Investigating the effect of sampling bias on SARS-CoV-2 phylogenetic inference

David H Yang

CMMB 530

Dr. Paul Gordon, Dr. Michael Hynes, Dr. Quan Long

Abstract:

The SARS-CoV-2 virus quickly became a global health threat as it rapidly spread across the world. To date, there have been more than 130 million reported cases and almost 3 million reported deaths worldwide. Part of the difficulty containing SARS-CoV-2 is identifying infected individuals that exhibit mild or no symptoms but contribute significantly to disease spread. Traditionally, phylogenetic analysis has been a valuable tool in mapping out the transmission of viruses and aiding with containment. Unfortunately, most of the phylogenetic studies conducted thus far for SARS-CoV-2 have been limited by the presence of sampling bias due to differential sequencing rates and SARS-CoV-2 prevalence in different regions around the world and time frames. In our study, we devised a weighted sampling strategy to minimize the aforementioned sampling bias. Our weighted sampling strategy significantly altered the number of sequences obtained from each region during each time frame and produced representative sequence samples. Consequently, phylogenetic trees produced from our weighted sampling strategy demonstrated signs of minimized sampling bias compared to phylogenetic trees produced from random sampling.

Introduction:

In December 2019, the SARS-CoV-2 virus was detected in Wuhan, China[(1)]. The SARS-CoV-2 virus quickly became a global health threat as it rapidly spread across the world. SARS-CoV-2 is a positive sense, single-stranded RNA virus in the family *Coronaviridae*. The SARS-CoV-2 genome is about 29.9 kb long. Other viruses in the family *Coronaviridae* include SARS-CoV, MERS-CoV and four other strains of coronavirus that cause the common cold in humans (OC43, NL63, 229E, HKU1). Unlike the other viruses, SARS-CoV-2 has spurred a global pandemic due to its high infectivity (R0 = 2.24-3.58) compared to the other coronaviruses during the early stages of the pandemic[(2)]. To date, there have been more than 130 million reported cases and almost 3

million reported deaths worldwide[3]. Unfortunately, these numbers only represent a fraction of the cumulative number of cases and deaths due to SARS-CoV-2 as many cases go unreported. Estimates claim that the true number of cases and deaths due to SARS-CoV-2 are 10.5 times and 1.47 times the reported amount respectively[4]. Thus, it is of the utmost importance to contain the spread of SARS-CoV-2. Part of the difficulty containing SARS-CoV-2 is identifying infected individuals that exhibit mild or no symptoms but contribute significantly to disease spread[5].

Traditionally, phylogenetic analysis has been a valuable tool in mapping out the transmission of viruses including in a previous SARS-CoV outbreak in 2009[6]. In phylogenetic analysis, phylogenetic trees are produced based on the genetic relatedness between different strains of viruses. Through such, a corresponding viral transmission network can be inferred based of the predicted evolutionary relationships between the various variants. These transmission networks are essential for allowing us to better understand the spread of viruses and coordinating preventative measures. Early phylogenetic studies conducted by Forster and colleagues have been able to identify transmission links between China and other geographic locations across the world[6]. Currently, the Nextstrain group has also done a lot of phylogenetic analyses for SARS-CoV-2[7]. Nextstrain has mapped their predicted evolution and transmission of various SARS-CoV-2 variants which is shown Figure 1. In the figure, clades 19A and 19B emerged in Wuhan, China and dominated the early outbreak. Meanwhile, clade 20A later dominated the European outbreak in March and spread across the world. Finally, clades 20B and 20C are subclades that evolved from 20A in early 2020 and eventually further evolved into other variants during later 2020.

