The impact of sequencing errors and contaminating viruses on SARS-CoV-2 variant detection by sequencing wastewater-sourced viral RNA

by

"M.J. van der Lugt"

Supervisor(s):

J.A. Baaijens

A Dissertation

Submitted to EEMCS faculty

Delft University of Technology,

In Partial Fulfillment of the Requirements

For the Bachelor of Computer Science and Engineering

January 28, 2022

The impact of sequencing errors and contaminating viruses on SARS-CoV-2 variant detection by sequencing wastewater-sourced viral RNA

MART VAN DER LUGT[1]
Supervisor: J.A. Baaijens[1]
[1]EEMCS, Delft University of Technology, The Netherlands
M.J.vanderlugt@student.tudelft.nl, J.A.Baaijens@tudelft.nl

Since the start of the SARS-CoV-2 pandemic, the monitoring of SARS-CoV-2 by way of viral RNA sequencing of wastewater has proven to be an efficient and effective way of estimating COVID-19 cases in population groups. A recently developed pipeline also enables us to estimate SARS-CoV-2 variant abundance using viral samples from wastewater. This is done by repurposing an RNA-seq quantification algorithm to quantify reads, belonging to variants, from DNA-sequencing data. However, the impact of sequencing errors and contaminating viruses on this process is unknown. Here I show that, in simulated data, the credibility of the prediction results is dependent on the error rate of the sequencing machines used. I also show that contaminating the simulated dataset with certain human coronaviruses has a significant effect on prediction accuracy. However, most viruses currently found in wastewater have no effect. Furthermore, adding a reference genome for these human corona-viruses to the reference set removes any impact. The results demonstrate that it is important to assess the credibility of the pipeline on a case by case basis and to tailor the testing setup and reference set to this assessment.

Early December 2019, a new coronavirus was discovered in Wuhan, China[19]. This virus, which is closely related to the SARS-CoV virus, the causative agent behind the SARS outbreak in the early 2000s, had already claimed the lives of 80 people by the 26th of January the next year.

The virus, cleverly dubbed SARS-CoV-2, would proceed to go worldwide in early 2020, causing hundreds of millions of infections and more than 5 million deaths to date. The absence of a cure and the long hospital admission for more serious cases meant that the disease caused by the virus, COVID-19, took hospitals around the world by storm.

Luckily though, less than a year later vaccines were being distributed and people around the world were getting immunised to the virus at a rapid pace. This process caused a significant drop in the number of active cases around the world in early to mid-2021.

However, viruses, like any replicating entity with a genetic code, change over time. Any time the virus replicates, there is a small chance that a mutation happens in its genetic makeup. Most of the time these mutations are detrimental to the survival of the virus, but in a very small amount of cases, this mutation can be beneficial, increasing the transmissibility and effectiveness of the virus. Any mutated genome is called a variant.

The high worldwide prevalence of SARS-CoV-2 caused it to mutate at an alarming rate, creating different variants, such as the Delta variant in early 2021[1], and more recently, the Omicron variant. These variants spread even more rapidly than the original virus, and more alarming, are more resistant to the vaccines[11].

This makes the tracking of these viruses crucial not only to crisis management teams, for devising government policy, but also to vaccine researchers and other healthcare professionals. The tracking and detection of these variants can be done using DNA sequencing. The viral RNA is first reverse-transcribed back into DNA, which can be read using a sequencing machine and then matched to reference genomes of the variants. However, doing this for individual patients is not only time-consuming but also very expensive, especially if the goal is to test a larger sample size, for example, an entire city or country.

A faster and cheaper way to track variants in population groups is to use wastewater. This has the added benefit that viral information from people that did not get tested (for example, asymptomatic patients) is also included. Viral RNA fragments are secreted by patients and end up in wastewater. The viral information is then filtered out of the wastewater and can be sequenced in one big batch. This sequencing happens in so-called sequencing reads, each of which is a few hundred base-pairs long. An RNA-seq quantification algorithm, which is normally used for quantifying gene expression, can then be used to quantify the prevalence of the different variants in these reads.

This is the main idea behind the pipeline developed by Baaijens et al.[1]. This pipeline uses Kallisto[4], as an RNA-seq quantification algorithm, for variant abundance prediction.

[1] The first sample of the Delta genome was collected in October of 2020, but it was only marked as a Variant of Interest by the WHO in March of 2021.

DNA sequencing machines are far from perfect, however. Different types of errors can sneak into the data, which sometimes are indistinguishable from actual mutations. In rare cases parts of different genomes can also end up on the same read, forming a chimeric read. Wastewater is also highly contaminated with other viruses, which can also influence the algorithm. It is therefore important to determine the impact of these errors on prediction accuracy to assess the reliability of the process.

The main question this research paper will thus answer reads: *What is the impact of sequencing errors and contaminating viruses on SARS-CoV-2 variant prediction accuracy?*

There has not been much research into the impact of sequencing errors on Kallisto. The authors of Kallisto have written that any regular amount of errors should have little to no impact on prediction accuracy[4]. Nevertheless, the precise impact that different types of errors have on prediction, and more specifically SARS-CoV-2 variant detection, remains to be seen. The disparity in Kallisto prediction accuracy between different error types is also unknown.

The effects of contaminants on SARS-CoV-2 variant prediction are also uncharted. [6] does list several viruses that have been reported in wastewater in previous literature. The effect of these contaminants is most likely dependent on the likeness between the contaminant and the prediction target. A virus that is highly similar to SARS-CoV-2 will have a higher impact on prediction accuracy than a virus that is completely different. The exact impact of these viruses, however, remains unknown.

Methodology

The main research question can be divided into two parts, the assessment of the impact of sequencing errors and the assessment of the impact of contaminating viruses. The general idea behind these experiments is laid out in the next sections.

All of the experiments are run with four different variants of concern, alpha (B.1.1.7), beta (B.1.351), gamma (P.1) and delta (B.1.617.2). Notably, omicron is not included in this list, as, at the time of designing the experiments, it had not yet been classified as a variant of concern. In the rest of the paper, variants of concern will often be referred to as VoC.

Sequencing errors

The different types of sequencing errors looked at in this research are *Insertion*, *Deletion*, and *Substitution* errors. *Chimeric reads*, where parts of different genomes end up on the same read, are not strictly sequencing errors, as this type of error is not by fault of the sequencing machine, but rather happens earlier in the process, mostly during PCR[2]. Chimeric reads are included in the experiments nevertheless, as they could arise during some sequencing processes and could influence results.

To assess the impact of each of these three basic sequencing errors, an experiment was devised for each error type. For each of these experiments, the goal was to obtain data to compare prediction accuracy at different error and variant abundance levels. To create a reproducible and verifiable experiment, this was done by simulating the reads with ART[8], a program that can simulate sequencing reads at different abundances and error levels.

For every variant of concern, a certain abundance was simulated by simulating a certain number of VoC reads and a certain number of background reads, which are reads from other SARS-CoV-2 variants. This way only a certain percentage of the reads in the dataset are VoC reads. It is important to note that in a single simulated test sample, only one VoC is present.

This data was thus simulated for every variant of concern and put into one dataset. The shape of the simulated data for each VoC was $VoC\_abundance \times induced\_error$. This dataset was then fed through Kallisto[4] to obtain prediction results, which could then be compared to datasets of different error types.

For the simulation of chimeric reads ART could not be used, as it has no support for simulating chimeric reads. For this dataset, the same procedure as above was used, but SimFFPE[15] was thus used instead of ART.

Wastewater contamination

It is not realistic to test all existing viruses as contaminants. Therefore, for the first experiment, only the viruses listed in [6] are used. [6] lists viruses that were reported to be present in wastewater in recent literature. These viruses are listed in C.1. Reads of these viruses were simulated using ART[8] and merged with simulated SARS-CoV-2 VoC data. By using different ratios of SARS-CoV-2 to contaminant viruses, the impact on prediction error can then be assessed. The goal is to test the impact on VoC abundance, which was kept static at 10% of the total SARS-CoV-2 coverage.

After feeding the simulated data through Kallisto, the resulting data should then give an indication of the effect of these contaminants on prediction accuracy.

