SARS-CoV-2 lineage abundance quantification in wastewater: a benchmark study for the identification of optimal reference set design

by

Ioanna Nika

Supervisor: Jasmijn Baaijens

A Dissertation

Submitted to EEMCS faculty

Delft University of Technology,

In Partial Fulfilment of the Requirements

For the Bachelor of Computer Science and Engineering

January 23, 2022

SARS-CoV-2 lineage abundance quantification in wastewater: a benchmark study for the identification of optimal reference set design

Ioanna Nika[1*], **Supervisor: Jasmijn Baaijens**[1]

[1]Delft University of Technology: Faculty of Electrical Engineering, Mathematics & Computer Science
i.nika@student.tudelft.nl, j.a.baaijens@tudelft.nl

Abstract

Lineage abundance estimation of SARS-CoV-2 in wastewater is a technique that aims to monitor the lineage prevalence in communities and help contain the COVID-19 pandemic. Lineages are collections of closely related mutants of a virus. It is suggested that the genome sequences of lineages differ across the globe due to random mutations or distinct immune responses of populations that mutate the virus. In order to estimate the lineage abundance in a specific community, wastewater data collected from the community are compared to reference SARS-CoV-2 genome sequences of different lineages. However, such region-related variation in the genome sequences of lineages could impact the abundance estimates. The main aim of this study is to identify an optimal way of sourcing reference genome sequences such that the lineage abundance estimates are improved. For the purpose of evaluating the performance of different reference sets, simulated wastewater data are used. We demonstrate that continent-specific reference sets are the most reliable option. The overall country interactions with other parts of the world could be considered for constructing an optimal reference set. Additionally, results show that considering immune-response related mutations for the reference set construction does not influence performance. Finally, it is suggested that a higher number of sequences per lineage and the inclusion of recently sourced sequences in the reference set improve results.

1 Introduction

COVID-19 is a highly contagious disease caused by SARS-CoV-2 and is responsible for more than 5 million deaths globally [1]. During the COVID-19 pandemic, many virus mutants and lineages have been identified. A virus mutant is a version of the original virus that contains an instance of mutation. Mutants do not necessarily have different properties than the original virus. A lineage is a collection of mutants that share predecessors. Continuous monitoring of the existing lineages is necessary and crucial for the efforts taken globally to contain the virus and avoid outbreaks. However, as clinical sequencing is not feasible in all situations, new methods can be considered additionally to existing methods in order to observe how the pandemic evolves and act early.

Lineage abundance estimation of SARS-CoV-2 in wastewater using RNA-Seq quantification methods is a promising technique that can facilitate the monitoring of different lineages especially when clinical sequencing is beyond reach [3]. Baaijens et al.[3] prove that such methods are effective in identifying trends in lineage prevalence. The workflow followed by the aforementioned study [3] can be described as follows. Multiple genome sequences from various lineages are provided to a tool that quantifies RNA-seq data named Kallisto [5]. Kallisto [5] is then responsible for aligning the reads from the genome sequences contained in wastewater data to the reference genome sequences. Finally, the alignment results are analyzed and the lineage abundance is estimated.

Many design choices need to be made regarding the selection of the genome sequences used for the reference set. Some of those choices should potentially be adjusted based on the specific wastewater data that need to be measured, in order to achieve higher performance. Some early findings of Baaijens et al. [3] suggest that state/country specific reference sets can improve performance results. In the same study, it is noted that such reference sets are likely to facilitate the alignment process due to random mutations that are dominant only in certain geographical regions [3].

Similarly, studies suggest that SARS-CoV-2 genome sequences sourced from different continents showed different mutation patterns [6; 12]. Specifically, different geographical regions show different percentages of mutations per genomic region. Results from Pachetti et al. [12] show also that the mutation frequencies per genomic region change over time as the pandemic evolves.

Viral mutations arise in several ways, one of which is the host immune response. Wang et al. [16] suggest that the immune response of several populations which are divided by continent, influences the evolution of SARS-CoV-2. Specifically, first in the study it is demonstrated that certain mutations are associated with the host-virus interaction and that the those mutation frequencies differ among geographical lo-

*Contact Author

cations [16]. Finally, Wang et al. [16] recommend that the different populations might have developed distinct immune responses to the viral infection that leads to different mutation frequencies being observed within the studied populations.

The studies considered, suggest that the SARS-CoV-2 genome sequences differ across the globe and that these differences can at least be grouped by continent. Also, it is mentioned that those differences can be explained by random mutations that are happening locally as well as the different immune responses of populations.

Aim of the study The main aim of this study, is to identify an optimal way of sourcing reference genome sequences from the different geographical regions and thus to be able to design a reference set such that the method described in Baaijens et al. [3] achieves the best results.

2 Experimental work

Different mutational patterns are observed in the genome sequences sourced in different geographical regions. Those differences involve the genomic region, the frequency of mutations as well as the type of mutations. It is thus expected that by making this variation common between the reference set and the lineages measured, alignment performance could be improved since the lineages measured and their corresponding sequences in the reference set will be more similar. Moreover, if this variation is not common among the different lineages, the alignment process could also be facilitated as it would make the different lineages to be more dissimilar between them.

In the following experiments, multiple reference sets are created in order to observe the effect of several factors on performance. The prediction accuracy of different reference sets is evaluated on given test sets created using simulated wastewater data. By understanding whether and how each of the factors studied influences results, better guidelines can be given on how the sequences used for the reference set should be sourced for geographical locations that can be found in the mainland of a continent as well as for remote areas.

2.1 Geographical proximity and its effects on prediction accuracy

In this experiment, we study how the geographical proximity between the source of collection of the reference genome sequences and the genome sequences used in the test set, influences prediction results. To do so, the test sets are formed using samples sourced from the states shown in Table 1. Then, the reference sets are constructed with increasing geographical proximity. Thus, starting with a global reference set and concluding with a state-specific reference set. A continent-specific reference set, a country-specific, and a reference set that includes lineages from the state and its nearby states are also formed. Table 1 shows the specific collection times and geographical locations for the reference set and test set construction for all proximity experiments conducted. The lineage measured in each experiment is also shown.

Table 1: Proximity experiment information regarding the geographical locations, the collection times of the reference sets and the test sets. The lineage measured is also shown.