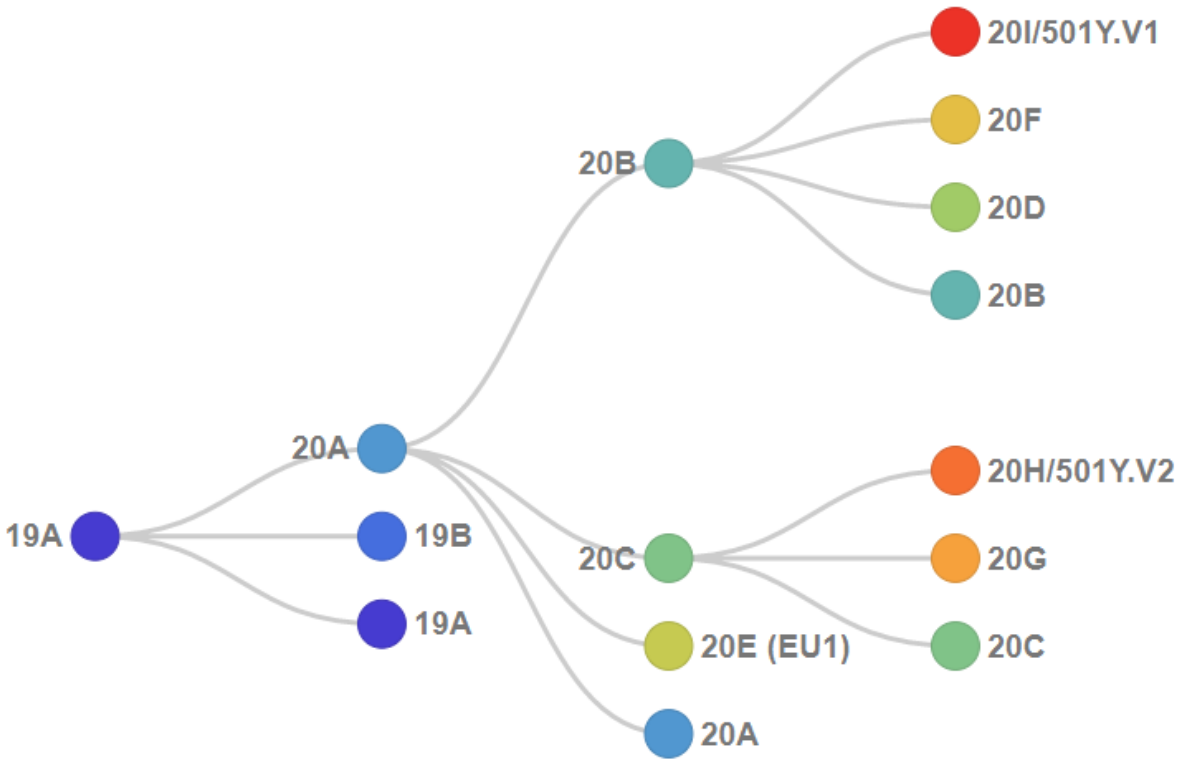

Figure 1: Nextstrain's SARS-CoV-2 variant evolution prediction grouped into clades 19A, 19B, and 20A to 20I. Each clade has their respective distinctive mutations[7].

Such an approach towards modelling the spread of SARS-CoV-2 has been particularly empowered in our current pandemic due to the unprecedented amounts of sequencing data made available by second generation and next generation sequencing technologies. Currently, more than 1 million SARS-CoV-2 full genome sequences are available on the Global Initiative on Sharing All Influenza Data (GISAID) database, which is one of the largest public datasets for SARS-CoV-2[8].

Unfortunately, most of the phylogenetic studies conducted thus far have been limited by the presence of sampling bias as they fail to account for the variation in infection rates and genome sequence contributions from different countries. For example, the R0 value, which indicates the

infectivity of the disease, varies region by region for SARS-CoV-2 due to factors such as government policies and access to public healthcare[9]. Consequently, cases per 1 million population varies greatly between countries such as the United States (95,993), United Kingdom (64,108) and India (9,729) as of April 11th, 2021[3]. Moreover, the number of sequences contributed by each of these countries are not representative of the number of cases in each country. For example, the United States has sequenced 294,361 sequences (~9.2 sequences/1000 cases), the United Kingdom has sequenced 367,102 sequences (~84 sequences/1000 cases), while India has only sequenced 8,110 sequences (~0.60 sequences/1000 cases)[3,8]. Consequently, this lack of consistency between the number of cases and sequences from each region introduces an intrinsic bias in SARS-CoV-2 genome databases that is problematic when standard sampling procedures such as random sampling is employed.

A traditional solution towards sampling bias would be to conduct weighted sampling where the number of genome sequences obtained from each region is proportion to the number of infections in each region. This would typically involve the use of reported case numbers as an indication of SARS-CoV-2 infection rates in each region. However, unfortunately, the reported case numbers in different regions across the world are both unrepresentative and noncomparable due to inconsistent testing rates and SARS-CoV-2 infectivity[8,9]. As such, these factors complicate weighted sampling for SARS-CoV-2 when trying to weigh sequences from different regions accordingly to its true SARS-CoV-2 prevalence.