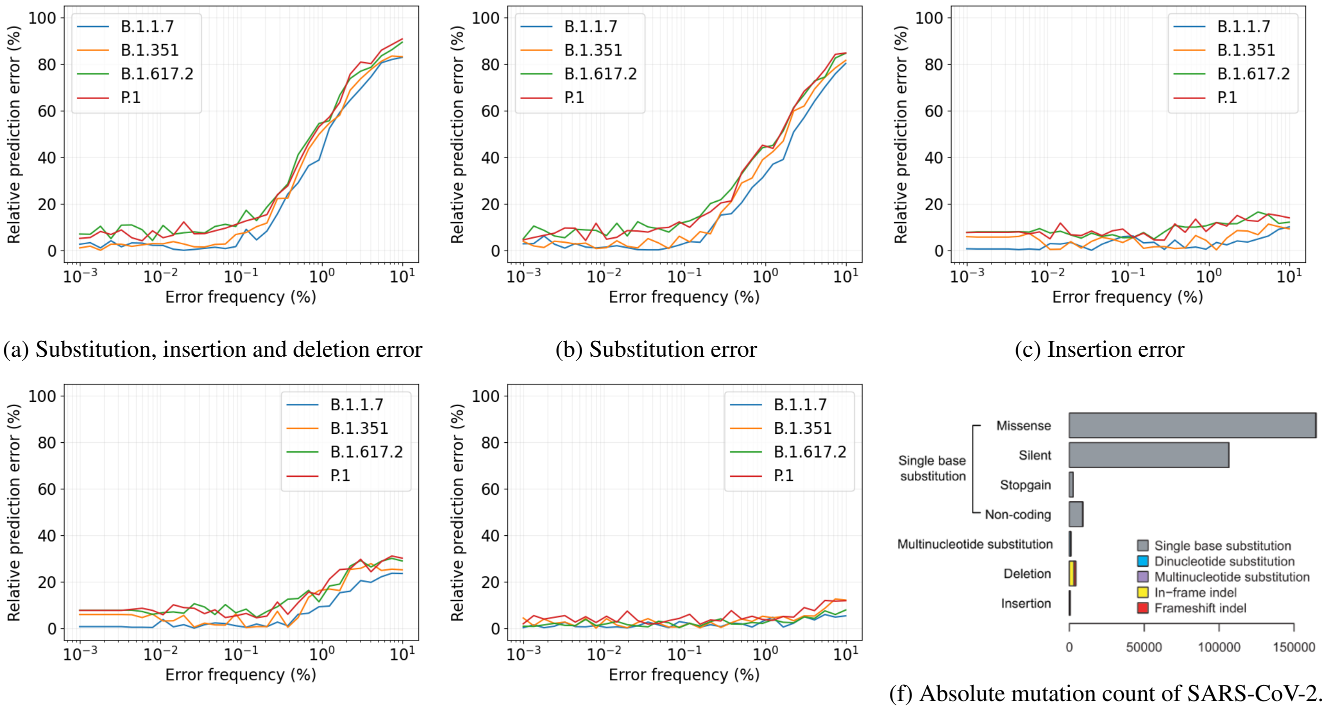
hCoV contamination

It is also important to test the impact of more similar viruses as contaminants. Even though this is a less realistic scenario, it is possibly still vital for understanding the contamination process. A new experiment was therefore designed, with six human coronaviruses as contaminants (Table C.3). This list includes SARS-CoV-1, MERS-CoV and four ”common” human coronaviruses[3]. These common coronaviruses are not as dangerous as the SARS-CoV's or MERS-CoV, but rather mostly cause mild illnesses that represent a common cold. An experiment where these contaminants were present in the reference set was also conducted. The strains in the reference set were not the same exact strains as the contaminants, to prevent 'overfitting' of the contaminants by Kallisto. The reference strains are listed in C.2.

Other than this, the experimental setup is identical to the wastewater contamination experiment.

[2] Polymerase chain reaction, which is used to create a large number of copies of a DNA sample. It is important to note that this process is not used by all DNA sequencing methods.

[3] hCoV-OC43, hCoV-NL63, hCoV-HKU1 and hCoV-229E.



(a) Substitution, insertion and deletion error

(b) Substitution error

(c) Insertion error

(d) Deletion error

(e) Chimeric reads

(f) Absolute mutation count of SARS-CoV-2.

Figure 1: a) b) c) d) e) Relative prediction error (%) plotted against induced error frequency (%) for substitution, insertion and deletion errors, and chimeric reads, plotted at a VoC abundance of 10.8%. f) Absolute mutation count of SARS-CoV-2 taken from 351,525 sequences collected between December 24 2019 and January 12 2021. Adapted from [18].

Results

The effect of errors

The combined effect of the three basic types of sequencing errors, substitution, insertion, and deletion on prediction accuracy is shown in figure 1a. At first, the relative prediction error remains steady at an average prediction error of 5.23%, at an error rate of 0.0108%. However, around an error rate of about 0.2%, the relative prediction error starts rising rapidly. At 10% error, the average relative prediction error reaches 86.3%. These error rates and all other results in this section are based on a VoC abundance of 10.8%.

As can be seen in figure 1, substitution errors have the largest impact on prediction. At an error rate of 1.25%, the average relative prediction error of the substitution dataset is 42.0%, while the insertion and deletion prediction errors are 6.83% and 16.4% respectively.

Figure 1c shows that while the predictions get a little noisier at higher insertion error rates, the prediction doesn't get much worse, going from an average relative prediction error of 3.90% at 0.0108% insertion error to 11.3% at 10.0% insertion error.

This can be compared to the increase in prediction error of the deletion error, figure 1d. This error climbs from an average relative prediction error of 4.80% at 0.0108% deletion error to 26.9% at 10.0% deletion error. This amounts to an increase that is 3.00× as big as the increase in prediction error of insertion errors.

Doing this same calculation on substitution errors (figure 1b), which have an average relative prediction error of 3.39% at 0.0108% substitution error, and 82.8% at 10.0% substitution error, leads to a 10.7× increase over insertion errors and a 3.59× increase over deletion errors.

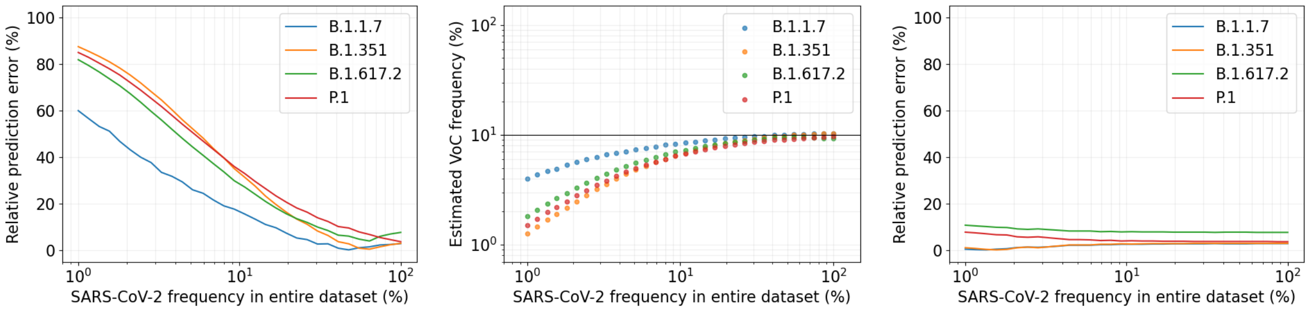
In figure A.1 all of the data points of the four VoC datasets are plotted in scatter plots. It is visible here that the insertion error and deletion error plots within VoC's are very similar, only differing slightly in the spread between error levels. The overestimation is also very similar between all three different error types. However, there are differences in underestimation, not only within VoC's but also between VoC's.

Chimeric reads appear to have no significant impact on prediction accuracy. Figure 1e only shows a minimal increase in relative prediction error, and only when reaching chimeric read frequencies close to 5%.

