Delft University of Technology

Bachelor Thesis
CSE3000

Effect of Output Granularity on SARS-CoV-2 Variant Abundance Estimates using Domestic Wastewater Sequencing.

Author:
Yash Kalia

A thesis submitted to the Faculty of Electrical Engineering, Mathematics and Computer Science at Delft University of Technology for the Degree of Bachelor's of Science in Computer Science and Engineering.

January 24, 2022

Effect of Output Granularity on SARS-CoV-2 Variant Abundance Estimates using Domestic Wastewater Sequencing.

Yash Kalia

November 2021

Abstract

Monitoring of SARS-CoV-2 variants is crucial to efforts in combating the COVID-19 pandemic. Lineage level abundance estimates for SARS-CoV-2 can be obtained from viral material present in domestic wastewater. The abundance predictions can be made at different levels of granularity- individual lineage level(high granularity) or variant level(low granularity). The question this paper answers is to what extent abundance predictions are more accurate at lower granularity. Here we show that when wastewater samples contain only one lineage low granularity predictions are in general more accurate than high granularity for all lineages across Alpha, Delta and Mu variants. No variant level overestimation was observed for this experiment, which was thought to be something that could have made low granularity predictions less accurate than those at high granularity. When lineages of a variant were combined into a wastewater sample, the prediction error rose because of the smaller relative abundances of the genome sequences. Overestimation due to predictions of all lineages being pooled into one lineage was observed here with the overestimated high granularity lineage being more accurate than the low granularity predictions. If samples are expected to contain a very small amount of lineages then it is better to make predictions at low granularity. On the other hand, as the relative abundances of lineages decrease in a sample due to a large number of lineages, the chances of lineage level predictions having a smaller relative prediction error rate increases- making high granularity the better choice for more accurate predictions.

1 Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first discovered in Wuhan, China in December 2019[1]. It is an RNA virus and it is responsible for causing the disease known as COVID-19. Over the course of the pandemic, the virus has undergone numerous changes in its genetic code which are termed mutation. Mutations produce different versions of the original virus which are considered "variants" of the parent virus.

Groups of viruses that have mutated from a common ancestor and which are closely related to each other are called lineages[2]. In [3] a method of naming the different SARS-CoV-2 lineages is outlined. At the root of the SARS-CoV-2 phylogeny are 2 lineages denoted as A and B. The names of the lineages begin with a letter and the lineages descending from these root lineages are assigned numerical values (for example A.1). According to the rules of [3] these lineages can further act as ancestors for other lineages resulting in lineages like A.1.1.

Genome sequences of the virus are represented by a string of letters with each letter representing a nucleotide, which is essential for DNA replication among other functions[4].

Not all mutations change the characteristics of the virus but some give rise to variants that are designated as 'Variants of Concern' that may require public health action. Some variants are more transmissible, cause more severe disease, or cause vaccines and other public and social health measures to be less effective[5]. The Delta variant, for example, originating in India in December 2020, is characterized by an increase in transmission compared to the original SARS-CoV-2 genome and has been reported in 187 out of 194 World Health Organization member countries as of 2021-11-23[6].

Given the fact that differences in the characteristics of variants affect the epidemiology of the coronavirus, it is of prime importance to ascertain which variants of the virus are prevalent in a specific region.

Following reports of detection of SARS-CoV-2 in feces in January 2021, a [7] that SARS-CoV-2 detection through wastewater surveillance could be used for viral tracking. This method is especially useful in areas where clinical estimates are unreliable or impractical to obtain.

Proceeding this study an experimental study by Baaijens et al. in September 2021 showed that quantitative analysis of the RNA material present in wastewater provides reliable abundance estimates for the different Variants of Concern(VoCs) of the SARS-CoV-2 virus[8]. Baaijens et al. introduced a novel method to estimate SARS-CoV-2 variant abundances from wastewater using an algorithm for RNA-Seq quantification. RNA-Seq(RNA-Sequencing) is a next-generation technique that can be used to identify and examine the sequences of RNA from wastewater samples.

In [8] since the sequencing reads(see methodology) obtained from wastewater are predicted at a particular level of the hierarchy of virus classification, for example-lineage or Variant of Concern/Interest, it is important to think about the level in the hierarchy at which predictions should be made since the prediction accuracy changes with the level under consideration. This level is one of the many aspects of the Baaijens pipeline that affects prediction accuracy with some others being reference set design, sequencing errors and genome region being sequenced. This level is indicated by the output granularity with the output referring to the output of the Baaijens pipeline- for example lineages, and granularity referring to how granular(specific) the output level is in terms of the hierarchy of classification. Understanding the effect of output granularity on prediction accuracy for the Baaijens pipeline is the research gap that this paper aims to fill. High granularity predictions correspond to abundance predictions

being made at the individual virus lineage/sub-lineage level. On the other hand, low granularity predictions correspond to predictions being made for a combination of lineages or at the even broader variant level. The predictions at different granularities can provide additional information about the presence of specific lineages of the virus from the wastewater samples. With improvements to this pipeline, we can make better predictions about new variants and have a more accurate idea of which variants are prevalent in a region. The insights from this algorithm open up the possibility to undertake variant-specific measures to curb the spread of the virus in a particular region.