Fortunately, many metrics that can indicate the true prevalence of SARS-CoV-2 are readily available due to intensive SARS-CoV-2 data collection. These metrics include percent positive ratios (Equation 1), case fatality ratios (Equation 2), incident fatality ratios (Equation 3), as well as seroprevalence studies.

Equation 1:

$$Percent\ positive\ ratio\ (in\ \%) = \frac{\#\ of\ positive\ test\ results}{\#\ of\ RTPCR\ tests}\ x\ 100\%$$

Equation 2:

$$Case\ fatality\ ratio\ (in\ \%) = \frac{number\ of\ deaths\ from\ disease}{number\ of\ confirmed\ cases\ of\ disease}\ x\ 100\%$$

Equation 3:

$$Infection\ fatality\ ratio\ (in\ \%) = \frac{number\ of\ deaths\ from\ disease}{number\ of\ infected\ individuals}\ x\ 100\%$$

Percent positive ratios can serve as a measure for the transmission of SARS-CoV-2 relative to the number of tests performed in a region. A high percent positive ratio (>5%) for instance, would indicate that the testing rate is too low to monitor the SARS-CoV-2 transmission within the population. Case fatality ratios (CFR) and incident fatality ratios (IFR) can also be used to gauge the prevalence of SARS-CoV-2 via comparison. This is because the two ratios differ in its denominator where the IFR accounts for both reported and unreported SARS-CoV-2 cases while the CFR only includes reported cases. Finally, serology studies can also be used to verify reported case counts as it provides an indication of the total number of individuals that have been previously infected by SARS-CoV-2 and possess corresponding antibodies. Though these metrics unveil important information regarding the infection rate of SARS-CoV-2, directly applying these metrics to estimate SARS-CoV-2 prevalence can be challenging. Fortunately, many studies have incorporated these metrics into epidemiological models to provide concrete SARS-CoV-2 prevalence estimates for regions across the world. For example, The Institute for Metrics and Evaluation's (IHME) model creates COVID-19 case predictions by fitting a death estimate to a statistical death model based off a country's IFR[10]. Such epidemiological models provide a more

accurate estimation of the SARS-CoV-2 infection rate in regions around the world and allow weighted sampling to be performed with greater confidence.

Ultimately, the purpose of this study was to develop a weighted sampling strategy to weigh SARS-CoV-2 sequences according to the prevalence of SARS-CoV-2 in each region to minimize sampling bias during phylogenetic analyses. We hypothesised that the sequences found on GISAID would be both unrepresentative of the prevalence of SARS-CoV-2 spatially and temporally. Thus, in this study, the true prevalence of SARS-CoV-2 in each location and timeframe (month) as well as the presence of intrinsic bias in GISAID was investigated. This was conducted by comparing various SARS-CoV-2 prevalence metrics to the number of sequences found in GISAID for the time range February 1st, 2020 to October 31st, 2020. Through this investigation of SARS-CoV-2 prevalence and sequencing contribution, a weighted sampling strategy was developed accordingly to the IHME SARS-CoV-2 epidemiological model. To evaluate the effectiveness of the weighted sampling strategy, phylogenetic trees produced from weighted and random sampling were verified with NextClade literature. Our results from the study suggest that significant bias was found in GISAID and that adopting a weighted sampling strategy significantly altered the number of sequences obtained from each region during each time frame. Furthermore, phylogenetic trees produced from weighted sampling demonstrated signs of minimized sampling bias compared to phylogenetic trees produced from random sampling.

Material and Methods:

Data Collection

The following data was collected: the total number of cases, total number of deaths, total number of tests, percent positivity ratios, case fatality ratios, total number of sequences on GISAID,

sequencing rate, number of participants involved in the seroprevalence studies, average seroprevalence, Imperial College London case estimation, Institute for Health Metric and Evaluation (IHME) case estimation, London School of Hygiene and Tropical medicines case estimation, and Youyang Gu's case estimation. SARS-CoV-2 cases, deaths, and testing data for each country and date was obtained from the *Our World in Data* COVID-19 dataset which compiles COVID-19 data from Johns Hopkins University and various official reports[11]. SARS-CoV-2 genomics data and associated metadata was obtained from GISAID as of March 1st, 2021[8]. SARS-CoV-2 serology test results were obtained from SeroTracker (December 2nd, 2020) which compiles all SARS-CoV-2 serosurveys across the world [12]. Finally, the epidemiological models from the Imperial College London (ICL), IHME, Youyang Gu (YYG), and The London School of Hygiene & Tropical Medicine (LSHTM) were also obtained from *Our World in Data* which compiled their respective estimates[10].