Wastewater virus contamination

As can be seen in figure 2a, introducing contaminating viruses can have a major impact on prediction performance. Even at a total SARS-CoV-2 frequency of 10.8%, which corresponds to a VoC frequency of 1.08%, the average relative prediction error still amounts to 31.9%. This error can be attributed to an underestimation of the VoC, as can be seen in figure 2b. It is noteworthy to mention that B.1.1.7 performs significantly better than the other VoC's in this experiment.

Most of this error is caused by a single contaminant, namely SARS-CoV-1. As is shown in figure 2c, removing



(a) All contaminants

(b) All contaminants

(c) All contaminants minus SARS-CoV-1

Figure 2: a) b) Relative prediction error (%) plotted against total SARS-CoV-2 frequency (%) in the entire dataset. VoC frequency is kept constant at 10%. Contaminating viruses are *not* present in the reference set. c) Predicted VoC frequency (%) plotted against total SARS-CoV-2 frequency (%) in the entire dataset. Contaminating viruses are *not* present in the reference set.

SARS-CoV-1 from the contaminants yields almost perfect prediction accuracy.

hCoV contamination

As was also found in the last experiment, figure 4a shows that SARS-CoV-1 has a substantial impact on prediction accuracy, with an average relative error rate of 46.0%. hCoV-HKU1 also has a significant impact on prediction accuracy, with an average relative error rate of 23.4%. All other contaminants tested in this experiment have no effect, all having an average relative error rate of 4.22%.

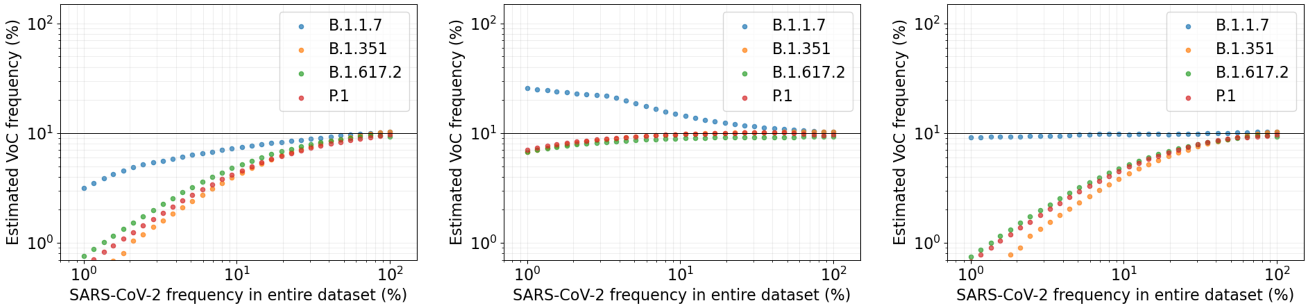
It is necessary to mention that the SARS-CoV-2 frequency as indicated in the graph, is the frequency out of the entire dataset with all six contaminants. In figure 4f the simulated read count is shown. As is visible, the absolute read count of SARS-CoV-2 and of individual contaminants, is simulated to remain equal. Following this, the x-axis on the plots in this section thus corresponds with the SARS-CoV-2 percentage in the entire dataset (figure 4d). This was done to allow for comparison between all graphs in this section.

Figure 4b shows that the results for SARS-CoV-1 as a contaminant corresponds with the results in the last experiment (Figure 2a). B.1.1.7 is better performing than the other variants when SARS-CoV-1 is the only contaminant, with an average relative prediction error of 28.5%, compared to an average of 51.8% for the other variants. This can be attributed to less underestimation of B.1.1.7 compared to the other variants (Figure 3a). As can be seen in table B.1 however, Kallisto aligns a significant percentage more reads than it should in the ideal case. As abundance is expressed in percentage points and non-aligned reads are not taken into account for this calculation, more pseudo-aligned reads will lead to underestimation.

On the contrary, when hCoV-HKU1 is the contaminant, B.1.1.7 performs significantly worse than the other VoC's, as can be seen in figure 4c. B.1.1.7 has an average relative prediction error of 64.0%, compared to the 9.80% of the other three variants averaged. In this case, the other variants perform well, but B.1.1.7 is highly overestimated (Figure 3b). In table B.2 it can be seen that, contrary to the SARS-CoV-1 experiment, the difference between the percentage of reads pseudo-aligned in the experiment and the ideal case is insignificant. The same table also shows a significant misprediction as B.1.1.7 for the other variants.

All other contaminants seem to not affect prediction, as can be seen in figure 4a.

When all contaminants are combined in one dataset (Fig-



(a) SARS-CoV-1

(b) hCoV-HKU1

(c) All contaminants

Figure 3: Estimated VoC frequency (%) plotted against total SARS-CoV-2 frequency (%) in the entire dataset. VoC frequency is kept constant at 10%. Points above the black line are thus overestimated and points below the line are underestimated.

ure 4d), B.1.1.7 performs excellent, with an average relative prediction error of 3.40%. The other variants perform comparable to the SARS-CoV-1 dataset, with an average error of 51.1%. This error is caused by underestimation (Figure 3c). As can be seen in table B.3, for every VoC there is almost an equal percentage of reads qualified as B.1.1.7 as the correct variant.

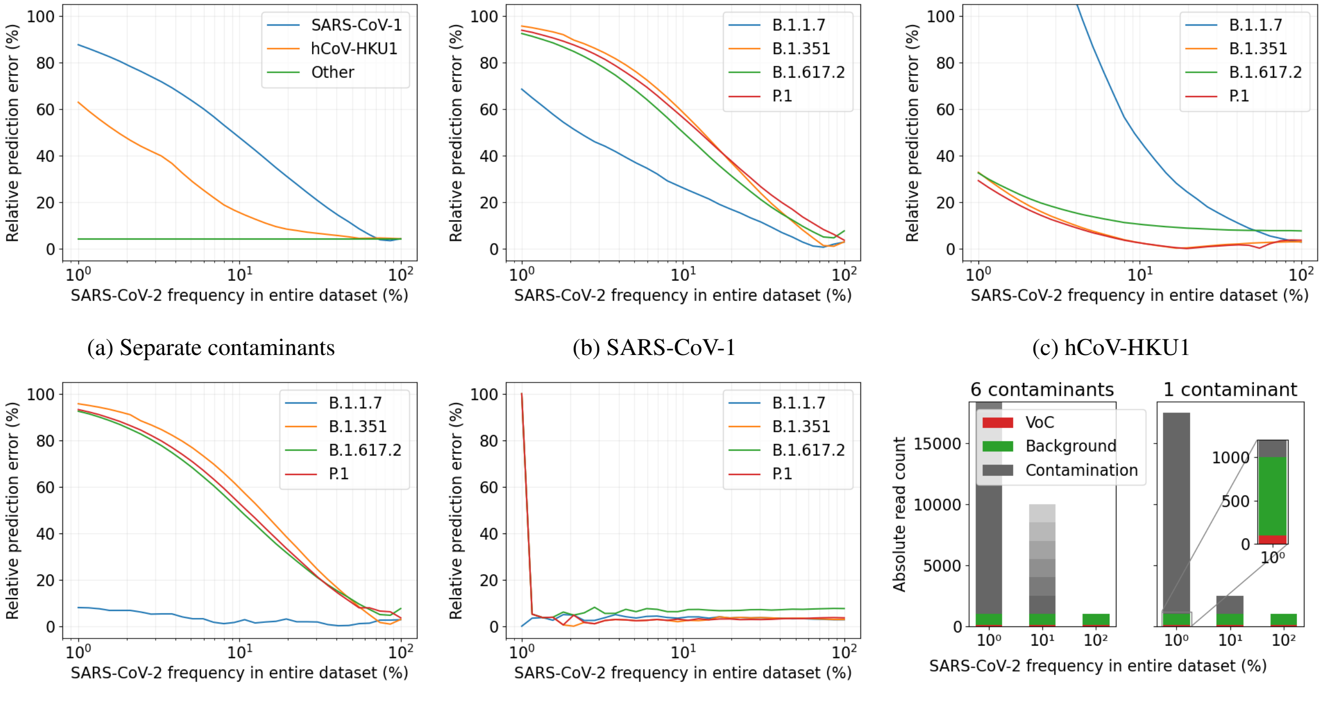
It appears that the good prediction for B.1.1.7 can be explained by the under- and overestimation of this variant in the SARS-CoV-1 and hCoV-HKU1 datasets respectively (Figure 3). As these are present in the same amount in the complete dataset, these seem to cancel each other out and produce an almost perfect prediction for B.1.1.7.