The main research question is 'Do predictions become more accurate at lower granularity? IF so, to what extent?' Results from this research are expected to provide an insight into which granularity is ideal for a variety of variants and how to interpret predictions at different granularities.

- Sub-question 1 - What are the differences in prediction accuracy at different granularities?
- Sub-question 2 - What do the results at different granularities theoretically indicate about the prediction pipeline's strengths and weaknesses?

2 Methodology

The effect of output granularity on prediction accuracy is analyzed through experiments with different configurations like different reference sets and wastewater sample compositions. The results at different granularities are compared are used to obtain an answer the research question.

The input to the pipeline consists of a data set of wastewater sequencing files and a reference set. The pipeline uses quantification software to identify and calculate the abundances for each lineage. The output is a list of abundances, in the percentage of the whole sample for each lineage the user is interested in observing.

The first data set contains the actual wastewater sequencing reads that need to be quantified for variant abundance. Sequencing reads are the result of RNA-Seq quantification on wastewater samples containing viral RNA and they include what genome sequence strings were identified from the wastewater. This data set can be of 2 different kinds. One option is to use a real-life data set of sequencing reads extracted from an actual wastewater sample taken from a sewage plant. The advantage of using this type of data is that it represents a real-life instance of wastewater so the experiment results are grounded in reality and are more of a reality check for the pipeline. The disadvantage however is that the user has no control over what the composition of the wastewater sample is and therefore cannot perform a wider range of experiments- concerning lineages that happen to not be present in the sample.

The other option is to artificially construct a data set where the user has complete control over which genome sequences are present in the wastewater(file) and in what specific abundances and then test how well the pipeline performs

on the data. Simulated data is used for experiments over real-life data because to analyze the effect of output granularity on prediction accuracy as a function of true frequency, not only is simulating the data much more convenient because the abundances simulated are known precisely but the abundances can also be adjusted as needed with ease, which would have been difficult if real samples were used. Simulation tools like ART(see below) offer excellent functionality to fine-tune aspects of the data making simulated data the obvious choice for benchmarking data.

The reference set acts as a dictionary of 'identifiable' genome sequences. Another purpose the reference set fulfills is that some of the reference set sequences are used as background sequences in a wastewater sequencing file i.e. sequences other than the ones we are interested in simulating and testing for abundance. For example in a wastewater sequencing file with 50% simulated abundance of a genome sequence from the B.1.1.7 Delta lineage, the other 50% of the file is made up of background sequences from the reference set.

To perform tasks like RNA-seq quantification and creating simulated data sets for experiments external tools are used. Kallisto is an RNA-seq quantification program [9]. It is responsible for analyzing the wastewater sequencing reads and classifying the RNA sequences found within into specific lineages and calculating final counts for each lineage the user is interested in. Kallisto extracts k-mers, which are substrings of length 'k' of the sequencing read genome string. These k-mers are compared them with the k-mers extracted from the reference genomes using pseudoalignment. Pseudoalignment is a concept where instead of matching sequencing reads to genome sequences that most closely resemble the reads in terms of genome sequence characters, the focus is on identifying which genome sequences in the reference set the reads could have potentially originated from[9] to identify which lineage is most likely present in the sample/data set. Kallisto was used over other similar tools such as TopHat2[10] or Cufflinks[11] because of its lower runtime and comparable accuracy[9].

ART is used as a simulation tool for experiments using simulated wastewater sequencing reads, [12]. ART needs as input an input genome sequence to be simulated and then it mimics the actual sequencing read process including sequencing errors that occur in real-life to produce next-generation sequencing reads.

3 Experimental Setup and Results

The next subsections describe the experimental configurations used and the results obtained.

3.1 Experimental Setup

Multiple reference sets are used for the different experiments performed. When choosing a reference set it is important to consider the location and time frame

to ensure that the specific variants or lineages being analyzed in experiments are present in the reference set so Kallisto can recognize them.

A reference set comprising of genome sequences of SARS-CoV-2 lineages downloaded from GISAID[13] found in Connecticut between October 2020 and September 2021 is used. The reference set was chosen from Connecticut because there is a wastewater treatment plant in New Haven, Connecticut which can be used to obtain real samples which can be used in a later experiment to compare with simulated samples from the same location and provide insights into prediction accuracy at different granularities for simulated versus real-life data.

In addition to a Connecticut reference set, a much larger reference set of sequences from all over the United States from GISAID, collected between October 2020- September 2021 is used to include contrast in sizes of reference sets between a local and global reference set. The results with the two reference sets are expected to differ as the bigger reference set offers a larger and more diverse set of reference sequences for Kallisto, but now the reference sequences are not necessarily found from the same geographic location to where the simulated sequences were found, which offers an interesting interplay to observe the effect of output granularity under. The expectation is that for the US reference set Kallisto would generally perform better for both granularities as with a larger set of reference sequences the Transcriptome-De Bruijn Graph(T-DBG) will have many more reference sequences for the same lineage so it has a higher chance of making correct predictions for lineages. The T-DBG is a graph with k-mers as nodes and vertices which connect different k-mers to form different genome sequence strings which are then used for pseudoalignment.

