-cov-2 and its variants in wastewaters in the eu", 2021. [Online]. Available: `https://op.europa.eu/s/unTm`.

[2] European Centre for Disease Prevention and Control, "Detection and characterisation capability and capacity for sars-cov-2 variants within the eu/eea", 2021. [Online]. Available: `https://www.ecdc.europa.eu/en/publications-data/detection-and-characterisation-capability-and-capacity-sars-cov-2-variants`.

[3] J. A. Baaijens, A. Zulli, I. M. Ott, *et al.*, "Variant abundance estimation for sars-cov-2 in wastewater using rna-seq quantification", *medRxiv*, 2021. DOI: `10.1101/2021.08.31.21262938`. eprint: `https://www.medrxiv.org/content/early/2021/09/02/2021.08.31.21262938.full.pdf`. [Online]. Available: `https://www.medrxiv.org/content/early/2021/09/02/2021.08.31.21262938`.

[4] "A pan-European study of SARS-CoV-2 variants in wastewater under the EU Sewage Sentinel System", *medRxiv*, p. 2021.06.11.21258756, Jun. 2021. DOI: 10.1101/2021.06.11.21258756. [Online]. Available: https://www.medrxiv.org/content/10.1101/2021.06.11.21258756v1%20https://www.medrxiv.org/content/10.1101/2021.06.11.21258756v1.abstract.

[5] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic rna-seq quantification", *Nature Biotechnology*, vol. 34, no. 5, pp. 525–527, May 2016, ISSN: 1546-1696. DOI: 10.1038/nbt.3519. [Online]. Available: https://doi.org/10.1038/nbt.3519.

[6] N. D. Rochman, Y. I. Wolf, G. Faure, P. Mutz, F. Zhang, and E. V. Koonin, "Ongoing global and regional adaptive evolution of SARS-CoV-2", *Proceedings of the National Academy of Sciences*, vol. 118, no. 29, e2104241118, Jul. 2021. DOI: 10.1073/pnas.2104241118. [Online]. Available: http://www.pnas.org/content/118/29/e2104241118.abstract.

[7] D. S. Smyth, M. Trujillo, D. A. Gregory, *et al.*, "Tracking cryptic sars-cov-2 lineages detected in nyc wastewater", *medRxiv*, 2021. DOI: 10.1101/2021.07.26.21261142. eprint: https://www.medrxiv.org/content/early/2021/07/29/2021.07.26.21261142.full.pdf. [Online]. Available: https://www.medrxiv.org/content/early/2021/07/29/2021.07.26.21261142.

[8] Y. A. Helmy, M. Fawzy, A. Elaswad, A. Sobieh, S. P. Kenney, and A. A. Shehata, "The COVID-19 Pandemic: A Comprehensive Review of Taxonomy, Genetics, Epidemiology, Diagnosis, Treatment, and Control", *Journal of Clinical Medicine*, vol. 9, no. 4, 2020, ISSN: 2077-0383. DOI: 10.3390/jcm9041225. [Online]. Available: https://www.mdpi.com/2077-0383/9/4/1225.

[9] S. Elbe and G. Buckland-Merrett, "Data, disease and diplomacy: Gisaid's innovative contribution to global health", *Global Challenges*, vol. 1, no. 1, pp. 33–46, 2017. DOI: https://doi.org/10.1002/gch2.1018. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gch2.1018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/gch2.1018.

[10] P. Danecek, A. Auton, G. Abecasis, *et al.*, "The variant call format and VCFtools", *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, Aug. 2011, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr330. [Online]. Available: https://doi.org/10.1093/bioinformatics/btr330.

[11] P. Danecek, J. K. Bonfield, J. Liddle, *et al.*, "Twelve years of SAMtools and BCFtools.", eng, *GigaScience*, vol. 10, no. 2, Feb. 2021, ISSN: 2047-217X (Electronic). DOI: 10.1093/gigascience/giab008.

[12] H. Li, "Minimap2: pairwise alignment for nucleotide sequences", *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, Sep. 2018, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty191. [Online]. Available: https://doi.org/10.1093/bioinformatics/bty191.

[13] B. Grüning, R. Dale, A. Sjödin, *et al.*, "Bioconda: sustainable and comprehensive software distribution for the life sciences", *Nature Methods*, vol. 15, no. 7, pp. 475–476, 2018, ISSN: 1548-7105. DOI: 10.1038/s41592-018-0046-7. [Online]. Available: https://doi.org/10.1038/s41592-018-0046-7.

[14] W. Huang, L. Li, J. R. Myers, and G. T. Marth, "ART: a next-generation sequencing read simulator", *Bioinformatics*, vol. 28, no. 4, pp. 593–594, Feb. 2012, ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btr708`. [Online]. Available: `https://doi.org/10.1093/bioinformatics/btr708`.