To gain further insights into this anomaly, Kallisto's pseudo-alignment was analysed for all VoC's in the dataset containing all contaminants. This analysis was done at a SARS-CoV-2 percentage of 10.8% and the results are listed in table B.5. It is important to note that the numbers in this table are not comparable with the results in table B.1 to B.3. The results in table B.5 are obtained from the pseudo-alignment step of Kallisto, after which reads can be aligned to multiple genomes and are not yet quantified. The numbers and ratios in this table are thus not directly equal or proportionate to the final prediction results. With that said, the results can give insights on what reads are not being aligned, and, in general, which reads are possibly being aligned to which genome.

From this table, it is first of all clear that hCoV-HKU1 is mostly being aligned to B.1.1.7, but very sparingly. This could explain the overestimation of B.1.1.7 in the hCoV-HKU1 dataset (Figure 3b). The very sporadic alignment of hCoV-HKU1 also matches with the data in table B.2, where the percentage of reads pseudo-aligned is not significantly different from the ideal. The misprediction as B.1.1.7 in the datasets of the other variants can also be explained by this mechanism.

For SARS-CoV-1 the data is less clear. It is evident that there is a significantly bigger number of reads that can be aligned to the VoC's than for hCoV-HKU1. This correlates with the high percentage of aligned reads in table B.1. The alignment of SARS-CoV-1 is also spread out over all different VoC's and the background, unlike hCoV-HKU1, which only aligns to B.1.1.7 and the background. It is, however, not clear enough from this data whether the good prediction of B.1.1.7 can indeed be explained by an overestimation due to misprediction of hCoV-HKU1 and an underestimation due to the excessive number of mispredicted SARS-CoV-1 reads.

Analysing the likeliness between the VoC's and the contaminant coronaviruses gives little clarity. As can be seen in



(a) Separate contaminants

(b) SARS-CoV-1

(c) hCoV-HKU1

(d) All contaminants

(e) All contaminants in reference set

(f) Absolute read count

Figure 4: a) Average relative prediction error (%) plotted against total SARS-CoV-2 frequency (%) in the entire dataset. VoC frequency is kept constant at 10%. Contaminating viruses are *not* present in the reference set. b) c) Relative prediction error (%) plotted against total SARS-CoV-2 frequency in the entire dataset. VoC frequency is kept constant at 10%. Plotted for contaminants SARS-CoV-1 and hCoV-HKU1. d) Relative prediction error (%) plotted against total SARS-CoV-2 frequency (%) in the entire dataset. VoC frequency is kept constant at 10%. Contaminating viruses are *not* present in the reference set. e) Relative prediction error (%) plotted against total SARS-CoV-2 frequency (%) in the entire dataset. VoC frequency is kept constant at 10%. Contaminating viruses are present in the reference set. f) Absolute read count for datasets with six contaminants and datasets with one contaminant.

table B.4, the four common human coronaviruses can not be aligned to the VoC's by bbmap.sh, as they are too different from the VoC genomes. This might mean that there is only a very small part of the hCoV-HKU1 genome that is similar enough to B.1.1.7 to align to it. SARS-CoV-1 can be aligned to the VoC's, but the differences in match percentage between different VoC's are statistically not significant.

Regardless, when the contaminating viruses are introduced into the reference set, they cease to have any significant effect on prediction, as can be seen in figure 4e.

Discussion

In this paper, I have analysed the effect of various sequencing errors and contaminating viruses on SARS-CoV-2 variant quantification prediction accuracy. For sequencing errors, this was done by simulating sequencing reads with varying error levels. The impact of contaminant viruses was tested by adding reads originating from various viruses to the, otherwise normally simulated, dataset. The next sections will go into detail on the impact and applicability of the results.

Sequencing errors

From the results, it shows that, for a VoC frequency of 10.8%, the pipeline remains able to make good predictions up until at least 0.112% error rate, which is the median error rate of the machine this simulation is based upon[14], the Illumina HiSeq 2500. It is, however, important to note that in this experiment the rate of substitution, insertion and deletion errors was equal. This is not true to the real Illumina sequencing machines, for which the rate of substitution errors is several orders of magnitude higher than the rate of insertion and deletion errors[13]. This means that the experiment which just looks at substitution errors (figure 1b) is probably a better indicator of the real world, at least when using Illumina sequencing machines.

The Illumina HiSeq 2500 is, however, one of the better performing sequencing machines, in terms of error rate, of the machines listed in [14]. This means that while this machine might be able to make acceptable predictions at its median error rate, other, less sophisticated, sequencing machines may not be. At the rate the industry is moving forward, however, technical advances could make this a non-issue in a couple of years. The Illumina HiSeq X Ten, which at the time of writing is already 8 years old, already has a median error rate of 0.087%; however, when using a less accurate sequencing machine or an entirely different sequencing method, the error rate can be as high as 1%. In these cases, the prediction error of the pipeline would be a detriment to performance.

It is also important to note that the results mentioned above are observed at a VoC frequency of 10.8%. This number was chosen to nicely accentuate the differences between the different error types. It is, however, not directly applicable to all variant prediction use cases. When the goal is to predict newer, upcoming variants, which could have a variant abundance of 1% or lower, the error graph might look more like figure 5a. The lower variant abundance seems to not drastically influence the average prediction error, but it does make the results substantially noisier and thus less trustworthy.

When the goal is to predict variants of concern though, the error plot might look more like figure 5c. Variants like delta and omicron have, in the USA, reached variant abundances surpassing 95%[4], at which point the noise in the prediction results disappears and results are, as long as error levels are kept in check, highly accurate.

It is therefore important to, on a case by case basis, assess the credibility of the prediction results. The plots in figure A.1 could be of help when doing this.

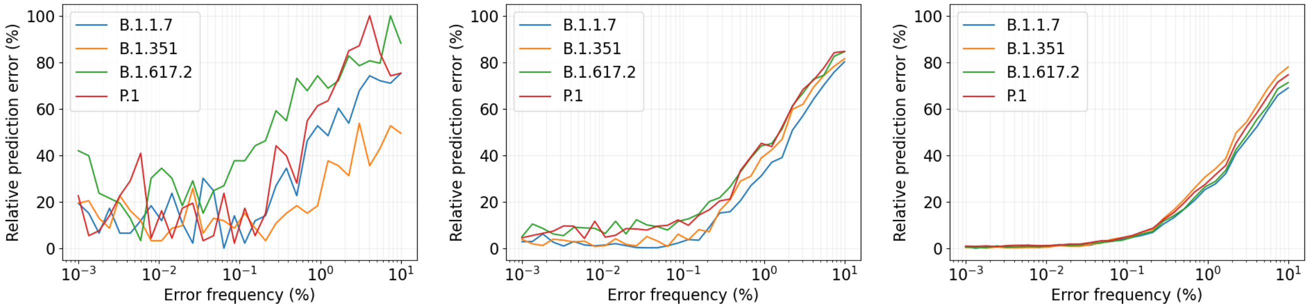
Differences between error types

The results show that, at least for SARS-CoV-2 variant prediction, substitution errors have the highest impact by far. The impact of deletion errors is still noticeable, but the impact of insertion errors and chimeric reads is minimal.

It is also clear that there are slight differences in over- and underestimation between the different VoC's (figure A.1). This is possibly due to the selection of the background genomes. When there is more similarity between the background genomes and the VoC, Kallisto might be more likely to wrongly qualify some reads. Regardless, the exact cause of the difference remains to be seen.