Test set: state	Test set: country	Test set: collection time	Reference set: collection time	Lineage measured
Massachusetts	USA	May, 2021	01/2021 - 03/2021	B.1.1.7
Connecticut	USA	May, 2021	01/2021 - 03/2021	B.1.1.7
Indiana	USA	May, 2021	01/2021 - 03/2021	B.1.1.7
Maharashtra	India	March, 2021	11/2020 - 01/2021	B.1.1.7
Kerala	India	March, 2021	11/2020 - 01/2021	B.1.1.7
Telangana	India	March, 2021	11/2020 - 01/2021	B.1.1.7

2.2 Effects of ancestry and immune response-related mutations in prediction accuracy

The immune response of the host influences how the virus evolves. Nedelec et al. [11] suggest that such immune response differences observed in various populations are related to the ancestry of the population. In this experiment, the aim is to identify the influence of ancestry and thus immune response related mutations in performance. This experiment is particularly interesting for remote communities or for cases where genome sequences are lacking from an optimal geographical region, thus sequences need to be sourced elsewhere.

A study on the demographics of Argentina suggests that the population of Argentina has European ancestors for 67%. A significant number of Argentinians have European ancestry from Spain, Portugal, Italy, and Greece [8]. The population of Argentina has East Asian ancestors for 0.014% [8]. Hence, the test set is constructed using samples sourced in Argentina. The first reference set is constructed using samples sourced from south Spain, Portugal, Italy, and Greece, and the second reference set is constructed using samples sourced from East Asia. A continent-specific reference set is constructed for the purpose of comparison.

Table 2: Information regarding the geographical locations, the collection times of the reference sets and the test sets for the ancestry experiments conducted. The lineage measured is also shown.

Test set: geographical location	Test set: collection time	Reference set: collection time	Lineage measured
Argentina	March, 2021	11/2020 - 01/2021	B.1.1.7
Argentina	April, 2021	12/2020 - 02/2021	B.1.1.7
Argentina	May, 2021	01/2021 - 03/2021	B.1.1.7

Table 2 shows the collection times, for the reference sets and the test sets as well as the lineage measured for each of the ancestry experiments conducted.

2.3 Effects of population interactions in prediction accuracy

Interactions between two populations could help local versions of a lineage to spread in other geographical regions. The purpose of this experiment is to observe how the interactions between geographical areas influences results. This

experiment is particularly interesting for islands or remote geographical areas.

Cyprus is an island in the Mediterranean sea. For the period of August 2020 anyone traveling to Cyprus had to fill in the Cyprus flight pass which includes their country of residence. The majority of travelers declared as country of residence the UK (39%) and Germany (12%) [13]. No travelers from France were reported for that period or during the previous month [13; 14]. The island reported various percentages of travelers from European countries for the period of July and August [13; 14]. All travelers reaching Cyprus for those two months declared that they reside in Europe.

Genome sequences collected from Cyprus during October 2020 are used for the construction of the test set. Next, three reference sets are created for the time period of August 2020. The first reference set contains samples sourced from the UK and Germany. The second reference set contains sequences sourced from France. Finally, the third reference set contains sequences sourced from Europe (not restricted to the countries Cyprus reported having interactions with). The B.1.258 lineage is measured.

2.4 Collection times for the reference set and the test set

As previously discussed it is suggested that mutation frequencies and the genomic location where mutations occur change over time [12]. In this experiment, it is tested how shorter time intervals between the collection of genome sequences for the reference set and the test set influence results. For this experiment, SARS-CoV-2 genome sequences collected in Connecticut are used for the creation of the test set. North American genome sequences are used to build the reference sets. The data used for the test set are collected during May 2021. The reference sets are constructed with a time span of approximately fifteen days for consecutive time periods from April till February 2021. The B.1.1.7 lineage is measured.

2.5 Effects of within lineage variation in the reference set

As described by Baaijens et al. [3], in order to improve predictions multiple genome sequences are used as reference per lineage in order to capture within lineage variation. In this experiment, it is evaluated how the number of sequences provided as a reference for the lineage measured influences predictions.

For the purpose of this experiment, five reference sets are constructed that contain two, four, six, eight, and ten B.1.1.7 sequences. Genome sequences collected from Connecticut sourced in May 2021 are used for the creation of the test set. Subsequently, four reference sets are constructed. All reference sets contain sequences collected from the USA for the time period between January and March of 2021. The lineage B.1.1.7 is measured.

3 Results

In this section, the results for the experiments described in section 2 are presented. First, in subsection 3.1 is shown that while most reference sets examined in the proximity experiments show high variability in their performance, continent-specific reference sets perform reliably well. In subsection 3.2 the results for the ancestry experiments are presented and it is suggested that immune response-related mutations do not influence performance. Moreover, in subsection 3.3 the findings of the country interactions experiment are presented. In subsection 3.4 is suggested that by including recently sourced genome sequences, better performance is achieved. Finally, in subsection 3.5 is shown that with a higher number of reference sequences, performance improves.

3.1 Variability in performance of the reference sets used in the proximity experiments

Figure 1 and Figure 2 show the average relative prediction error (Equation 2) of the reference sets constructed per proximity experiment. Figure 1 illustrates the results obtained for the experiments done for the states in India. Figure 2 shows the results obtained for the experiments done for the states in the USA. As can be observed in both figures, there is not a consistent type of reference set that achieves best results in all experiments and most reference sets show high variability in their performance. However, in most cases, continent-specific reference sets achieve the best results.