[15] J. D. Hunter, "Matplotlib: A 2d graphics environment", *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. DOI: `10.1109/MCSE.2007.55`.

[16] T. pandas development team, *Pandas-dev/pandas: Pandas*, version latest, Feb. 2020. DOI: `10.5281/zenodo.3509134`. [Online]. Available: `https://doi.org/10.5281/zenodo.3509134`.

[17] A. A. T. Naqvi, K. Fatima, T. Mohammad, *et al.*, "Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach.", eng, *Biochimica et biophysica acta. Molecular basis of disease*, vol. 1866, no. 10, p. 165 878, Oct. 2020, ISSN: 1879-260X (Electronic). DOI: `10.1016/j.bbadis.2020.165878`.

A Used EPI accession numbers

A.1 Reference set

EPI_ISL_1139261 EPI_ISL_1139313 EPI_ISL_1139566 EPI_ISL_1139567 EPI_ISL_1158943
EPI_ISL_1158960 EPI_ISL_1159051 EPI_ISL_1161151 EPI_ISL_1163342 EPI_ISL_1193800
EPI_ISL_1193449 EPI_ISL_1193475 EPI_ISL_1194199 EPI_ISL_1203821 EPI_ISL_1225731
EPI_ISL_1254412 EPI_ISL_1254419 EPI_ISL_1254189 EPI_ISL_1255296 EPI_ISL_1139575
EPI_ISL_1267153 EPI_ISL_1273319 EPI_ISL_1273336 EPI_ISL_1273337 EPI_ISL_1293137
EPI_ISL_1297685 EPI_ISL_854866 EPI_ISL_855084 EPI_ISL_886399 EPI_ISL_886543
EPI_ISL_967249 EPI_ISL_1016681 EPI_ISL_1016688 EPI_ISL_1017472 EPI_ISL_1017474
EPI_ISL_1017476 EPI_ISL_1017481 EPI_ISL_1017495 EPI_ISL_1017498 EPI_ISL_1021202
EPI_ISL_1021232 EPI_ISL_1021258 EPI_ISL_1021259 EPI_ISL_1032158 EPI_ISL_1032353
EPI_ISL_1032430 EPI_ISL_1087053 EPI_ISL_1086734 EPI_ISL_1086751 EPI_ISL_1086789
EPI_ISL_1087102 EPI_ISL_1087440 EPI_ISL_1087798 EPI_ISL_1087849 EPI_ISL_1090567
EPI_ISL_1090602 EPI_ISL_1090621 EPI_ISL_1090898 EPI_ISL_1090918 EPI_ISL_1091865
EPI_ISL_1314947 EPI_ISL_1315168 EPI_ISL_1339920 EPI_ISL_1401504 EPI_ISL_1401552
EPI_ISL_1067693 EPI_ISL_1315200 EPI_ISL_1163004 EPI_ISL_1163669 EPI_ISL_1163684
EPI_ISL_1339946 EPI_ISL_1293215 EPI_ISL_1667600 EPI_ISL_1540810 EPI_ISL_1681339
EPI_ISL_1680780 EPI_ISL_1738612 EPI_ISL_1753640 EPI_ISL_1790008 EPI_ISL_1818233
EPI_ISL_1826486 EPI_ISL_1826533 EPI_ISL_1834545 EPI_ISL_1963788 EPI_ISL_2182642
EPI_ISL_2011029 EPI_ISL_2035068 EPI_ISL_2133124 EPI_ISL_2230038 EPI_ISL_2268674
EPI_ISL_2296394 EPI_ISL_2306914 EPI_ISL_2371884 EPI_ISL_2384069 EPI_ISL_2399109
EPI_ISL_2440795 EPI_ISL_2500436 EPI_ISL_2501721 EPI_ISL_2501722 EPI_ISL_3104752
EPI_ISL_3354882 EPI_ISL_4366185 EPI_ISL_4366845 EPI_ISL_4368112 EPI_ISL_4369028
EPI_ISL_4966775 EPI_ISL_4985293 EPI_ISL_4985349 EPI_ISL_4985696 EPI_ISL_3320661
EPI_ISL_3320681 EPI_ISL_3324747 EPI_ISL_3347254 EPI_ISL_3370111 EPI_ISL_3370255
EPI_ISL_3455461 EPI_ISL_3493185 EPI_ISL_3493212 EPI_ISL_3493225 EPI_ISL_3500569
EPI_ISL_3500598 EPI_ISL_3500698 EPI_ISL_3500878 EPI_ISL_3500894 EPI_ISL_3500922
EPI_ISL_3512895 EPI_ISL_3512896 EPI_ISL_3512967 EPI_ISL_3512972 EPI_ISL_2716210
EPI_ISL_2716215 EPI_ISL_2860314 EPI_ISL_3605753 EPI_ISL_3670173 EPI_ISL_3670532
EPI_ISL_3670202 EPI_ISL_2860305 EPI_ISL_2860313 EPI_ISL_2869056 EPI_ISL_2876851
EPI_ISL_2930238 EPI_ISL_2930270 EPI_ISL_2930312 EPI_ISL_2930321 EPI_ISL_2960038