It is also important to zoom in on the experiments and look

[4] https://covariants.org/per-country



(a) 1.16% VoC frequency

(b) 10.8% VoC frequency

(c) 80.0% VoC frequency

Figure 5: Relative prediction error (%) plotted against induced substitution error frequency (%). Plotted at VoC abundances of 1.16%, 10.8% and 80.0%.

at the impact of errors on the Kallisto algorithm[4]. Kallisto computes the compatibility of a read to a genome by hashing groups of base-pairs, so-called k-mers. These hashed k-mers are then aligned to the hashed k-mers of the reference genome, after which redundant k-mers are skipped. Redundant k-mers are, for example, k-mers which are the same for all reference genomes. The inclusion of these k-mers would thus not change results. Utilising this hashing and skipping of redundant k-mers, Kallisto can be incredibly fast while still producing highly accurate prediction results.

Due to this hashing process though, even a single error in a k-mer would change the value of the entire k-mer. One would think that this would then upset the prediction of the entire read. From the results, it turns out that this is not necessarily the case. According to the creators of Kallisto, this can be contributed to the fact that there is a high probability that an error in a k-mer will cause the k-mer to not match with any k-mer in the reference genomes. This causes the k-mer to simply be ignored, and Kallisto thus only produces a possibly less accurate, but not necessarily completely wrong, prediction. Kallisto's authors also mention that the skipping of redundant k-mers also helps ameliorate errors, as there is a significant chance that an error would end up in one of these skipped k-mers. [4]

All of this means that most errors will be ignored by Kallisto. Errors that correspond with a mutation in a reference genome, however, will cause the read to look like, and align to, this reference genome, introducing prediction errors.

The difference between the impact of different error types could thus be explained by the mutational spectrum of SARS-CoV-2 variants. As can be seen by comparing the error plots in figure 1 to figure 1f, there is a correlation between the relative impact of different sequencing errors and the abundance of the corresponding mutation in the mutations of the variants. As errors that correspond with mutations in the reference set can not be ignored by Kallisto, this could be the mechanism behind the disparity in results between different error types. For example, as the number of substitution mutations is substantially higher than the number of insertion mutations in the variants, a substitution error in a read has a significantly higher chance to "mimic" a substitution mutation in a different sequence in the dataset thus causing a misprediction.

Further research will however have to be done to verify if this correlation is indeed also causation. To test for this, one could check if this correlation also exists for a different virus with a different mutational spectrum. One could also simulate new SARS-CoV-2 "variants" with a different mutational spectrum to the current variants and compare the results. Care will have to be taken in making sure that the mutational spectrum between the new variants is similar, with not only similar mutation frequencies but also similar mutation sites, just like the real virus.

Furthermore, further research could be done on the impact of different substitution error types. Substitution mutations in SARS-CoV-2 are heavily favoured towards $C > U$ and, slightly less heavily, towards $G > U$ mutations. Adding this bias in the simulated data could give new insights into, for example, the impact of the mutational spectrum of the virus on the effect of errors.

Whether the selected background sequences are an accurate reflection of the real world can also be up for debate. Care was taken to ensure that all background sequences were collected in Connecticut, USA, around the same time as the VoC sequences. All sequences which belong to the same variant group as the VoC's were removed. However, no detailed assessment of the likeliness between the VoC's and the selected background set was done.

Contamination

From the results, it appears that most viruses normally present in wastewater have no significant impact on prediction accuracy. Viruses that bear a significant likeness to the predicted virus may, however, to a greater or lesser extent, impact prediction. It is clear that when there is a virus that is similar enough, its impact on prediction accuracy is substantial.

The only two viruses that had a significant effect on prediction error in the experiments in this paper are SARS-CoV-1, which is part of the same species as SARS-CoV-2, and hCoV-HKU1, for which only a small part of the genome seems to be similar enough to SARS-CoV-2 to influence results.

The mechanism of the error of these two viruses is slightly different, however. It seems that in the SARS-CoV-1 dataset all VoC's were underestimated due to the high abundance of (miss)aligned SARS-CoV-1 reads. Contrary, for the hCoV-HKU1 dataset, there is a little underestimation, due to the small amounts of hCoV-HKU1 reads that can be aligned to reference genomes. B.1.1.7 is however highly overestimated, most likely due to the aligning of this small amount of hCoV-HKU1 reads to B.1.1.7.

It is however clear that in both cases B.1.1.7 is predicted to have a higher abundance than the other VoC's, possibly due to a higher likeliness of the B.1.1.7 reference genomes to both hCoV-HKU1 and SARS-CoV-2. Further research will however have to be done to confirm this.

The great performance of B.1.1.7 in the dataset containing all human corona-viruses as contaminants seems like a great anomaly compared to the other VoC's. It is however highly likely that the overestimation due to hCoV-HKU1 and the underestimation due to SARS-CoV-1, as can be seen in their respective separate datasets, cancel each other out for B.1.1.7. As the other VoC's are underestimated in both datasets, this error is not reduced, but rather amplified.

It is important to note that for this experiment all sequencing errors were disabled, to assess the raw influence of the contaminants on prediction. The absence of errors explains the smooth lines in all graphs, as there is considerably less randomness, and thus less noise, in the dataset. In the real world, however, errors can not be disabled. The direct applicability of these results is therefore questionable, as results might improve or worsen when errors are introduced. However, the results in this paper are a good indicator of the kinds of viruses one should watch out for and the possible impact of these viruses on the Kallisto algorithm.

It is also noteworthy to mention that the contaminant in this experiment with the biggest impact on prediction error, SARS-CoV-1, is a virus that infected a relatively low amount of people. There have also been no confirmed cases since

May of 2004[17]. There is thus a negligible chance that SARS-CoV-1 is present in wastewater today.

However, hCoV-HKU1 is a highly prevalent virus. Analysis of blood serum samples in Canada resulted in antibodies against hCoV-HKU1 being found in 60.7% of samples. The virus has also been identified in wastewater before[3]. Even though SARS-CoV-1 theoretically has a higher impact on prediction accuracy, hCoV-HKU1 will have a higher real-world impact.

An important takeaway from this experiment is that when a reference genome for the contaminating viruses is added to the reference set, the contaminants seize to have any effect on prediction error. This means that when the contaminating viruses in the wastewater to be tested are known, the problem of this contamination can greatly be ameliorated.

It needs to be kept in mind, however, that in this experiment only a single reference sequence and a single, different, test sequence were used. Using multiple sequences for both reference and test, as is the case for SARS-CoV-2 in the experiments, could change results. However, as the differences between different viruses will be greater than the differences within viruses, the results in this paper should hold up. Further research will have to confirm this.

Conclusion

To conclude, the credibility of the pipeline depends on the sequencing machine and method used. For most methods used today, the error rate lies between 0.1% and 1%. At the lower end of this range, sequencing errors should have little impact on prediction error. When the error rate of the sequencing method used lies in the upper end of the range, however, the sequencing errors produced could be detrimental to performance. It is therefore of high importance to, on a case by case basis, evaluate the error levels cultivated during the sequencing process and, following that, decide on the credibility and validity of the prediction results.

Contamination is more likely to be a problem. hCoV-HKU1 presence in the dataset can influence prediction and this virus is highly prevalent. Adding this virus to the reference set was however enough to remove the effect of this contaminant. It is therefore important to analyse the prevalence of viruses that are similar to SARS-CoV-2 in the testing data. Adding these viruses to the reference set could be a simple way of making testing results more realistic and trustworthy.

Responsible Research

It was not feasible to upload all benchmarks for public access. Per the Netherlands Code of Conduct for Research Integrity[10] section 3.3:25 however, all code used in the experimentation process can be found on GitHub[5]. Together with the usage of Snakemake[12], this can be used to reconstruct all benchmarks. The Snakemake setup and configuration for all experiments can be found in the folder *experiments*.

All genome data used in this research is publicly available on GISAID[7][6] and GenBank[2]. For all experiments, the data used is also listed on GitHub under the folder *experiments*. It was not possible to test all data used from these sources for legitimacy, but both sources are credible, widely trusted and heavily used by other researchers. All data on these platforms is also uploaded by registered and trustworthy laboratories around the globe.

Care was taken to ensure conformity with the Netherlands Code of Conduct for Research Integrity section 3.4:37,38, which states that you should be clear about results, conclusions and uncertainties, as well as their scope, and you should not draw unsubstantiated conclusions[10].

It is therefore also important to disclose that all the experiments in this paper are based on SARS-CoV-2 data from the USA. Any conclusions and statements made might thus not be generalisable to other regions and viruses.