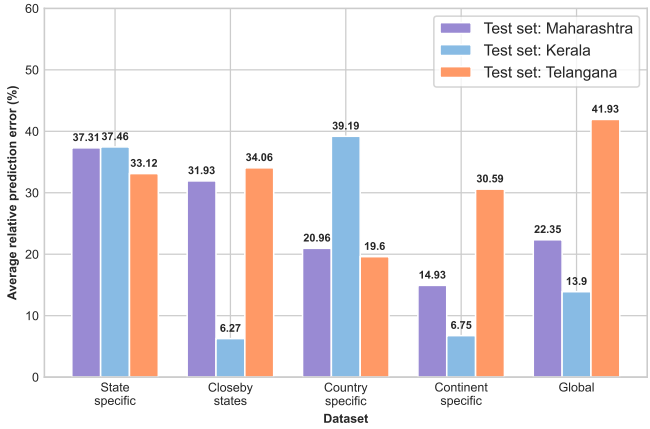


Figure 1: Average relative prediction error for all proximity experiments conducted for states in India

Figure 1 shows that in the experiment done for the state of Maharashtra, the continent-specific reference set achieves the best results. For the experiment conducted for the state of Kerala, the reference set constructed for the nearby states achieves the best results while the continent-specific reference set achieves marginally worse results. Finally, in the experiment conducted for the state of Telangana best results are achieved from the country-specific reference sets, and the continent-specific reference set follows. The state-specific reference sets show high error rates in all experiments conducted for the states in India.

Similarly, Figure 2 shows that the continent-specific reference set achieves the best results in the experiment done for the state of Connecticut. In the experiment conducted

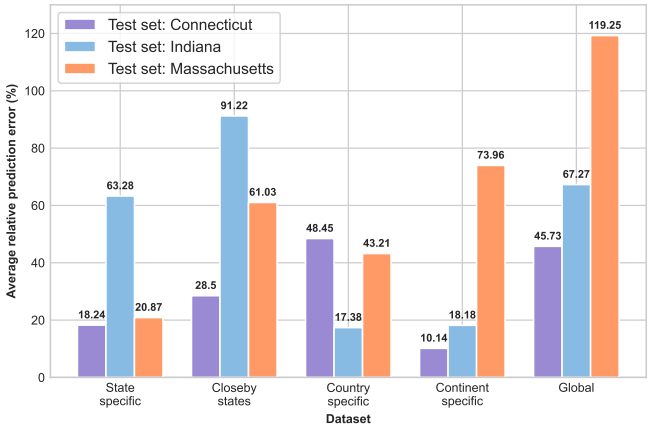


Figure 2: Average relative prediction error for all proximity experiments conducted for states in the USA

for the state of Indiana, the country-specific reference set achieves the best results, and the continent-specific reference set performs marginally worse. Finally, for the state of Massachusetts, the state-specific reference set performs best and the continent-specific reference set has one of the worst performances.

Overestimation and underestimation analysis

Figure 3 shows the simulated frequency compared to the estimated frequency for the proximity experiment done for the state of Maharashtra. For true frequencies ranging from 1-10%, the global, country-specific, and continent-specific reference sets show overestimation as well as some underestimation. For true frequencies ranging from 10-100% mainly underestimation is observed for those reference sets. The other reference sets suffer from underestimation across the whole range of simulated B.1.1.7 frequencies.

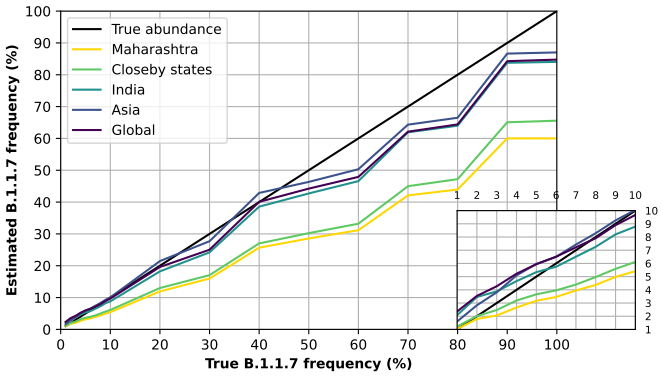


Figure 3: True abundance compared to the estimated abundance for the proximity experiment done for the state of Maharashtra in India.

Figure 4 shows the simulated frequency compared to the estimated frequency for the proximity experiment done for the state of Connecticut. The state-specific and the global reference sets show some overestimation up until the 50% to 60% true frequencies and for the following abundances, some underestimation is observed mainly for the Global reference set. Mostly, underestimation is observed for the rest of the data sets in all abundances.

Similar findings to what have been discussed for Figure 1 and Figure 2 were made for the rest of the proximity experiments conducted.

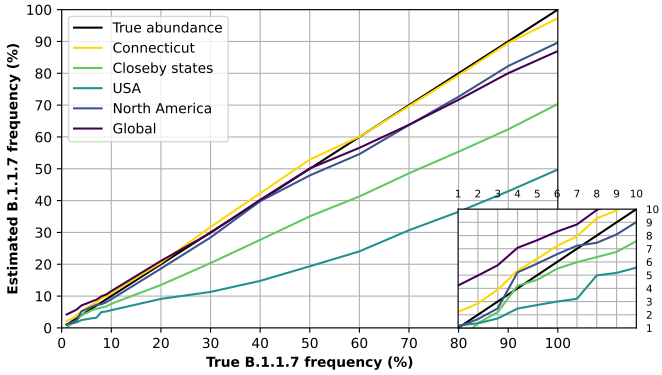
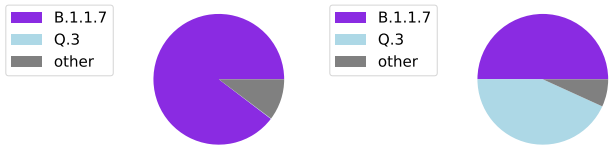


Figure 4: True abundance compared to the estimated abundance for the proximity experiment done for the state of Connecticut in the USA.

Abundance estimation analysis

Figure 5 shows the abundance distribution across the lineages in the reference sets used. In both subfigures, the B.1.1.7 sequence sourced in Connecticut is simulated at 100% abundance. Figure 5(a) corresponds to the North American reference set which achieves the best results in the proximity experiment done for Connecticut while figure 5(b) corresponds to the USA reference set which has the higher average relative error in the same experiment. What is observed is that, even though both reference sets contain the same sub-lineage of B.1.1.7, the USA reference set is affected much more by its presence.

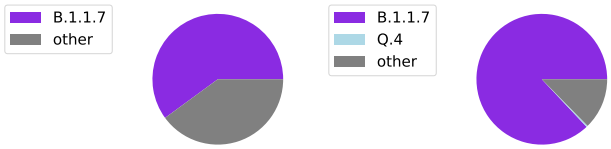


(a) North American reference set (b) USA reference set

Figure 5: Abundance estimation predictions for the simulated abundance of 100% of the B.1.1.7 lineage sourced in Connecticut.