All experiments are performed with simulated data and 3 variants area analyzed- Alpha(B.1.1.7), Delta(B.1.617.2) and Mu(B.1.621). These variants were chosen because genome sequences belonging to these variants were present in abundance in the Connecticut reference set constructed from GISAID, whereas variants such as Beta were not present in the reference set. The experiments aim to find and analyze the differences in prediction accuracy at different granularities- specifically lineage(high granularity level) versus variant (low granularity level), for all variants at different simulated abundances. It is important to know that since Kallisto can only make predictions at the lineage level, the predictions at high granularity for a variant were obtained by summing up the predictions for the simulated lineages.

A wastewater sample containing 50% abundance of a single genome sequence from a lineage out of all sequences present contains 50% the genome sequence and the remaining 50% is comprised of background sequences.

For every variant, only selective lineages are analyzed. This is because some variants, like Delta, have hundreds of associated sub-lineages and it is impractical to construct a benchmarking dataset containing multiple sequences from each of the lineages, especially since many of the lineages are emerging recently and do not have ample sequences on GISAID at the time the experiments were performed. Additionally, it is also the case that the results are expected to not differ significantly across every lineage, which would make many of the lineages

simulated redundant in terms of results. The genome sequences for every lineage are simulated at 33 different abundances ranging from 0.05% to 100% of the sequencing file.

Delta variant corresponding to lineage B.1.617.2 which was discovered in India in late 2020 has become a dominant strain globally. It is considered one of the most transmissible respiratory viruses known and it has over 200 hundred sub-lineages for the B.1.617.2 lineage as of 2021-12-19. The Delta lineages and sub-lineages simulated are B.1.617.2, AY.3, AY.14, AY.25 and AY.26 and they correspond to a realistic wastewater sample that could have been taken from Connecticut between 1 July-8 July 2021 as the sequences simulated were collected between those dates.

The Alpha variant was first discovered in November 2020 and is characterized by an increase in transmissibility compared to the wild-type SARS-CoV-2. The following lineages and sub-lineages of the Alpha variant were simulated - B.1.1.7(root), Q.1, Q.4 and Q.3. The wastewater sample corresponds to a real-life sample that could have been taken from Connecticut between April 1 and April 30, 2021.

The Mu variant was detected in Columbia, in January 2021 and was considered as a Variant of Interest. The Mu variant corresponds to the lineage B.1.621 and has a sub-lineage B.1.621.1 as of December 2021. The genome sequences simulated were from Connecticut collected between 2021-01-01 and 2021-05-31.

For constructing a benchmarking dataset 2 genome sequences are simulated for every lineage under consideration. Since lineages contain multiple genome sequences, including only one sequence would not necessarily sufficiently represent the variety existing within a lineage and could lead to worse performance. In general, the more sequences simulated per lineage the better the performance is expected to be, the downside being an increase in runtime. So two is chosen as the bare minimum. Additionally, having the same number of genome sequences simulated for every lineage is important as it ensures that the results are normalized so no lineage has an advantage over another with lesser genomes simulated which may result in more inaccurate predictions.

As a key performance indicator accuracy is used and not precision because precision concerns how close different measurements for the same true value are to each other, which is not relevant here because there is only one measurement(abundance prediction) for every true value(true abundance). Accuracy, however, indicates how close the predicted abundance is to the true abundance which is more meaningful as it only concerns the true and predicted abundance. The measure of accuracy being utilized here is relative prediction error(RPE). This measure calculates the error of the model without concern for what the true abundance the error was made at. This is an appropriate measure for this pipeline because there is no reason to punish the RPE depending on what true abundance the error was made at. This is especially important for errors at high true abundance to be normalized by abundance. This would not have been the case if measures such as root mean square error or Mean Absolute error was used instead of Relative Prediction error and error rates from those metrics would have been much higher. The formula for relative error is:

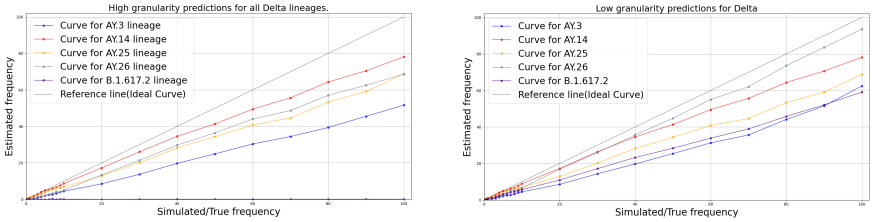
$$RPE = |(T - E)|/T * 100$$

where T is the true abundance of a lineage/variant and E is the estimated abundance.