[5] https://github.com/MartLugt/wastewater_analysis

[6] Requires registration.

Methods

For all research in this paper, the pipeline developed by Baaijens et al.[1] was used. This pipeline uses Kallisto[4] for variant abundance prediction. Snakemake[12] was used for workflow management.

Genome data sources

SARS-CoV-2 genome data was obtained from GISAID[7]. The VoC lineages and their respective GISAID accession id's used are; B.1.1.7 (EPI_ISL_1008906), B.1.351 (EPI_ISL_1001460), B.1.617.2 (EPI_ISL_1924762) and P.1 (EPI_ISL_1194849). Other viral genomes were sourced from the US National Library of Medicine[2]. The viruses and respective accession numbers can be found in table C.1 and C.3. The accession id's of all viruses used can also be found on GitHub[7] under the folder *experiments*.

Reference set construction

The reference set includes SARS-CoV-2 genome data collected in the USA between *2020-10-01* and *2021-09-24*. The reference set was constructed using the pipeline described in [1]. First, a quality filter was applied to the sequences. Variants were called compared to the original SARS-CoV-2 reference genome (WIV04 [19][8]). Sequences were then selected per lineage such that all mutations with an allele frequency of at least 50% were captured at least once. This thus means that not all sequences are captured in the reference set, and the reference set is significantly smaller than the full dataset. All lineages included in the reference set are listed on GitHub[9] under the folder *experiments/reference_set*. A Kallisto index was then created out of this reference set.

Background genome selection

Background sequences were collected from the full dataset. Only sequences collected in Connecticut, USA were considered as background sequences. VoC sequences and VoC sub-lineages were excluded. All sequences included in the background set are listed on GitHub[10] under the folder *experiments/background*.

Insertion error benchmark construction

For the construction of the insertion error benchmark, ART Illumina[8] was used to simulate reads. For every variant of concern – at the time of running the experiment[11]: B.1.1.7, B.1.351, B.1.617.2, P.1 – *32* different VoC abundance percentages and *32* different insertion error percentages were simulated, both equally spaced on a logarithmic scale and ranging from 10^{-1} to 10^{2} and from 10^{-3} to 10^{1} respectively. The rest of the dataset was filled up with the background sequences. A total fold coverage of 1000 was used. Paired-end reads were simulated with a read length of 150, a median fragment size of 250, and a fragment size standard deviation of 10. The ART quality profile[12] used was *HS25*. The resulting reads were shuffled.

Deletion error benchmark construction

For the construction of the deletion error benchmark the same general procedure as for the insertion error benchmark was followed. Instead of insertion errors, deletion errors were simulated.

Substitution error benchmark construction

For the construction of the substitution error benchmark the same general procedure as for the insertion error and deletion error benchmark was followed. Instead of insertion or deletion errors, substitution errors were simulated.

For substitution errors, ART doesn't take a single parameter, but a quality profile, which can be manipulated by using the *quality_shift* parameter. The ART quality profile[13] used was *HS25*, which is based on the Illumina HiSeq 2500. This machine has a median error rate of 0.112%[14][14]. To take this into account the ART quality shift was calculated using the following formula:

$$quality\_shift = 10 \times \log_{10} \frac{0.112}{error(\%)}$$

After verifying some of the resulting reads with BBMap[5] this proved to indeed produce reads with the correct substitution error rate.

Combined error benchmark construction

For the construction of the combined error benchmark the same general procedure as for the other error benchmarks was followed. Instead of simulating a single error type all errors were simulated at the same time, all at the same error percentage in the range from 10^{-3} to 10^{1}.

Chimeric read benchmark construction

For the construction of the chimeric read benchmark, the R package SimFFPE[15] was used. This package allows for the simulation of chimeric reads[16]. The same procedure as for the InDel and substitution benchmark was followed. The SimFFPE parameters *chimMutRate*, *noiseRate*, and *highNoiseRate* were all set to 0 to disable simulated sequencing errors. This means that any arbitrary *PhredScoreProfile*[15] can be used, as it will have no effect. For this experiment, the *"PhredScoreProfile1.txt"* included in SimFFPE was used. A total fold coverage of 1000 was used. Paired-end reads were simulated with a read length of 150, a median fragment size of 250, and a fragment size standard deviation of 10. For the SimFFPE reference a file with all background sequences and all VoC's was used, which means that SimFFPE is able to generate distant chimeric reads between VoC's and background sequences, and also between two different VoC's.

[7] https://github.com/MartLugt/wastewater_analysis/

[8] In other literature, NC_045512.2 is often used. This is the same genome with a different identifier.

[9] See footnote 7.

[10] See footnote 7.

[11] Omicron (B.1.1.529) is not present as it had not yet been discovered.

[12] A way of setting different substitution levels for different bases and genome locations.

[13] See footnote 12.

[14] It is important to note that this is the median error rate of all errors, not just substitution errors. However, for most Illumina sequencing machines, the substitution error rate is several orders of magnitude higher than the insertion and deletion rate[13]. Therefore it can be said that $substitution\_rate \approx error\_rate$.

[15] Comparable to an ART quality profile. See footnote 12.

Running sequencing error benchmarks

Benchmarks were fed through Kallisto[4], using the index created out of the reference set. This process was done on the TU Delft HPC.

Contamination reference set construction

For the construction of the contamination reference set, the same general procedure as for the default reference set was followed. Before the Kallisto index was created the genomes for the contaminants (C.2) were added to the reference sets. All lineages included in the reference set are listed on GitHub[16] under the folder *experiments/reference_set*. The Kallisto index was then created out of the whole reference set, including the contaminants.

Wastewater viruses contamination benchmark construction

For the construction of the wastewater contamination benchmark ART[8] was used for sequencing read simulation. Paired-end reads were simulated with a read length of 150, a median fragment size of 250, and a fragment size standard deviation of 10. The contamination dataset was created in three parts; the contaminants, the VoC's, and the SARS-CoV-2 background sequences. The same VoC's and SARS-CoV-2 background sequences as for the sequencing error benchmarks were used. A total SARS-CoV-2 fold coverage of 1000 was used, with a constant VoC abundance of 10%, which is thus a fold coverage of 100. The contaminants used are listed in table C.1. The fold coverage of these 16 contaminants was varied such that the total SARS-CoV-2 frequency varied from 10^0 to 10^2 by using the formula listed below.

$$total\_sars2 \times \frac{1}{sars2\_rate} - total\_sars2$$

For this benchmark all ART simulated sequencing errors were disabled.

This same benchmark was created without SARS-CoV-1 as a contaminant, keeping the fold coverage of the other contaminants and SARS-CoV-2 equal. This is also shown in figure 4f.

hCoV contamination benchmark construction

The same general procedure as for the wastewater viruses contamination benchmark construction was used. The viruses listed in C.3 were used as contaminants.

A benchmark was created containing all contaminating viruses. Six, otherwise identical, benchmarks were created, each with a single, different, contaminating virus. The fold coverage of independent viruses was kept identical between benchmarks. Benchmarks with fewer contaminants thus have a lower total fold coverage. This is also highlighted in figure 4f.

Running contamination benchmarks

Contamination benchmarks were fed through Kallisto[4] twice, once using the default index (excluding the contaminating genomes), and once using the contamination index (including the contaminating genomes). This process was done on the TU Delft HPC.

Plotting

Data was plotted using Matplotlib[9].