Figure 6 shows the abundance distribution across the lineages for the reference sets used. In both subfigures, the B.1.1.7 lineage sourced in the state of Maharashtra is simulated at 100% abundance. Figure 6(a) corresponds to the state-specific reference set which achieves the worst results in the proximity experiment conducted for the state of Maharashtra, while figure 6(b) corresponds to the continent-specific reference set which has the lowest average relative

error in the same experiment. What is observed is that the state-specific reference set suffers from underestimation due to similar lineages that it contains. Even though the Asian reference set contains sub-lineages it is not affected by their presence and it thus performs better.



(a) Maharashtra reference set (b) Asian reference set

Figure 6: Abundance estimation predictions for the simulated abundance of 100% of the B.1.1.7 lineage sourced in Maharashtra.

Figure 7 shows the estimated abundance compared to the true abundance for the continent-specific reference sets for all proximity experiments conducted. It can be observed that all continent-specific reference sets follow similar trends. In the experiments done for Massachusetts and Telangana, the continent-specific reference sets achieve higher error rates. Those error rates are caused by overestimation in lower abundances.

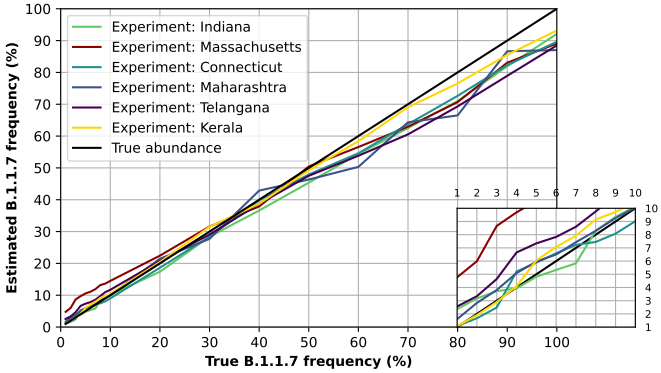


Figure 7: True abundance compared to the estimated abundance for the continent-specific reference sets for all proximity experiments conducted.

3.2 Immune response-related mutations do not influence performance

In Figure 8, the average relative prediction errors (Equation 2) for the ancestry experiments described in subsection 2.2 are presented. The South American reference sets perform well in all experiments conducted. The results obtained by the south European and east Asian reference sets are mostly comparable. Thus, it is suggested that ancestry and immune response-related mutations do not influence performance.

Figure 9 shows the estimated abundance over the true abundance for the reference sets used in the ancestry experiment done for the time period of May. It is shown that all reference sets suffer mainly from underestimation in higher simulated

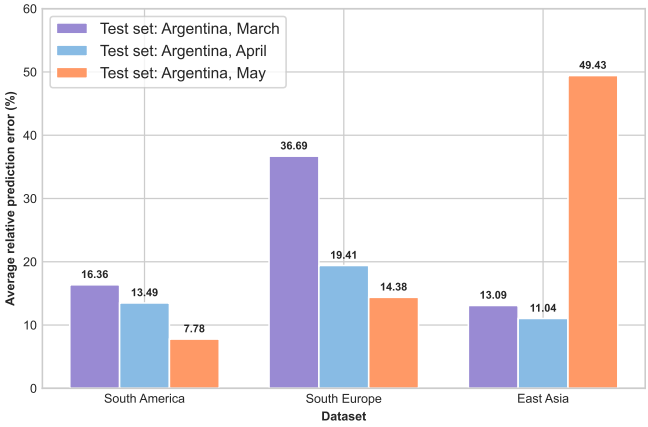


Figure 8: Average relative prediction error percentage for the South American, the South European, and the East Asian reference sets. Each experiment presented covers a different time period. Thus, the test sets are built with samples sourced from Argentina for the time periods of March, April and May.

abundances. Similar observations are made for the remaining experiments conducted for the other time periods.

Figure 10 shows that the underestimation that is observed in the experiment done for the period of May (Figure 9) is mainly caused by sub-lineages that are present in the reference sets. Likewise, Figure 11 shows that in the experiment done for the month of March, the sub-lineages that are present in the South European reference set cause the higher error rates presented in Figure 8. No sub-lineages are present in the East Asian reference set and it thus performs better. Both the East Asian and the South European reference sets seem to be similarly affected by sub-lineages they might contain.

3.3 German-British reference set has the worst performance

Figure 12 shows that the European reference set performs best. A marginal difference in performance is observed between the French and German-British reference sets. However, the British-German reference set has the highest average relative prediction error. None of the reference sets contains sub-lineages of the B.1.258 lineage that is measured.

The results presented in Figure 12 suggest that, it is not enough to just consider the majority of interactions of a country and that it is safer to consider the entire continent the country had interactions with. More such experiments will add confidence to the results.

3.4 Including the latest sequences available in the reference set improves results

As shown in Figure 13 the error rate follows a decreasing trend in most cases as the time interval shortens between the collection times for the reference set and the test set. The data set that corresponds to the time span between the first and the fifteenth day of April even though being the second data set closest to the collection date of the test set, has the highest error rate. Another exception to that trend is the data set

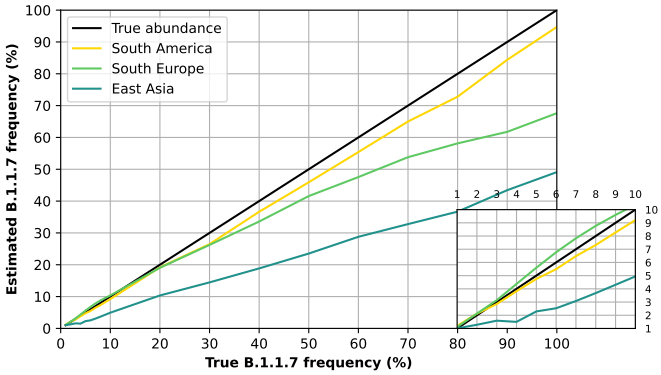
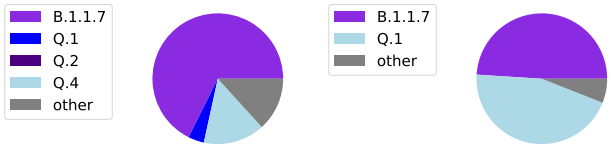


Figure 9: True abundance compared to the estimated abundance for the experiment that covers immune response-related mutations. The results correspond to the experiment done for the month of May. The reference sets are constructed using genome sequences collected from the geographical locations of South America, South Europe, and East Asia.