3.2 Results

The following experiments are carried out with the Connecticut Reference set.

Figures 1, 2 and 3 show the plots for the Delta, Alpha and Mu variants respectively for samples with one genome sequence simulated at a time(Single lineage experiment). Sub-figures (a) and (b) correspond to prediction accuracy plots at high and low granularity respectively, as a function of true abundance. Figure 4 shows the RPE plot for the Delta and Alpha variants.



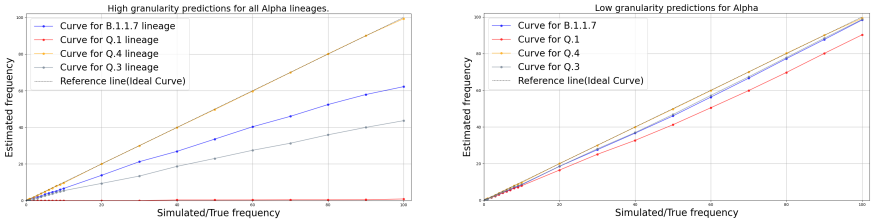
(a) High Granularity (b) Low Granularity

Figure 1: Delta Variant, Single Lineage Sample(Connecticut)

For the Delta variant although both AY.14 and B.1.617.2 have roughly the same number of genome sequences representing them in the reference set, at all abundance levels Kallisto classifies a B.1.617.2 sequencing read as an AY.14 sequence, therefore leading to the high granularity curve for the root lineage in Figure 1(a) to stagnate at 0% for all abundances. Results for B.1.617.2, AY.3 and AY.26 are better when considering lower granularity compared to their high granularity results. The biggest improvement with low granularity was observed for the root lineage B.1.617.2 with the relative prediction error(RPE) decreasing to 40% at most abundances above 0.1% (Figure 4(a)).

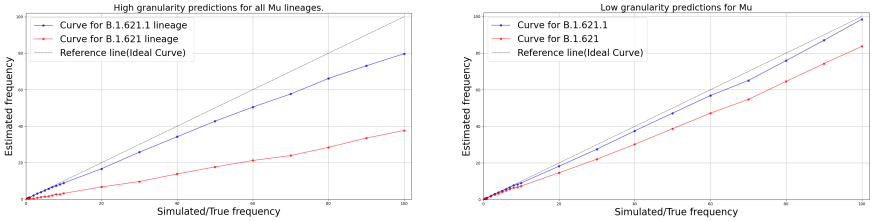
Figure 2(a), 2(b) indicate that Alpha is one of the variants that benefit greatly in terms of prediction accuracy when considering low granularity predictions. Lineages B.1.1.7 and Q.3 see significantly improved results at low granularity with the biggest improvement being for Q.1 from 100% RPE to only 20-30% at most abundances greater than 0.1%(Figure 4(b)).

The B.1.621.1 sub-lineage was consistently more accurate than the root lineage, and it was noted that sequences from B.1.621 were being misclassified as B.1.621.1 leading to more inaccurate predictions for B.1.621. At low granularity RPE for B.1.621 improved from 60%(at high granularity) to about 20% at most abundances above 0.1% true abundance(see Appendix).



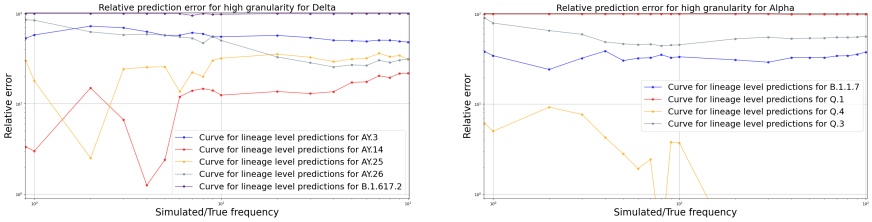
(a) High Granularity (b) Low granularity

Figure 2: Alpha Variant, Single Lineage Sample(Connecticut)



(a) High Granularity (b) Low granularity

Figure 3: Mu Variant, Single Lineage Sample(Connecticut)



(a) RPE for Delta. (b) RPE for Alpha

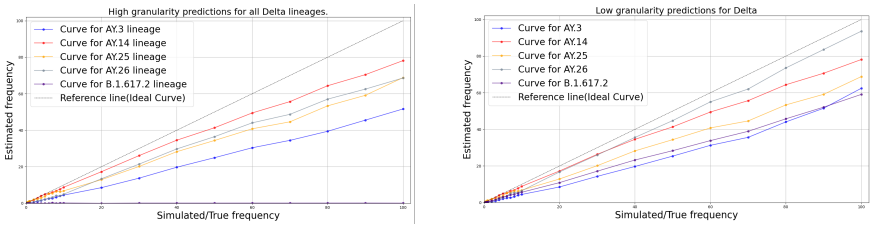
Figure 4: High Granularity RPE Plots, Single Lineage Sample(Connecticut)

The same experiment was then carried out with a US reference set to analyze to what extent the results specifically Q.1(Alpha) and B.1.617.2(Delta) were influenced by the size and diversity of the reference set.