Genome likeness analysis

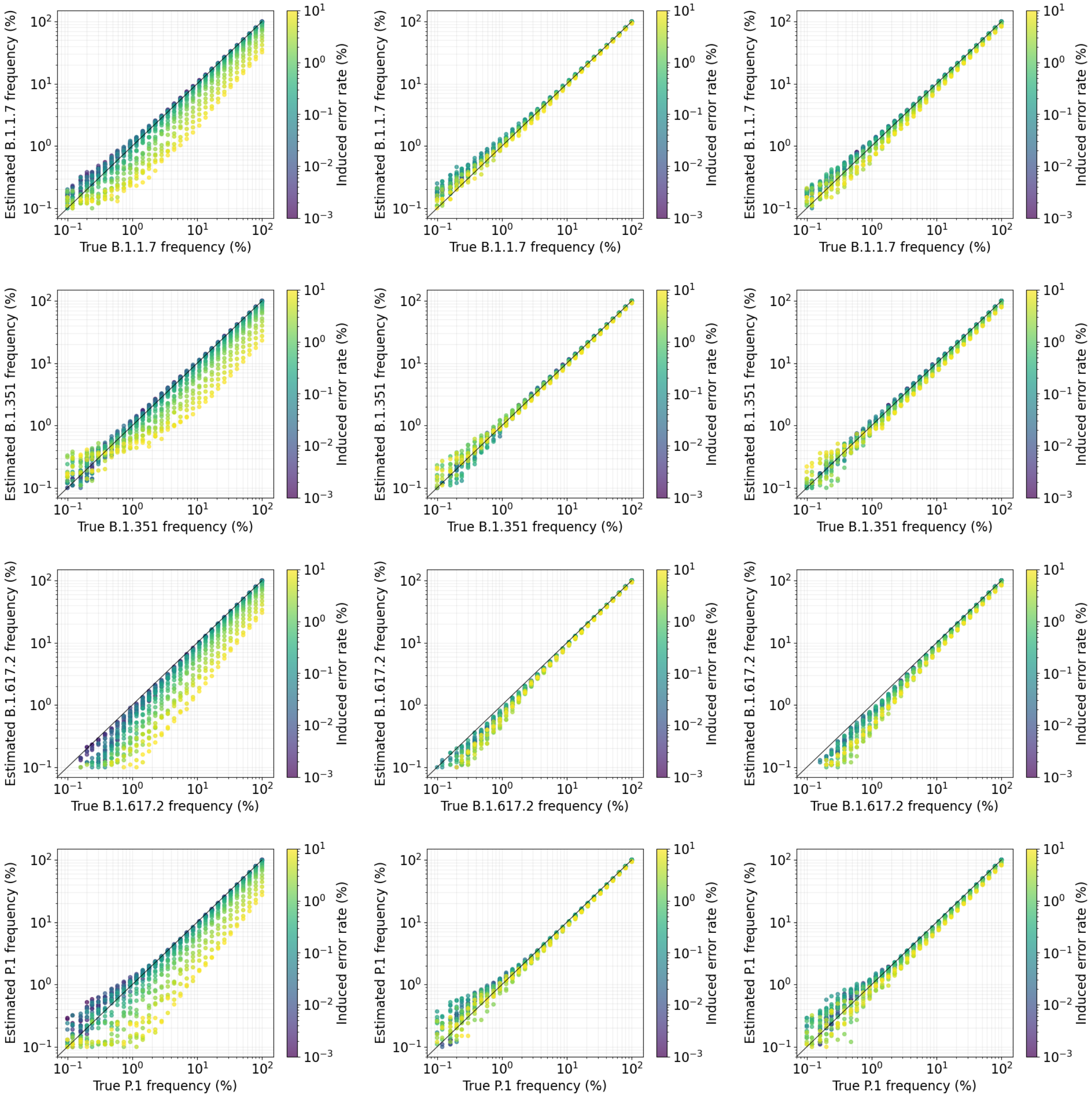
The likeliness between genomes was calculated using bbmap.sh[5].

Pseudo-alignment analysis

For the pseudo-alignment analysis, the alignment was first extracted by using the Kallisto *pseudobam* parameter. This file was then analysed line by line. For every alignment, the lineage the read originated from and the alignment target were determined. The count was kept for every possible alignment.

[16] See footnote 7.

A Supplementary figures



(a) Substitution error

(b) Insertion error

(c) Deletion error

Figure A.1: All datapoints of the substitution, insertion, and deletion datasets for B.1.1.7, B.1.351, B.1.617.2, P.1. Predicted VoC frequency (%) is plotted against true VoC frequency (%). Error rate (%) is depicted by colour. Points under the black line are thus underestimated, points above this line are overestimated.

B Supplementary tables

Table B.1: Average Kallisto prediction results for all VoC's with SARS-CoV-1 as contaminant.

Variant	% of reads psuedoaligned	Abundance correct (%)	Abundance B.1.1.7 (%)	Abundance other (%)
B.1.1.7	61.3	7.18	-	0.920
B.1.351	61.2	4.70	0.212	0.943
B.1.617.2	61.2	5.09	0.154	0.940
P.1	61.2	4.68	0.000	0.946
Ideal	45.7	10.0	0.000	0.000

Table B.2: Average Kallisto prediction results for all VoC's with hCoV-HKU1 as contaminant.

Variant	% of reads psuedoaligned	Abundance correct (%)	Abundance B.1.1.7 (%)	Abundance other (%)
B.1.1.7	45.6	16.4	-	0.833
B.1.351	45.6	9.32	1.57	0.887
B.1.617.2	45.6	8.63	1.61	0.893
P.1	45.6	9.30	1.60	0.888
Ideal	45.7	10.0	0.000	0.000

Table B.3: Average Kallisto prediction results for all VoC's with all viruses listed in table 4d as contaminants.

Variant	% of reads psuedoaligned	Abundance correct (%)	Abundance B.1.1.7 (%)	Abundance other (%)
B.1.1.7	26.4	9.73	-	0.894
B.1.351	26.4	4.63	4.65	0.899
B.1.617.2	26.4	5.08	4.57	0.897
P.1	26.4	4.96	4.63	0.897
Ideal	22.4	10.0	0.000	0.000

Table B.4: Full genome likeliness (%)

	B.1.1.7	B.1.351	B.1.617.2	P.1
WIV04	99.66	99.44	99.68	99.46
SARS-CoV-1	86.05	86.04	86.03	86.02
MERS-CoV	0.000	0.000	0.000	0.000
hCoV-OC43	0.000	0.000	0.000	0.000
hCoV-NL63	0.000	0.000	0.000	0.000
hCoV-HKU1	0.000	0.000	0.000	0.000
hCoV-229E	0.000	0.000	0.000	0.000

Table B.5: Kallisto pseudo-alignment counts for simulation files containing a VoC, background sequences and all viruses listed in 4d. These viruses, listed in the leftmost column, are aligned to the viruses listed in the header row. Please note that these are just the results from Kallisto's pseudo-alignment step, and are thus not comparable to the results in table B.1 to B.3.

(a) B.1.1.7

	B.1.1.7	B.1.351	B.1.617.2	P.1	Background	Not aligned	% aligned
B.1.1.7	19660	13130	0	0	1220502	0	100.0
SARS-CoV-1	38182	25507	0	0	2519017	192444	93.07
MERS-CoV	0	0	0	0	0	276200	0.00
hCoV-OC43	0	0	0	0	0	280344	0.00
hCoV-NL63	0	0	0	0	0	252724	0.00
hCoV-HKU1	1898	0	0	0	1050	272922	1.069
hCoV-229E	0	0	0	0	0	249962	0.00
Background	120928	126214	0	0	12086998	870	99.99

(b) B.1.351

	B.1.1.7	B.1.351	B.1.617.2	P.1	Background	Not aligned	% aligned
B.1.351	13280	35990	0	0	1313772	0	100.0
SARS-CoV-1	26244	51014	0	0	2555333	192444	93.19
MERS-CoV	0	0	0	0	0	276200	0.00
hCoV-OC43	0	0	0	0	0	280344	0.00
hCoV-NL63	0	0	0	0	0	252724	0.00
hCoV-HKU1	1898	0	0	0	1050	272922	1.069
hCoV-229E	0	0	0	0	0	249962	0.00
Background	120928	257872	0	0	12186922	870	99.99

(c) B.1.617.2

	B.1.1.7	B.1.351	B.1.617.2	P.1	Background	Not aligned	% aligned
B.1.617.2	12022	12412	19388	0	1220500	0	100.0
SARS-CoV-1	26244	25507	27591	0	2555347	192444	93.19
MERS-CoV	0	0	0	0	0	276200	0.00
hCoV-OC43	0	0	0	0	0	280344	0.00
hCoV-NL63	0	0	0	0	0	252724	0.00
hCoV-HKU1	1898	0	0	0	1050	272922	1.069
hCoV-229E	0	0	0	0	0	249962	0.00
Background	120928	126214	120018	0	12209934	870	99.99