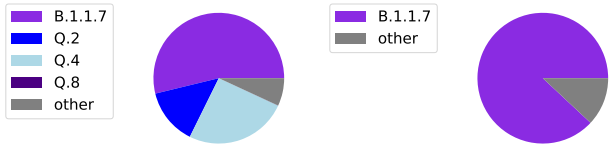
(a) South European reference set (b) East Asian reference set

Figure 10: Abundance estimates for the simulated abundance of 100% of the B.1.1.7 lineage sourced in Argentina for the month of May 2021.

that corresponds to the time span between the first and the fifteenth day of February since it performs better than some data sets that have collection dates closer to the collection dates of the lineage measured in the test set. Despite the aforementioned exceptions, results suggest that considering sequences collected closest to the day of measuring the lineage abundance in a specific geographical location, can benefit results.

3.5 A higher number of B.1.1.7 sequences in the reference set improves results

As shown in Figure 14, in most cases, the results improve with more B.1.1.7 sequences. However, the error rate does not follow a linearly decreasing trend. The reference sets that contain two to eight B.1.1.7 sequences have similar error rates. A rapid decrease in the error rate is observed for the reference set with ten sequences thus when two more sequences are added in the reference set that contains eight B.1.1.7 sequences. When only those last two sequences are included in the reference set, the performance is better than in the case of the reference sets with two up to eight B.1.1.7 sequences. Those findings suggest that performance improves with a higher number of reference sequences for the lineage measured.



(a) South European reference set (b) East Asian reference set

Figure 11: Abundance estimation predictions for the simulated abundance of 100% of the B.1.1.7 lineage sourced in Argentina for the month of March 2021.

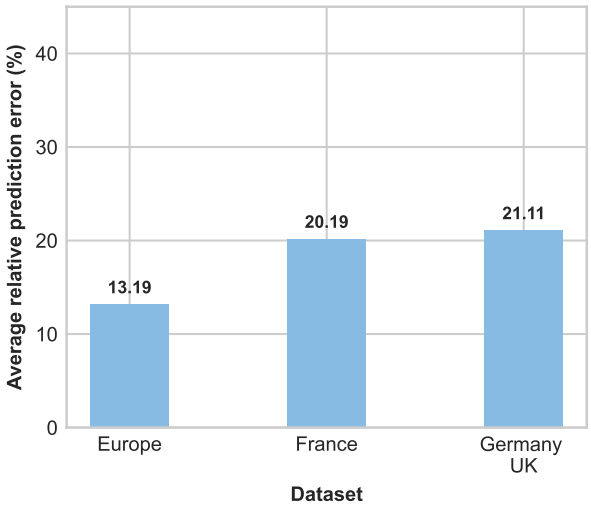


Figure 12: Average relative error percentage for the abundance predictions when using each of the reference sets presented in the figure. The test set is built with samples sourced from Cyprus.

4 Discussion

In this study, we aim to identify an optimal way of sourcing reference genome sequences so that the lineage abundance estimation of SARS-CoV-2 in wastewater improves. Previously, in section 2 the experiments conducted in order to achieve this objective are described. In this section, the results presented in section 3 for each of those experiments are discussed.

4.1 Underestimation when sub-lineages are present in the reference set

The presence of sub-lineages seems to be responsible for the abundance underestimation that is observed mainly in high simulated abundances across the experiments presented in section 3. What happens is that some of the reads of the lineage measured in the simulated wastewater data are assigned to sub-lineages that are present in the reference set. However, not all reference sets that contain sub-lineages suffer from this. This can be seen in Figure 5. Both reference sets contain the same sub-lineage but in the case of the continent-specific reference set, results show that is not as affected by its presence. This is not the case for the country-specific reference

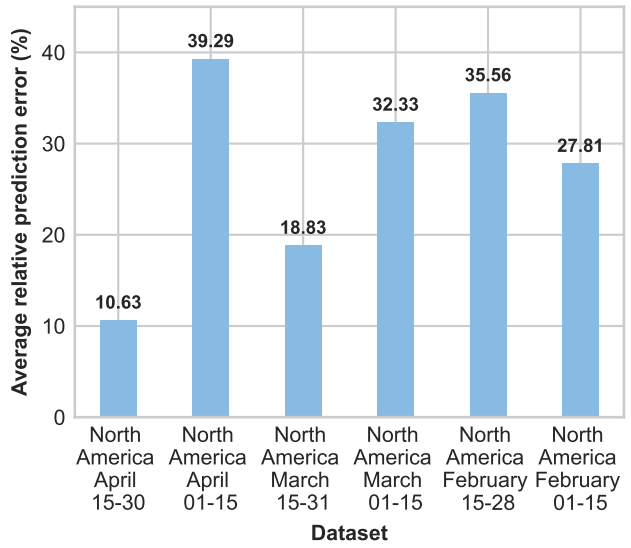


Figure 13: Average relative prediction error per reference set. The reference sets are constructed with a time span of approximately fifteen days for consecutive time periods from April till February 2021. All samples are sourced from North America. The test set is formed from samples sourced in Connecticut during May 2021.

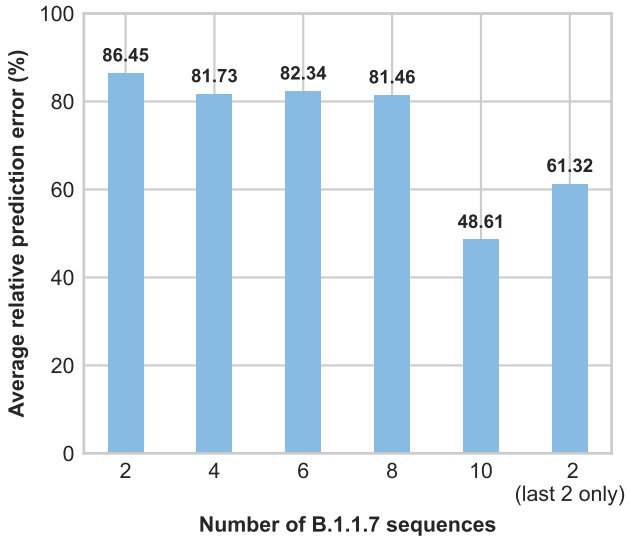


Figure 14: Average relative prediction error for the reference sets with varying amounts of reference samples of the lineage measured.

set, which is significantly affected.