EPI_ISL_2960040 EPI_ISL_2960044 EPI_ISL_3018282 EPI_ISL_3025429 EPI_ISL_3025436 EPI_ISL_3025441 EPI_ISL_3061868 EPI_ISL_3329472 EPI_ISL_3066877 EPI_ISL_3104720 EPI_ISL_3104721 EPI_ISL_3104726 EPI_ISL_3104730 EPI_ISL_3104731 EPI_ISL_3104735 EPI_ISL_3104742 EPI_ISL_3104745 EPI_ISL_3111155 EPI_ISL_3112333 EPI_ISL_3114605 EPI_ISL_3151580 EPI_ISL_3151599 EPI_ISL_3152940 EPI_ISL_3216766 EPI_ISL_3236376 EPI_ISL_3236382 EPI_ISL_3236409 EPI_ISL_3236431 EPI_ISL_3236494 EPI_ISL_3319122 EPI_ISL_3319126 EPI_ISL_3319134 EPI_ISL_3324307 EPI_ISL_2874058 EPI_ISL_3511309 EPI_ISL_3347246 EPI_ISL_4158306 EPI_ISL_3693128 EPI_ISL_3693178 EPI_ISL_3815852 EPI_ISL_3829721 EPI_ISL_3829758 EPI_ISL_3840809 EPI_ISL_3841117 EPI_ISL_3841118 EPI_ISL_3841134 EPI_ISL_3841861 EPI_ISL_3841892 EPI_ISL_3841904 EPI_ISL_3848252 EPI_ISL_3945566 EPI_ISL_4006132 EPI_ISL_4029875 EPI_ISL_3841049 EPI_ISL_3841121 EPI_ISL_3746356 EPI_ISL_4164270 EPI_ISL_3740144 EPI_ISL_4331988 EPI_ISL_4332360 EPI_ISL_4198005 EPI_ISL_4198122 EPI_ISL_4239450 EPI_ISL_4366957 EPI_ISL_4370508 EPI_ISL_4454359 EPI_ISL_4483074 EPI_ISL_4483410 EPI_ISL_4537170 EPI_ISL_4537222 EPI_ISL_4239653 EPI_ISL_4367582 EPI_ISL_4367013 EPI_ISL_4164570 EPI_ISL_4164542 EPI_ISL_4331983 EPI_ISL_5112247 EPI_ISL_5170105 EPI_ISL_5196015 EPI_ISL_4576913 EPI_ISL_4576919 EPI_ISL_4576935 EPI_ISL_4824364 EPI_ISL_5021823 EPI_ISL_5088486 EPI_ISL_4950050 EPI_ISL_4576895 EPI_ISL_5320698 EPI_ISL_5584461 EPI_ISL_5644981 EPI_ISL_5682681 EPI_ISL_5721359 EPI_ISL_5868737 EPI_ISL_5868779 EPI_ISL_5875671 EPI_ISL_5878404 EPI_ISL_5927401 EPI_ISL_5794677 EPI_ISL_6173628 EPI_ISL_6398289 EPI_ISL_6473826 EPI_ISL_6550464 EPI_ISL_6568294 EPI_ISL_6166451 EPI_ISL_6307072 EPI_ISL_6683259 EPI_ISL_6483496 EPI_ISL_6914356 EPI_ISL_6940296 EPI_ISL_7408235 EPI_ISL_7247647 EPI_ISL_7252218 EPI_ISL_7447104 EPI_ISL_7447173 EPI_ISL_6940378 EPI_ISL_7674106 EPI_ISL_7696883 EPI_ISL_7696955 EPI_ISL_7783721 EPI_ISL_8017670 EPI_ISL_8029377 EPI_ISL_8185077

A.2 Background lineages

EPI_ISL_1265971 EPI_ISL_1265987 EPI_ISL_1266150 EPI_ISL_1266157 EPI_ISL_1266224 EPI_ISL_1266881 EPI_ISL_1267036 EPI_ISL_1267075 EPI_ISL_1267146 EPI_ISL_1267152 EPI_ISL_1273334 EPI_ISL_1273351 EPI_ISL_1273352 EPI_ISL_1273355 EPI_ISL_1273358 EPI_ISL_1273359 EPI_ISL_1273370 EPI_ISL_1289787 EPI_ISL_1293210

A.3 VOCs

alpha - EPI_ISL_4371224
beta - EPI_ISL_4372166
gamma - EPI_ISL_3104694
delta - EPI_ISL_2035068
omicron - EPI_ISL_8185077