Data Processing

The SARS-CoV-2 metadata from GISAID was grouped by month and location, and then tallied to obtain the number of sequences found from each location during each month. Following, SARS-CoV-2 cases, deaths, testing, epidemiological model and serosurvey data were also grouped by location and month. During processing, serosurvey data was grouped based on the date during the middle of the survey date range. Studies with the same dates were averaged using weighted averaging according to the number of participants that were involved in each study. The percent positive ratio, CFR and IFR were calculated respectively using the case, deaths, and testing data. All data obtained were then merged into the same comma-separated values (CSV) file. The scripts used to perform data processing and produce the resulting CSV file can be found in the following GitHub repository: https://github.com/davidhy8/covid-data-processor

Random Sampling

Random sampling was performed using the Software Augur (augur filter) from NextStrain on the FASTA file and associated metadata obtained from GISAID on March 1st, 2021[13]. The following parameters were used: --min-date 2020-02-01 --max-date 2020-10-31 –subsample-max-sequences 1000.

Weighted Sampling

The following steps were performed to determine the number of desired sequences to be sampled from each region and month. First, the number of sequences to be obtained from each month globally (monthly sequencing allowance) was determined based on the proportion of SARS-CoV-2 infections that occurred in each month out of all months. Next, the number of sequences to be obtained from each country in each month was determined based on the proportion of SARS-CoV-2 infections that occurred in each country out of all countries during each respective month. The proportion of infections in each country was then multiplied by the monthly sequencing allowance to ensure that the number of sequences obtained in each month and country was proportional to their respective infection proportions. In total, 25,000 sequences were obtained. The number of SARS-CoV-2 infections during each month and in each country was estimated using the IHME epidemiological model. Using the following strategy, a CSV file containing the number of sequences to be obtained from each region and month was generated and fed into the software Nybbler which allows this weighted sampling strategy to be performed on sequencing data and its associated metadata[14]. The Nybbler Github repository is found at the following link: https://github.com/nodrogluap/nybbler

For countries that did not contain enough sequences as the desired amounts specified in the weighted sampling strategy, sequences filled by uncalled bases (N) was used to reach the target amount and assigned to their respective locations. After 25,000 samples were obtained from weighted sampling, random sampling was performed to create subsamples of about 1000 sequences using the steps outlined in *Random Sampling.*

Phylogenetic Tree Production

The TransCovid pipeline was used to create phylogenetic trees for the SARS-CoV-2 sequence subsamples[15]. In the pipeline, multiple sequence alignment was performed on all sequence subsamples using Multiple Alignment using Fast Fourier Transform[16]. Next, Beauti from the BEAST2 software package was used to set-up the parameters used in the phylogenetic analyses. In the analysis, tip dates were activated, the Generalized Time Reversible (GTR) site model with a gamma category count of 4 was used as well as a strict clock model and the Birth Death Skyline Serial model as a prior[17]. Phylogenetic analyses were ran with MCMC chain lengths of 100 million for 10 separate runs for each sample with BEAST2. The corresponding tree and log output files were then combined with LogCombiner from BEAST2 with a resample frequency of 5000. Finally, phylogenetic trees were extracted with TreeAnnotator from BEAST2 using common ancestor node heights and a 10% burn in, then visualized via Figtree[17,18]. Sequences within each phylogenetic tree were classified into NextStrain clades using the Nextclade software[7].

Results:

SARS-CoV-2 Prevalence

SARS-CoV-2 prevalence from February 1st, 2020 to October 31st, 2020 was investigated. The confirmed case counts, and epidemiological models show that SARS-CoV-2 infections greatly increased from February 2020 to October 2020 (Figure 2). Over the span of the eight months, the results show that the ratio between the IHME estimate and the confirmed case counts decreased as the pandemic progressed. Furthermore, we found that as the pandemic progressed, the percent positive ratios generally decreased for most countries. Despite so, the estimates from the epidemiological models were generally around 5 to 15 fold greater than the reported case counts.