The results obtained by the continent-specific reference sets in the proximity experiments suggest that they are not prone to underestimation (Figure 7). Instead, they are more likely to show overestimation in lower abundances. This can be seen in the proximity experiments done for the state of Telangana in India and the state of Massachusetts in the USA. The higher error rates observed in those two cases are due to overestimation in lower abundances.

Preliminary findings presented in Appendix A, suggest that measuring the SARS-CoV-2 quantities in wastewater for a family of lineages improves the abundance estimates.

4.2 Geographical proximity

The continent-specific reference sets seem to be the most reliable option. What seems to matter is the similarity between the lineages in the reference set as well as between the lineages in the reference set and the test set.

Locality

In the case of local reference sets, they seem to be prone to overestimation but mainly to underestimation for all abundances simulated. If all lineages in a given geographical location are more similar between them, this results in less divergence between lineages in the reference set and in the test set. Some reads of similar lineages that exist in the background could be assigned to the lineage measured thus, some overestimation is observed. In most cases, however, what seems to happen is that reads of the lineage measured are similar to multiple lineages in the reference set, thus underestimation is observed.

Global reference sets

Global reference sets suffer from high error rates due to overestimation in low abundances but perform well on high abundances. Perhaps a global reference set does not contain as many lineages that are found exclusively in certain geographical locations. Again, lineages that are in the background could be similar to the lineage measured. In such cases, in low simulated abundances of the lineage measured, those lineages that exist in the background are highly abundant in the simulated wastewater data. If their corresponding lineage does not exist in the reference set and/or are similar to the lineage measured, then they can be wrongly assigned to the corresponding sequences for lineage measured in the reference set.

Continent specific reference sets

The continent-specific reference sets perform well in most experiments conducted. This is perhaps because they can provide the right amount of contrast between the lineages in the reference set while having the lineage measured being similar enough to the corresponding sequences in the reference set. Enough contrast in the reference set means that the lineages are more divergent between them due to non-defining region-specific mutations since multiple countries in the continent are considered. At the same time, continent-specific reference sets are ”local” enough to contain the corresponding lineages for the lineages that exist in the background. Thus, those background lineages are not so often being falsely considered as the lineage measured.

Other factors

Measures taken to contain the virus as well as population density could result in states or countries experiencing the pandemic differently. Depending on such factors, certain reference sets could perform better or worse at given times.

Stringency levels for measures implemented It is possible that the states studied had implemented different stringency levels of measures to control the spread of the virus for the time periods considered. Such different levels of measures influence mutation frequencies [2] as well as the state interactions with the rest of the world. This situation could influence the similarity of the lineages of a given state with the contents of reference sets constructed using sequences from other parts of the world.

Population density Population density positively correlates with the number of infections. Research conducted specifically for India and the USA indicates that higher population density facilitates disease transmission and it thus leads to higher infection rates [15; 4]. Higher infection rates could result in lineages showing more non-defining mutations. This can impact the performance of certain reference sets.

4.3 Interactions between countries

Cyprus reported having travelers in July and August mainly from European countries. Even though the majority of those, were coming from the UK and Germany for the month of August, the fact that the German and British reference set performs marginally worse than the French reference set even though no travelers were reported from France, suggests that interactions can be hard to capture. The European data set performs best, perhaps because it has better chances to capture the interactions that brought the most abundant version of a specific lineage at a specific time. That being said, it could be the case that there is a more specific geographical location that could achieve better results. However, this experiment shows that there is no reliable way of finding such location. It is safer to consider the entire continent the country had interactions with.

4.4 Increased number of the B.1.1.7 sequences in the reference set

The findings from this experiment presented in subsection 4.4, can be partially explained due to the fact that for the wastewater simulation only one genome sequence is used. When working with real wastewater data, it is expected that there will be a greater variety in the lineages measured. Then, perhaps the decreasing error rate will follow a more linear trend.

Another interesting observation is that the reference set with four B.1.1.7 sequences has slightly better performance than the reference set with six B.1.1.7 sequences. This reveals that Kallistos' [5] likelihood function that uses the expectation-maximization algorithm to compute the probability that a specific read belongs to a specific reference transcript, distributes those probabilities slightly differently given a different number of sequences in the reference set.

All in all, the experiment shows that the number of reference sequences does not necessarily translate to higher performance. Perhaps by including more sequences, one increases the chances of having a good sequence included in the reference set. Finally, different sizes of reference sets could influence performance negatively or positively.

5 Methods

In this section, the experimental setup and methods are described. First in subsection 5.1 the workflow followed by this study is illustrated and the steps are explained. Next, in subsections 5.2, 5.3 and 5.4 design choices made such as the genomic region, the simulated abundances and the simulated coverage are documented and discussed. In 5.5, the metrics used to evaluate the reference sets created, are presented. Finally, in subsections 5.6 and 5.7, the data collection methods are described.

5.1 Pipeline

In this subsection, the steps of the pipeline used for the lineage abundance estimation are illustrated in Figure 15 and then described.

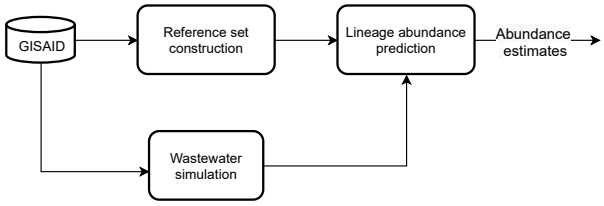


Figure 15: Steps of the pipeline used in this study to quantify the abundance of SARS-CoV-2 in wastewater data.

Reference set construction

First, the sequences and the corresponding metadata are provided as input to the pipeline. Quality filters are applied. Then the sequences are compared to the original SARS-CoV-2 reference (NC_045512.2, collected from the NCBI database). Next, the allele frequencies per lineage are computed and are used for the selection of the sequences for each lineage. All mutations with an allele frequency of at least 50% are captured at least once.