Figures 5, 6 and 7 are granularity plots for the single lineage experiments using a US reference set and sub-figures (a) and (b) are high and low granularity plots respectively. Figure 8 is the RPE plot for Delta(8(a)) and Alpha(8(b)) variants respectively.

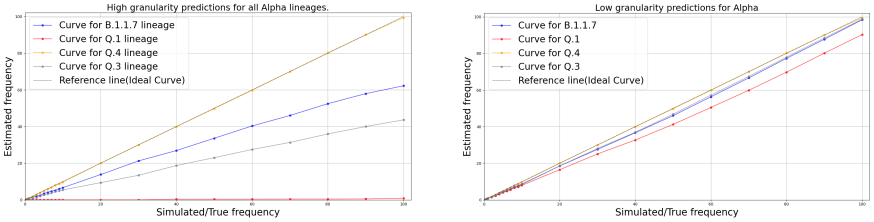
For the Delta variant, predictions for the B.1.617.2 lineage were significantly better at high granularity compared to the Connecticut experiment. The RPE consistently decreased from a 100% at 0.6% true abundance to 40% at 100% abundance(Figure 8(a)). At low granularity, results were not appreciably better

compared to higher granularity which implies that genome sequences from one lineage were not misclassified into another lineage for the Delta variant. For the Alpha variant except for lineage Q.3, the predictions are worse for all variants for the US reference set(Figure 8(b)), which is not as expected with a larger reference set. For the mu variant, results at both granularities with the US set were better for both lineages as expected but not by a very huge margin and the relative prediction errors are in a comparable range as their Connecticut counterparts(see Appendix).



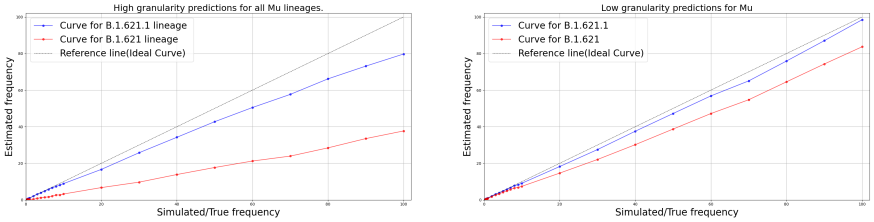
(a) High Granularity (b) Low granularity

Figure 5: Delta Variant, Single Lineage Sample(United States)



(a) High Granularity (b) Low granularity

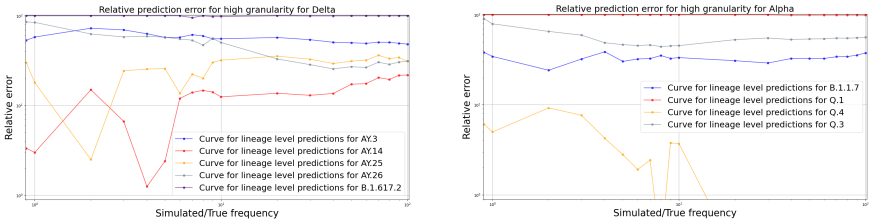
Figure 6: Alpha Variant, Single Lineage Sample(United States)



(a) High Granularity (b) Low Granularity Plot

Figure 7: Mu Variant, Single Lineage Sample(United States)

For the US reference set, a second set of experiments were performed with all lineages/sub-lineages of a particular variant present together in a single wastew-



(a) RPE for Delta.

(b) RPE for Alpha

Figure 8: High Granularity RPE Plots, Single Lineage Sample(United States)

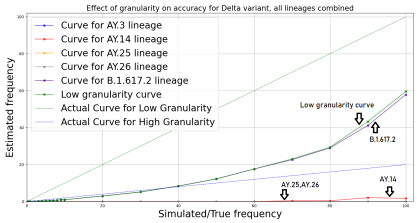
ater sample. This configuration was very different from how lineages had been simulated for the previous experiments as previously there was only one genome sequence of one lineage in one wastewater sample. This combined lineage experiment compelled Kallisto to not only detect which variant was present singularly, like in the previous experiments but now it had to also differentiate between different lineages/sub-lineages of a variant. This is where simulating multiple genome sequences for every lineage offers utility as this better arms Kallisto to derive more complex and detailed T-DBGs to better distinguish between different lineages in the same sample. This simulated data-set is more realistic compared to the previous experiments and the results offer an interesting insight into as to how they vary from single lineage samples.

An example composition of a wastewater sequencing sample with 50% abundance comprises 50% of the file consisting of background sequences, just like a sample with a single genome sequence simulated, but the remaining 50% is now divided over all the genome sequences simulated for a variant. For the Mu variant which has lineage B.1.621 and sub-lineage B.1.621.1, with 2 genome sequences per lineage simulated, each lineage was present in 25% abundance with each genome sequence within a lineage comprising of 12.5% of the sample, which for a total of 4 genome sequences makes up the 50% of the simulated sequences.