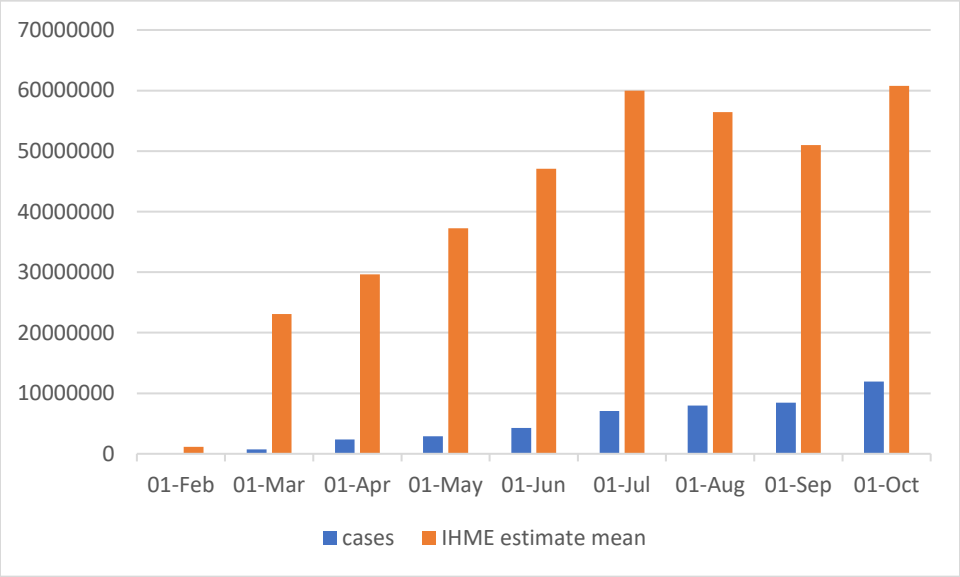

Figure 2: The number of reported SARS-CoV-2 cases and the Institute for Health Metrics and Evaluation's mean estimate of cases in the world from February 1st, 2020 to October 31st, 2020.

Due to the inconsistency between the estimates and reported case counts, other metrics were investigated to determine whether the epidemiological models were accurately estimating the

infection rate of SARS-CoV-2. In our study we found that about 106 countries had an overall percent positive ratio above the 5% standard provided by the World Health Organization (WHO). However, some countries with percent positive ratio below 5% such as Italy in May 2020 (percent positive ratio = 1.53%), also had an IHME estimate of 10-fold greater and seroprevalence estimate of 38.12% (Table 1). Similarly, this trend was also observed in Austria in April 2020 (percent positive ratio = 2.82%) which also had a 10-fold greater IHME estimate and seroprevalence estimate of 40.71%. Moreover, interestingly in our study, most countries did not have CFR ratios above their predicted IFR ratios. Meanwhile, seroprevalence estimates also varied greatly between the range of 0.1% to 44.7%. In total, 409 serosurveys were obtained. However, of those 409 estimates, 153 were from the United States and United Kingdom (Figure 3).

Table 1: SARS-CoV-2 prevalence metrics obtained for five countries (Ecuador, Austria, Italy, Pakistan, Iran) and months with the highest average seroprevalence estimate as of December 9th, 2020.

Location	***Month***	***Case count***	***Percent positivity***	***Serology tests***	***IHME estimate***
Ecuador	*May 2020*	*13896*	*36.54%*	*44.74% (N = 992)*	*573349.3915*
Austria	*April 2020*	*5746*	*2.82%*	*40.71% (N = 3076)*	*22240.91491*
Italy	*May 2020*	*29073*	*1.53%*	*38.12% (N = 17123)*	*244049.5844*
Pakistan	*June 2020*	*139841*	*19.00%*	*35.75% (N = 2045)*	*1693502.075*
Iran	*April 2020*	*52162*	*30.06%*	*33% (N = 528)*	*642326.7207*