Wastewater simulation

In order to test the performance of the reference sets created, wastewater sequencing data are simulated using ART [9]. Genome sequences collected in a course of a month for a specific geographical region are sourced from GISAID [7]. All sequences that are of the same lineage with the lineage to be measured are then removed from that data set. The remaining sequences are used to simulate the background lineages for the wastewater simulation. A genome sequence that is collected during that same time period for that same geographical location is sourced from GISAID [7] to represent the lineage to be simulated.

Lineage abundance predictions

The Kallisto index (de Bruijn graph) is built for the filtered reference set. Kallisto [5] quantifies the abundance per reference sequence for the simulated wastewater data. Finally, the Kallisto output is processed to obtain the abundance estimates per lineage.

5.2 Genomic region

For the purpose of the experiments, the whole genome is used as input. Even though, spike-only sequencing of the genome yields better results with simulated data [3], when working with real sequencing data, it is usually needed to sequence the whole genome in order to gather enough information. Moreover, it is suggested that there are mutations in the whole genome that could be related to immune responses [16]. Finally, it is recommended that mutations in the whole-genome change frequencies over time in different geographical locations [12]. Thus, the whole genome is used.

5.3 Simulated abundances and abundance threshold

The lineage chosen to be measured in each experiment is simulated in low (scale of one to ten) and high (scale of ten to hundred) abundances. The abundance threshold applied is 1% on the overall sequence abundance.

5.4 Coverage

The coverage simulated is 100x. Baaijens et al. [3] suggest that using coverage of 100x achieves comparable results with higher coverage. It is also computationally not as intensive as other higher coverage options.

5.5 Metrics

The metrics used to measure the performance of each reference set, are namely the relative prediction error (Equation 1) and the average relative prediction error (Equation 2).

$$\frac{|\text{true abundance} \ - \ \text{estimated abundance}|}{\text{true abundance}} \tag{1}$$

$$\frac{Sum_{a \ \in \ \text{Abundances}} \ (\text{Relative prediction error}_a)}{N_{\text{Abundances}}} \tag{2}$$

5.6 Data availability

All data used for the reference set and test set construction are downloaded from the GISAID database [7]. All GISAID identifiers are available on the Gitlab repository [1].

5.7 Data collection methods

All data are collected from GISAID. In order to reduce the number of sequences, filters available on GISAID were applied. Those filters exclude incomplete sequences, sequences where patient status is missing or the date is incomplete. This was mainly done for the country level and the continent-specific reference sets. For the construction of the close-by states reference sets and for cases where a combination of countries was needed, data were downloaded from GISAID and then merged. In the case of global data sets, the global data set available on GISAID (next-regions) was downloaded and then filtered to contain sequences only for the needed time periods.

If global reference sets were to be sourced in the same way as the rest of the data sets, results could have perhaps been different in some cases. However, the reference sets would have been much larger in size which would have resulted in the process being more computationally intensive. Most importantly, such global reference sets would have been heavily biased towards countries and continents that sequence and submit more especially at earlier time periods during the pandemic.

[1] https://gitlab.ewi.tudelft.nl/jbaaijens/CSE3000_wastewater_project.git

6 Responsible research

In order to ensure the reproducibility of the results produced by this study, the tool Snakemake [10] is used. Moreover, all code is available on the Gitlab repository. In any parts of the code where randomization is involved, random seeds where used. Therefore, the reader should be able to reproduce the exact same results by using the same random seeds which can be found on the Gitlab repository. The genome data used in this study can be downloaded from GISAID [7]. As it is forbidden to publish the genome data, the GISAID identifiers for all data sets used are available on the Gitlab repository.

In this study, an effort has been made so that the experiments will allow the reader to generalize over the conclusions and use them for a wide range of geographical regions. Also, there has been an effort to include various geographical locations into the experiments so that confidence in the results can be gained for more places around the globe. However, due to limitations in data sources, this was not entirely feasible, thus the most studied continents are Asia, Europe, North and South America. By conducting additional experiments that include other geographical regions, more confidence will be gained in the results of this study and their applicability in more geographical areas.

It is at the moment unclear whether the lineage classification and labeling for the data being submitted to GISAID [7] are updated and if so, how often the updating process takes place. In a situation where data are not frequently updated, new lineages could be mislabelled in the database as similar known lineages. The accuracy of the classification process for the genome sequences available in GISAID [7] is also unknown. Those uncertainties could impact the accuracy of the results obtained by this study.

Finally, in this study, wastewater data were simulated in order to test the performance of each reference set used. It is encouraged that the experiments be repeated using real wastewater data in order to gain more confidence in the results in real-world settings.

7 Conclusion

Continent specific reference sets are the most reliable and good performing choice for estimating the lineage abundance of SARS-COV-2 in the wastewater of a community. In case an optimal reference set is not available or in the case of a remote geographical location, it is worth considering the overall interactions of that geographical region. It is suggested that ancestry and thus immune response-related mutations do not influence performance. It is also recommended that a higher number of sequences per lineage in the reference as well as the inclusion of recently sourced sequences improve results.

Finally, preliminary results indicate that considering a family of lineages or the variants of concern/interest as they are defined by the World Health Organisation yields better results since most underestimation and overestimation happen due to sub-lineages in the data sets.

7.1 Future work

In this study, suggestions have been made mainly for the regions the samples should be sourced from given a geographical location that lineage abundance estimations should be made. One could further investigate topics such as optimal reference set size as well as the optimal amount of time the reference set should cover in order to achieve best results. It is recommended to repeat the experiments using real wastewater data. This will add confidence in the results obtained for real-world settings.

Detection of a new variant

The pipeline used by this study can only detect known lineages that are contained in the reference sets used. The pipeline aligns the reads from the wastewater data to the most similar lineages in the reference set. In this case, if a new lineage is more similar to a non-highly abundant lineage then an increase in the abundance of that lineage will be observed. Then, differential expression analysis could be conducted to identify if it is the already known lineage or a similar new one. In case the new lineage is most similar to an already highly abundant lineage that would make it harder to detect.