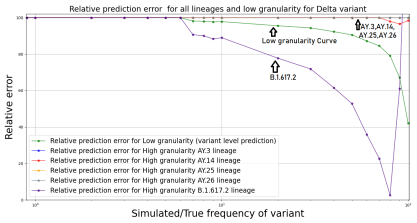
Figures 9, 10 and 11 detail the results obtained for the combined lineage experiments. Figures 9(a), 10(a) and 11(a) correspond to the estimated frequency on the y-axis and true frequency on the x-axis and show estimated frequency curves for high and low granularities for Delta, Alpha and Mu variants respectively. Figures 9(b), 10(b) and 11(b) detail the comparative performance of high and low granularities to evaluate which performed better in terms of RPE.

Since in this experiment the estimated abundance of a specific lineage at a simulated abundance should be equal to simulated abundance divided by the number of lineages simulated for that variant there are separate reference lines for high and low granularity values. The reference line for low granularity values(dotted blue line) remains the same as previous experiments but this experiment includes an additional reference line for high granularity values(dotted green line) which indicates what the actual frequency should be for the lineages.

Results for the combined lineage experiment were quite surprising. For the Delta variant, sequences from all sub-lineages were misclassified into the root

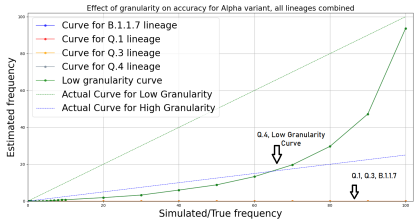


(a) Low versus High Granularity

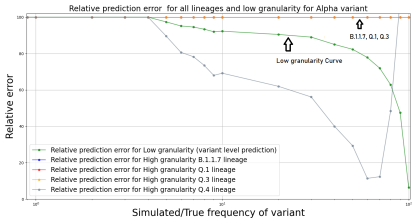


(b) RPE for Low and High granularity

Figure 9: Delta Variant, Combined Lineage Sample

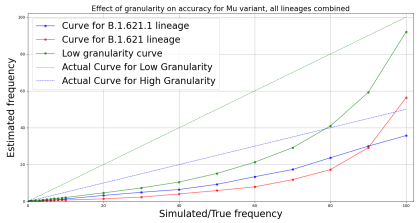


(a) Low versus High Granularity

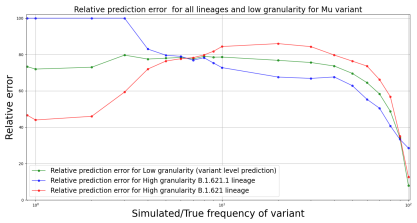


(b) RPE for Low and High granularity

Figure 10: Alpha Variant, Combined Lineage Sample



(a) Low versus High Granularity



(b) RPE for Low and High granularity

Figure 11: Mu Variant, Combined Lineage Sample

lineage B.1.617.2 with the predictions for B.1.617.2 being 57.88% at a 100% true abundance where B.1.617.2 was expected to be at a maximum abundance of 20%. It also means the remaining 38.88% of predictions were from the other sub-lineages all of which were simulated at 20% abundance each. For the alpha variant- lineages B.1.1.7, Q.1 and Q.3 had an RPE of 100% for all abundances. It was observed that for Delta and Alpha the best RPE curve for actually for B.1.617.2 and Q.4(respectively) at most abundances(Figure 9(b), 10(b)) and not the low granularity curve as seen before. For the Mu variant, the results at high granularity were much better compared to Delta and Alpha. The relative prediction error for B.1.621.1 was in the 30-80% range (averaging at 60%), for frequencies between 4 and 100%. B.1.621. had a similar prediction error

although slightly larger in magnitude. At 100% abundance, B.1.621 overtook B.1.621.1.

4 Discussion

To evaluate the effect of output granularity on prediction accuracy, we have conducted experiments using the Baaijens pipeline for Alpha, Delta and Mu variants. The true abundances for genomes simulated range from 0.05% to 100% of a wastewater sample. The experiment with the Connecticut and US reference set, as well as the combined lineage experiment have provided an insight into how accurate predictions are at different granularities. Additionally, the results have revealed the strengths and weaknesses of the prediction pipeline regarding its ability to calculate abundances accurately.

For the Connecticut Reference set, the delta variant sequencing reads from genome sequences from root lineage B.1.617.2 were misclassified as AY.14. This misclassification is mitigated when a US reference set is used which has 4 times as many reference genome sequences for the two lineages. Therefore part of the problem causing this misclassification seems to be a lack of diversity in reference sequences which can cause the sequences simulated to be more easily misclassified. For the delta variant using the US reference set yields more accurate results for all lineages observed.

For the Connecticut reference set, results for Q.1(Alpha) are interesting as even though in the reference set Q.1 has more reference sequences than Q.4, in some cases all lineages other than Q.1 seem to be favored for classification for a sequence that was from Q.1. Just like Delta's B.1.617.2 for the Connecticut reference set, the high granularity predictions for Q.1 are consistently almost 0 because of these misclassifications. Lineage Q.4 is also interesting as it seems to be predicting the abundances exactly with an unnoticeable margin of error indicating the pipeline is very accurate for this lineage.