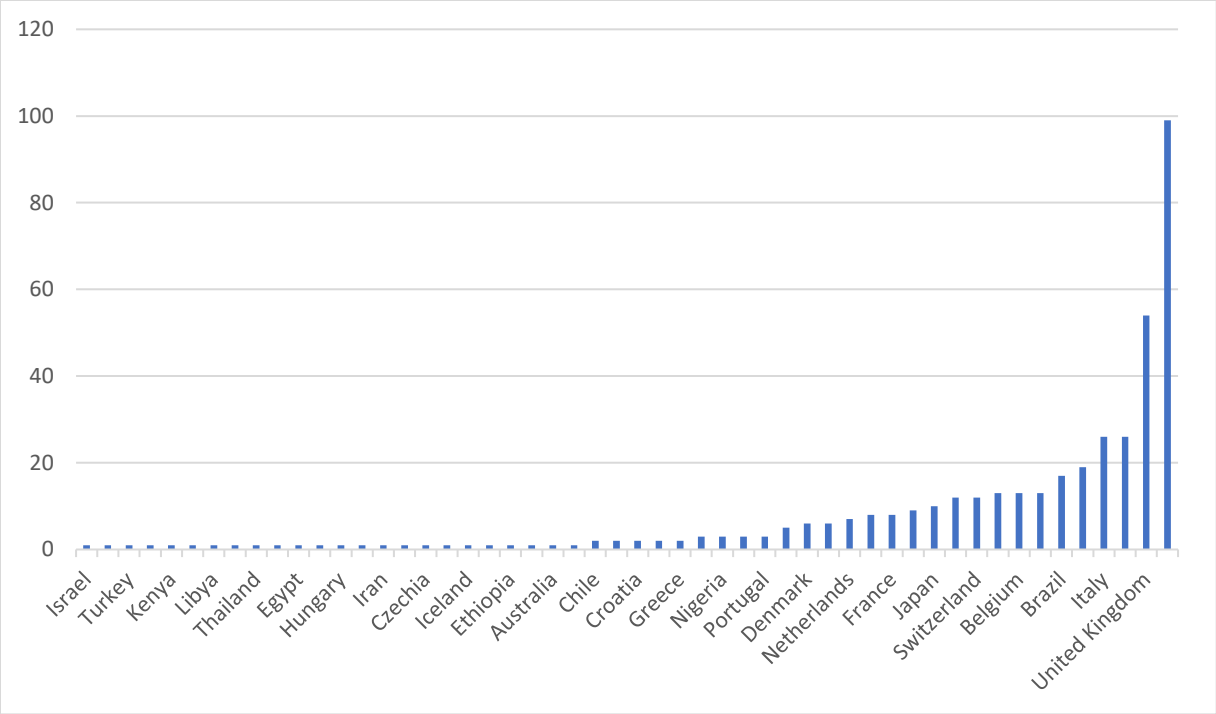

Figure 3: A tally of the number of seroprevalence studies across the world collected on SeroTracker on December 9th, 2020. The last bar and third last bar from the right on the chart represents the total number of studies from the USA and China respectively.

GISAID Sequencing Contributions

In our study, 274,182 sequences were obtained from GISAID from the date range of February 1st, 2020 to October 31st, 2020. Of those sequences, 36% (99,227) of the sequences were obtained from the UK, 25% (67,658) of the sequences were obtained from the USA, while only 1.8% (4884) and 0.9% (2503) of the sequences were obtained from India and Brazil respectively (Figure 4c). However, in terms of cases, only 0.86% of all cases were from the UK, 7.84% of all cases were from the USA, 7.05% of all cases were from India, and 4.78% of all cases were from Brazil (Figure 4b). On the other hand, different proportions were also observed for the IHME estimates. Only

1.04% of all infections were from the UK, 4.35% of all infections were from the USA, 6.09% of all infections were from India, and 5.49% of all infections were from Brazil (Figure 4a).