Another approach to this problem is the use of Generative Artificial Intelligence. Generative Artificial Intelligence could be used to predict and generate sequences of lineages that might appear in the future. This would require the evolution of the virus to be predicted and the genome sequences that would be produced by such a tool to be more similar to the new lineage than other known lineages.

References

[1] WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int, visited on 2022-01-09.

[2] Santiago Justo Arevalo, Daniela Zapata Sifuentes, César J. Huallpa, Gianfranco Landa Bianchi, Adriana Castillo Chávez, Romina Garavito-Salini Casas, Carmen Sofia Uribe Calampa, Guillermo Uceda-Campos, and Roberto Pineda Chavarría. Dynamics of sars-cov-2 mutations reveals regional-specificity and similar trends of n501 and high-frequency mutation n501y in different levels of control measures. *Scientific Reports 2021 11:1*, 11:1–11, 9 2021.

[3] 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, 9 2021.

[4] Arunava Bhadra, Arindam Mukherjee, and Kabita Sarkar. Impact of population density on covid-19 infected and mortality rate in india. *Modeling Earth Systems and Environment*, 7:623–629, 3 2021.

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

[6] Nitin Chitranshi, Vivek K. Gupta, Rashi Rajput, Angela Godinez, Kanishka Pushpitha, Ting Shen, Mehdi Mirzaei, Yuyi You, Devaraj Basavarajappa, Veer Gupta, and Stuart L. Graham. Evolving geographic diversity in sars-cov2 and in silico analysis of replicating enzyme 3clprotargeting repurposed drug candidates. *Journal of Translational Medicine*, 18:1–15, 7 2020.

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

[8] Julian R. Homburger, Andrés Moreno-Estrada, Christopher R. Gignoux, Dominic Nelson, Elena Sanchez, Patricia Ortiz-Tello, Bernardo A. Pons-Estel, Eduardo Acevedo-Vasquez, Pedro Miranda, Carl D. Langefeld, Simon Gravel, Marta E. Alarcón-Riquelme, and Carlos D. Bustamante. Genomic insights into the ancestry and demographic history of south america. *PLOS Genetics*, 11:e1005602, 2015.

[9] Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 12 2011.

[10] F Mölder, KP Jablonski, B Letcher, MB Hall, CH Tomkins-Tinch, V Sochat, J Forster, S Lee, SO Twardziok, A Kanitz, A Wilm, M Holtgrewe, S Rahmann, S Nahnsen, and J Köster. Sustainable data analysis with snakemake [version 1; peer review: 1 approved, 1 approved with reservations]. *F1000Research*, 10(33), 2021.

[11] Yohann Nédélec, Joaquín Sanz, Golshid Baharian, Zachary A. Szpiech, Alain Pacis, Anne Dumaine, Jean-Christophe Grenier, Andrew Freiman, Aaron J. Sams, Steven Hebert, Ariane Pagé Sabourin, Francesca Luca, Ran Blekhman, Ryan D. Hernandez, Roger Pique-Regi, Jenny Tung, Vania Yotova, and Luis B. Barreiro. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell*, 167(3):657–669.e21, 2016.

[12] Maria Pachetti, Bruna Marini, Francesca Benedetti, Fabiola Giudici, Elisabetta Mauro, Paola Storici, Claudio Masciovecchio, Silvia Angeletti, Massimo Ciccozzi, Robert C. Gallo, Davide Zella, and Rudy Ippodrino. Emerging sars-cov-2 mutation hot spots include a novel rna-dependent-rna polymerase variant. *Journal of translational medicine*, 18, 4 2020.

[13] Lucy Panayidou. Press releases, Sep 2020. https://www.pio.gov.cy/en/press-releases-article.html?id=15731, visited on 2022-01-09.

[14] Lucy Panayidou. Press releases, Aug 2020. https://www.pio.gov.cy/en/press-releases-article.html?id=15263, visited on 2022-01-09.

[15] Karla Therese L. Sy, Laura F. White, and Brooke E. Nichols. Population density and basic reproductive number of covid-19 across united states counties. *PLOS ONE*, 16:e0249271, 4 2021.

[16] Rui Wang, Yuta Hozumi, Yong-Hui Zheng, Changchuan Yin, and Guo-Wei Wei. Host immune response driving sars-cov-2 evolution. *Viruses*, 12, 10 2020.

A Abundance estimates in lower granularity

Figure 16 shows the average relative prediction error (Equation 2) for each of the proximity experiments done for states in the USA. These experiments are similar to the experiments described in subsection 2.1. The difference is that abundance predictions are made for a family of lineages.

The B.1.1.7 lineage is simulated. All B.1.1.7 sequences are removed from the test set. The sublineages of the B.1.1.7 lineage (Q.x) are also removed. Abundance is measured from both B.1.1.7 and Q.x sequences in the reference set for the final abundance estimates. As we observe in Figure 16 results mostly improve for all experiments and all reference set types. Continent specific reference sets continue to perform well. The results of the global reference sets are also significantly improved.

More such experiments and throughout analysis of the results will help identify the optimal granularity one should use for more accurate abundance predictions.

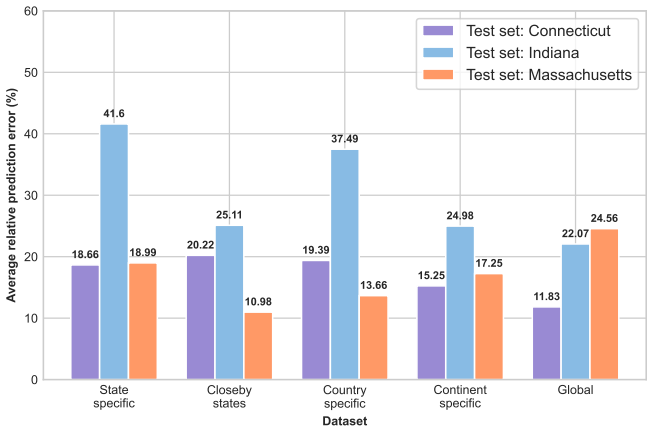


Figure 16: Average relative prediction errors for proximity experiments conducted for states in the USA. The B.1.1.7 lineage is simulated. Abundance estimates are measured from B.1.1.7 and Q.x sequences in the reference set