For the Connecticut reference set, for the Mu variant, it is interesting to note that Kallisto seems to favor predictions for sub-lineage B.1.621.1 with high granularity predictions for B.1.621, the root lineage being about twice as errorsome on average than the sub-lineage B.1.621.1. The results improve significantly when using a US reference set for Mu for both lineages although B.1.621.1 was still a lot more accurate. The results using a combined lineage data-set get better exponentially the higher simulated frequency with the results having a very small error margin at 100% true abundance

A hypothesis for why B.1.617.2 and Q.1 predictions at high granularity are so low for Connecticut is as follows. For lineages that like B.1.617.2 and AY.14 which are very closely related, the T-DBG will have fewer branches between the path covers of the of the two lineages making it more likely that a sequencing read from one lineage would be assigned to another, causing the B.1.617.2 flat curve observed in Figure 1(a) and Q.1 flat curve in Figure 2(a). With the US reference set, which has more genome sequences for both AY.14 and B.1.617.2 the T-DBG constructed has more diversion for the B.1.617.2 and AY.14 branches

so they are not misclassified as much anymore as seen in the results. This also explains the observation that the root lineage performance is comparatively worse than the sub-lineages in variants like Mu. All variants considered, the pipeline does reasonably well at differentiating between lineages/sub-lineages regardless of the sub-lineage bias when the lineages are simulated separately in wastewater samples and are not combined into one sample.

It is also interesting to note that not all lineages necessarily seem to benefit from low granularity predictions for example for Delta lineages AY.14 and AY.25, the low and high granularity curves overlap almost perfectly. This indicates that Kallisto is very accurate at predicting these lineages at high granularity.

Using the US reference set revealed some interesting results for the Alpha variant. Firstly, the predictions at both granularities for lineage B.1.1.7 were worse compared to the Connecticut reference set, which may be due to the genome sequences simulated from Connecticut not bearing as much resemblance to the reference sequences from the US reference set. More in-depth analysis, which is out of the scope of this project is required to confirm if this hypothesis is true.

For the combined lineage experiments, results for B.1.1.7 and Q.1 were not unexpected as in the experiment where B.1.1.7 and Q.1 are simulated individually with the US reference set, their estimated abundances at 20% are essentially 0%. Results for Q.3 were much worse compared to the US reference set results indicating that the accuracy does suffer in comparison when lineages are combined into the same sample. The results at lower granularity had smaller RPE rates implying that for combined lineage experiments, lower granularity is a much better choice compared to high granularity especially since accuracy for most Alpha and Delta lineages was quite poor because of their abundances being a maximum of 25% or 20% respectively so the more lineages that were simulated the worse the results as it meant a smaller abundance per lineage. It is expected that the results for Mu were better compared to Alpha and Delta for this very reason. For this experiment predictions of some lineages seem to be pooled into a single lineage of the variant, resulting in overestimation for that lineage. Also, because of the lower relative abundances of the lineages, the difference between the true high and low granularity curves could not be made up for when considering low granularity even at higher abundances. This resulted in the RPE of the overestimated lineage to being less than that of the lower granularity curve from 5% to 90% abundances(x-axis). For this experiment it was not expected that the aforementioned difference would not be made up when considering lower granularity.

5 Responsible Research

The code used for running the experiments can be found here[1]. The README file in the repository provides an explanation on how to set up the code base carry out the experiments in a step-by step-manner and instructions on how to download the genome sequences simulated from GISAID along with all filters used. Given GISAID's privacy policy the data downloaded is not available in the GitLab repository as GISAID registration is required for access to the data.

The analysis of the results was done in a disinterested manner and negative results have not been ignored but pointed out and analyzed. The final results are available to the public. Additionally, the principle of Independence from the Netherlands Code of Conduct for Research Integrity holds as specific results do not benefit the researchers materialistically.

It is important to note that GISAID is a secondary data source as the genome sequences used are not acquired directly from the patients by the researchers for this paper. Additionally, a limitation is that the credibility of the GISAID data being used cannot be verified as it is outside the scope of this research which has implications on how correct the conclusions drawn from this data can be.

6 Conclusion and Future Work

The purpose of this research was to evaluate and understand the effect of output granularity on prediction accuracy of the Baaijens pipeline with a focus on the research question "Do predictions become more accurate at lower granularity?".

Prediction accuracy was analyzed for a variety of variants with different reference sets and wastewater sample configurations. Results for single lineage experiments confirmed that predictions at low granularity- specifically at the variant level were consistently at least as accurate as high granularity results- at the lineage level. An increase in prediction accuracy was observed for most variants when granularity was lowered. Additionally, no overestimation was observed for the single lineage experiments which is a potential downside to considering results at lower granularity as the high granularity results are added up to obtain low granularity results.