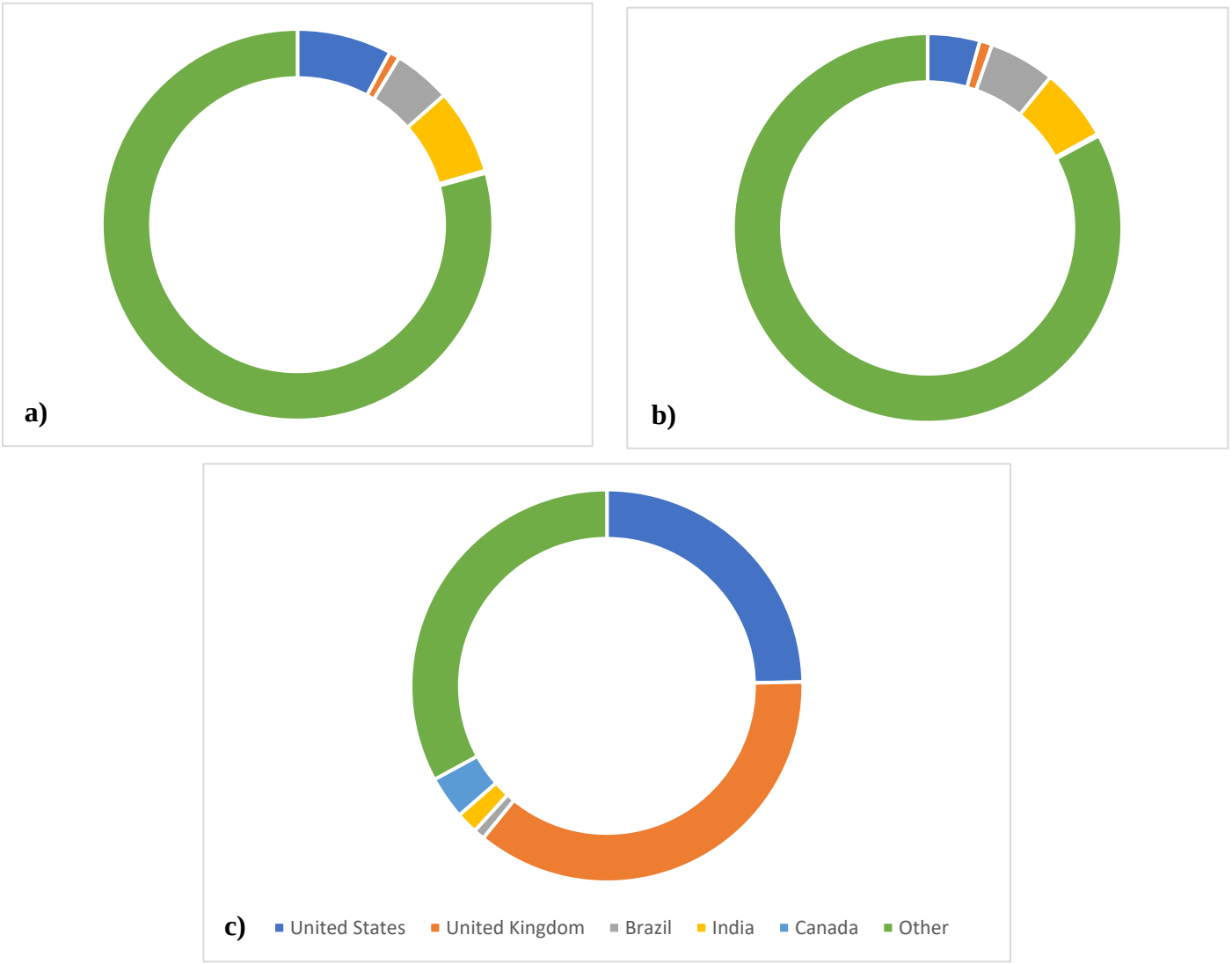

Figure 4: Doughnut charts for countries from February 1st, 2020 to October 31st, 2020. a) depicts the number of reported SARS-CoV-2 cases in each country, b) depicts the IHME estimate of the number of cases in each country, and c) depicts the number of sequences sequenced from each country.

These proportions further varied when individual months were isolated. For example, in March 2020, 26% of all sequences were from the United States, 25% of all sequences were from the UK, 1.1% of all sequences were from Brazil and 0.17% of all sequences were from India (Figure 5c).

Meanwhile, differences in the cases and IHME estimates were also seen for USA (case proportion = 23%; IHME proportion = 22%), UK (case proportion = 4.1%; IHME proportion = 12%), Brazil (case proportion = 0.6%; IHME proportion = 2.3%), and India (case proportion = 0.2%; IHME proportion = 0.8%) as well (Figure 5a & 5b).