(d) P.1

	B.1.1.7	B.1.351	B.1.617.2	P.1	Background	Not aligned	% aligned
P.1	12606	13238	0	19300	1204046	0	100.0
SARS-CoV-1	26244	25507	0	24392	2477755	192444	92.99
MERS-CoV	0	0	0	0	0	276200	0.00
hCoV-OC43	0	0	0	0	0	280344	0.00
hCoV-NL63	0	0	0	0	0	252724	0.00
hCoV-HKU1	1898	0	0	0	1050	272922	1.069
hCoV-229E	0	0	0	0	0	249962	0.00
Background	120928	126214	0	122026	11808676	870	99.99

C Accession numbers

Table C.1: Viruses and corresponding accession no's used in the first contamination experiment. [6]

Virus		Accession No.[2]
HEV	Hepatitis E virus, complete genome.	NC_001434.1
RoV-A	Rotavirus A segment 4, complete genome.	NC_011510.2
HAdV	Human adenovirus 2, complete genome.	AC_000007.1
NoV-GI	Norovirus GI, complete genome.	NC_001959.2
NoV-GII	Norovirus GII, complete genome.	NC_039477.1
HAV	Hepatitis A virus, complete genome.	NC_001489.1
EV	Human enterovirus D, complete genome.	NC_001430.1
JCPyV	JC polyomavirus strain #2, complete genome.	MF662181.1
PMMV	Pepper mild mottle virus, complete genome.	NC_003630.1
HPyV	Human papillomavirus type 16, complete genome.	NC_001526.4
Sapovirus	Sapovirus Mc10, complete genome.	NC_010624.1
BK Polyomavirus	BK polyomavirus, complete genome.	NC_001538.1
AiV	Influenza A virus(A/chicken/Soc Trang/5/2012(H5N1)) viral cRNA, segment 2, complete genome.	AB818503.1
Parechovirus	Human parechovirus 1, complete genome.	FM178558.1
Reovirus	Mammalian orthoreovirus 3 segment S4, complete genome.	NC_013234.1
SARS-CoV-1	SARS coronavirus Tor2, complete genome.	NC_004718.3

Table C.2: Viruses and corresponding accession no's used in the HCoV contamination experiment in the reference set.

Virus		Accession No.[2]
SARS-CoV-1	SARS coronavirus Tor2, complete genome.	NC_004718.3
MERS-CoV	Middle East respiratory syndrome coronavirus isolate, MERS-CoV/KOR/KNIH/002_05_2015, complete genome.	KT029139.1
HCoV-HKU1	Human coronavirus HKU1 isolate SI17244, complete genome	MH940245.1
HCoV-229E	Human coronavirus 229E isolate HCoV-229E/BN1/GER/2015, complete genome	KU291448.1
HCoV-NL63	Human Coronavirus NL63, complete genome	NC_005831.2
HCoV-OC43	Human coronavirus OC43 isolate MDS16, complete genome	MK303625.1

Table C.3: Viruses and corresponding accession no's used in the HCoV contamination experiment.

Virus		Accession No.[2]
SARS-CoV-1	SARS coronavirus Tor2 isolate Tor2/FP1-10895, complete genome	JX163928.1
MERS-CoV	Middle East respiratory syndrome-related coronavirus isolate D1189.5, complete genome	MW545527.1
HCoV-HKU1	Human coronavirus HKU1 isolate Caen1, complete genome	HM034837.1
HCoV-229E	Human coronavirus 229E strain 229E/China/01/2009, complete genome	MW532106.1
HCoV-NL63	Human coronavirus NL63 strain ChinaGD04, complete genome	MK334047.1
HCoV-OC43	Human coronavirus OC43 strain HZ-459, complete genome	MG197723.1

References

[1] Jasmijn A Baaijens, Alessandro Zulli, Isabel M Ott, Mary E Petrone, Tara Alpert, Joseph R Fauver, Chaney C Kalinich, Chantal B F Vogels, Mallery I Breban, Claire Duvallet, Kyle McElroy, Newsha Ghaeli, Maxim Imakaev, Malaika Mckenzie-Bennett, Keith Robison, Alex Plocik, Rebecca Schilling, Martha Pierson, Rebecca Littlefield, Michelle Spencer, Birgitte B Simen, Yale SARS-CoV-2 Genomic Surveillance Initiative, William P Hanage, Nathan D Grubaugh, Jordan Peccia, and Michael Baym. Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-Seq quantification. *medRxiv*, page 2021.08.31.21262938, 1 2021.

[2] Bethesda (MD): National Library of Medicine (US) and National Center for Biotechnology Information; [1988]. National Center for Biotechnology Information (NCBI)[Internet]. *Available from: https://www.ncbi.nlm.nih.gov/.*

[3] Kyle Bibby and Jordan Peccia. Identification of Viral Pathogen Diversity in Sewage Sludge by Metagenome Analysis. *Environmental Science & Technology*, 47(4):1945–1951, 2 2013.

[4] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016.

[5] Brian Bushnell. BBMap: A Fast, Accurate, Splice-Aware Aligner. *https://www.osti.gov/biblio/1241166*, 2014.

[6] Mary Vermi Aizza Corpuz, Antonio Buonerba, Giovanni Vigliotta, Tiziano Zarra, Florencio Ballesteros, Pietro Campiglia, Vincenzo Belgiorno, Gregory Korshin, and Vincenzo Naddeo. Viruses in wastewater: occurrence, abundance and detection methods. *Science of The Total Environment*, 745:140910, 11 2020.

[7] Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1(1):33–46, 1 2017.

[8] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)*, 28(4):593–594, 2 2012.

[9] J D Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[10] KNAW, NFU, TO2-federatie, Vereneging Hogescholen, and VSNU. Nederlandse gedragscode wetenschappelijke integriteit. . *DANS*, 2018.

[11] Jamie Lopez Bernal, Nick Andrews, Charlotte Gower, Eileen Gallagher, Ruth Simmons, Simon Thelwall, Julia Stowe, Elise Tessier, Natalie Groves, Gavin Dabrera, Richard Myers, Colin N.J. Campbell, Gayatri Amirthalingam, Matt Edmunds, Maria Zambon, Kevin E. Brown, Susan Hopkins, Meera Chand, and Mary Ramsay. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *New England Journal of Medicine*, 385(7):585–594, 8 2021.

[12] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with Snakemake. *F1000Research*, 10:33, 1 2021.

[13] Melanie Schirmer, Rosalinda D'Amore, Umer Z. Ijaz, Neil Hall, and Christopher Quince. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, 17(1):125, 12 2016.

[14] Nicholas Stoler and Anton Nekrutenko. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics*, 3(1), 1 2021.

[15] Lanying Wei. SimFFPE: NGS Read Simulator for FFPE Tissue, 2021.

[16] Lanying Wei, Martin Dugas, and Sarah Sandmann. SimFFPE and FilterFFPE: improving structural variant calling in FFPE samples. *GigaScience*, 10(9), 9 2021.

[17] WHO. China's latest SARS outbreak has been contained, but biosafety concerns remain. *https://web.archive.org/web/20200212205529/ https://www.who.int/csr/don/2004_05_18a/en/.*

[18] Kijong Yi, Su Yeon Kim, Thomas Bleazard, Taewoo Kim, Jeonghwan Youk, and Young Seok Ju. Mutational spectrum of SARS-CoV-2 during the global pandemic. *Experimental & Molecular Medicine*, 53(8):1229–1237, 8 2021.

[19] Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, Hui-Dong Chen, Jing Chen, Yun Luo, Hua Guo, Ren-Di Jiang, Mei-Qin Liu, Ying Chen, Xu-Rui Shen, Xi Wang, Xiao-Shuang Zheng, Kai Zhao, Quan-Jiao Chen, Fei Deng, Lin-Lin Liu, Bing Yan, Fa-Xian Zhan, Yan-Yi Wang, Geng-Fu Xiao, and Zheng-Li Shi. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798):270–273, 3 2020.