For the combined lineage experiments, the smaller relative abundances of the lineages resulted in overestimation of a specific lineage of the variants in the case of Alpha and Delta and this resulted in the RPE of the overestimated lineage to be less than that of the lower granularity curve at some abundances leading to the conclusion that in a wastewater sample with only a small abundance per lineage it is ,therefore, better to consider results at higher granularity for better accuracy.

All in all, as far the strengths and weaknesses of the pipeline are concerned, the accuracy depends heavily on the reference set design, and granularity under

[1] `https://gitlab.ewi.tudelft.nl/jbaaijens/CSE3000_wastewater_project/-/tree/yash`

consideration. Lineages like Q.4 from Alpha have been observed to have near-perfect accuracy at high granularity even for a small reference set whereas Q.1 and B.1.617.2 have an estimated abundance of 0% and results vary significantly from variant to variant and in some cases even for the different lineages/sub-lineages of the same variant. When a lot of lineages are combined into the same sample the performance suffers considerably and results indicate the fewer lineages that are combined the more predictable the results will be.

Future work includes running the experiment for a larger variety of reference sets and variants to get a larger variety of plots to analyze. Another avenue is using real-life wastewater samples instead of simulated data sets which the results for which would be grounded in reality.

References

[1] Kristian G Andersen et al. "The proximal origin of SARS-CoV-2". In: *Nature medicine* 26.4 (2020), pp. 450–452.

[2] *SARS-CoV-2 Variant Classifications and Definitions*. `https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html`. Accessed: 2021-12-18.

[3] Andrew Rambaut et al. "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology". In: *Nature microbiology* 5.11 (2020), pp. 1403–1407.

[4] Larry R Engelking. *Textbook of veterinary physiological chemistry, updated 2/e*. Academic Press, 2010.

[5] *Tracking SARS-CoV-2 Variants*. `https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/`. Accessed: 2021-11-25.

[6] *Global COVID-19 deaths hit 5 million as Delta variant sweeps the world*. `https://www.reuters.com/world/global-covid-19-deaths-hit-5-million-delta-variant-sweeps-world-2021-10-02/`. Accessed: 2021-11-23.

[7] Salmaan Sharif et al. "Detection of SARs-CoV-2 in wastewater using the existing environmental surveillance network: A potential supplementary system for monitoring COVID-19 transmission". In: *PloS one* 16.6 (2021), e0249568.

[8] Jasmijn A Baaijens et al. "Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-Seq quantification". In: *medRxiv* (2021).

[9] Nicolas L Bray et al. "Near-optimal probabilistic RNA-seq quantification". In: *Nature biotechnology* 34.5 (2016), pp. 525–527.

[10] Daehwan Kim et al. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome biology* 14.4 (2013), pp. 1–13.

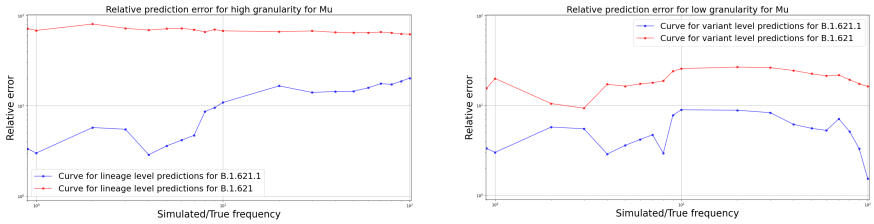
[11] Sreya Ghosh and Chon-Kit Kenneth Chan. "Analysis of RNA-Seq data using TopHat and Cufflinks". In: *Plant Bioinformatics*. Springer, 2016, pp. 339–361.

[12] Weichun Huang et al. "ART: a next-generation sequencing read simulator". In: *Bioinformatics* 28.4 (2012), pp. 593–594.

[13] Peter Bogner et al. "A global initiative on sharing avian flu data". In: *Nature* 442.7106 (2006), pp. 981–981.

7 Appendix

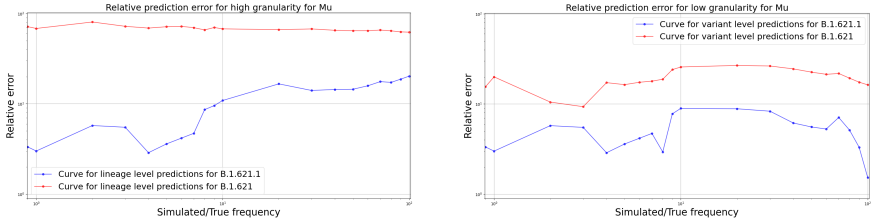
Relative Prediction Error Plots for Connecticut Reference Set:



(a) High Granularity RPE plot (b) Low Granularity RPE Plot

Figure 12: Mu Variant, Single Lineage Sample(Connecticut)

Relative Prediction Error Plots for US Reference Set:



(a) High Granularity RPE plot (b) Low Granularity RPE Plot

Figure 13: Mu Variant, Single Lineage Sample(US)