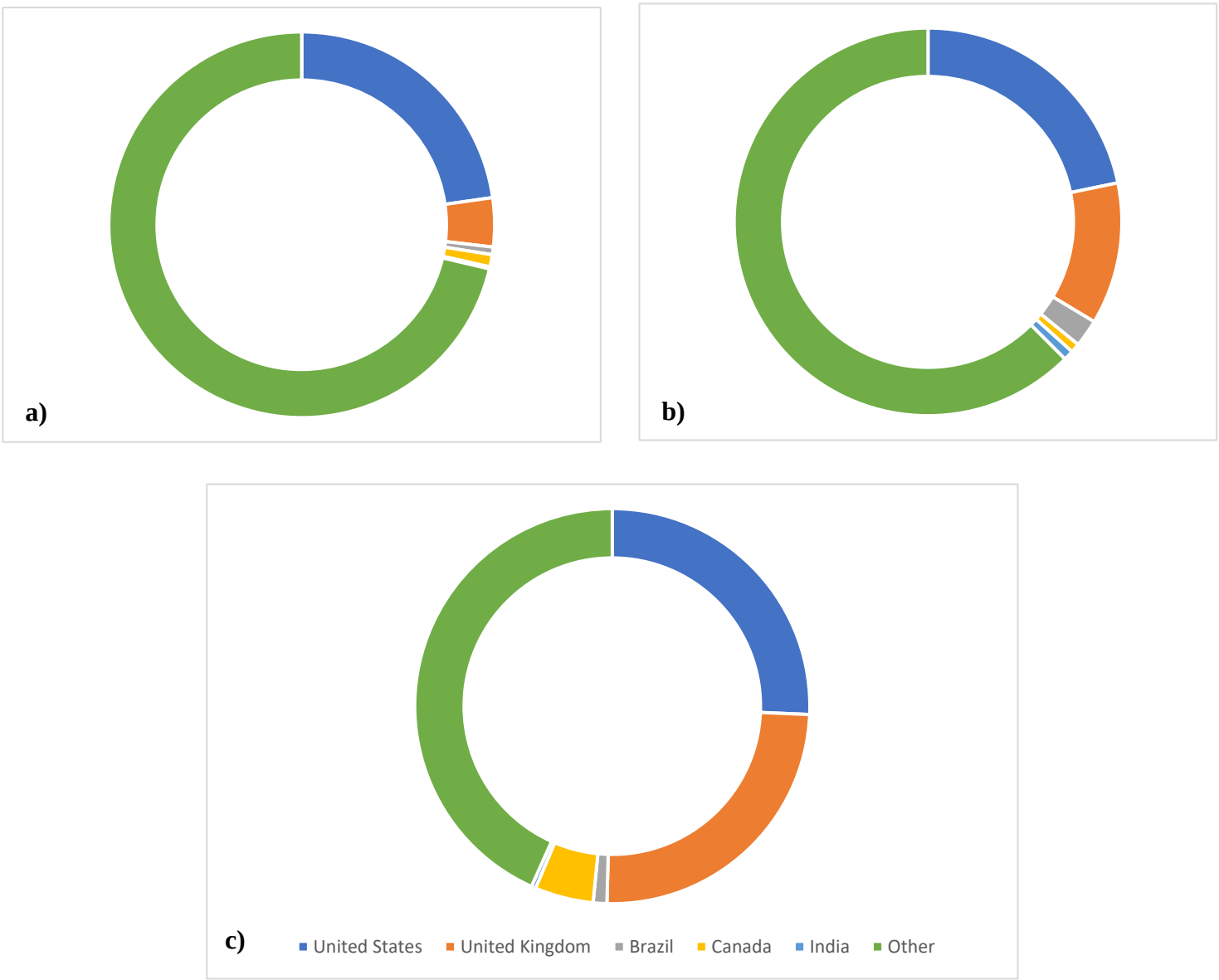

Figure 5: Doughnut charts for countries in March 2020. a) depicts the number of reported SARS-CoV-2 cases in each country, b) depicts the IHME estimate of the number of cases in each country, and c) depicts the number of sequences sequenced from each country.

Other than countries like the UK which contributed much more than other countries, during certain months of the pandemic, certain locations were also significantly undersequenced. Namely, in Poland in October 2020, only 1 sequence was obtained but 93 sequences were needed to reach the weighted sampling target (March 1st, 2021). Furthermore, this problem with underrepresentation was intensified during earlier stages of the pandemic when sequences were not uploaded to GISAID right away. For example, on December 1st, 2020, only 1 sequence was uploaded from August 2020 in Brazil. However, 1404 sequences were needed from Brazil in August 2020 to reach the weighted sampling target. Now, although this problem has been slightly alleviated as 140 sequences for August 2020, Brazil is now available (March 1st, 2021), the number of sequences available is still far from the 1404 sequence target needed to accomplish weighted sampling.

Sampling strategies

In order to account for the aforementioned differences in the sequence and estimate proportions in different countries and months, a weighted sampling strategy that accounts for temporal and spatial bias was devised. With the weighted sampling strategy, a sample of 22,238 sequences was selected from 274,182 sequences. The sample was then filled to 25,000 sequences with sequences filled by uncalled bases and assigned to their corresponding regions. After the implementation of the weighted sampling strategy, the overall proportion of sequences inside the sample from each country was consistent with their respective IHME case estimate proportions (Figure 4b & 6a). Furthermore, with the weighted sampling strategy, the proportion of sequences obtained from each country and proportion of infections (IHME) in each country remained consistent when individual months were isolated. For example, in March 2020, it is seen that the sequence and estimate proportions remained consistent (Figure 5b & 6b).

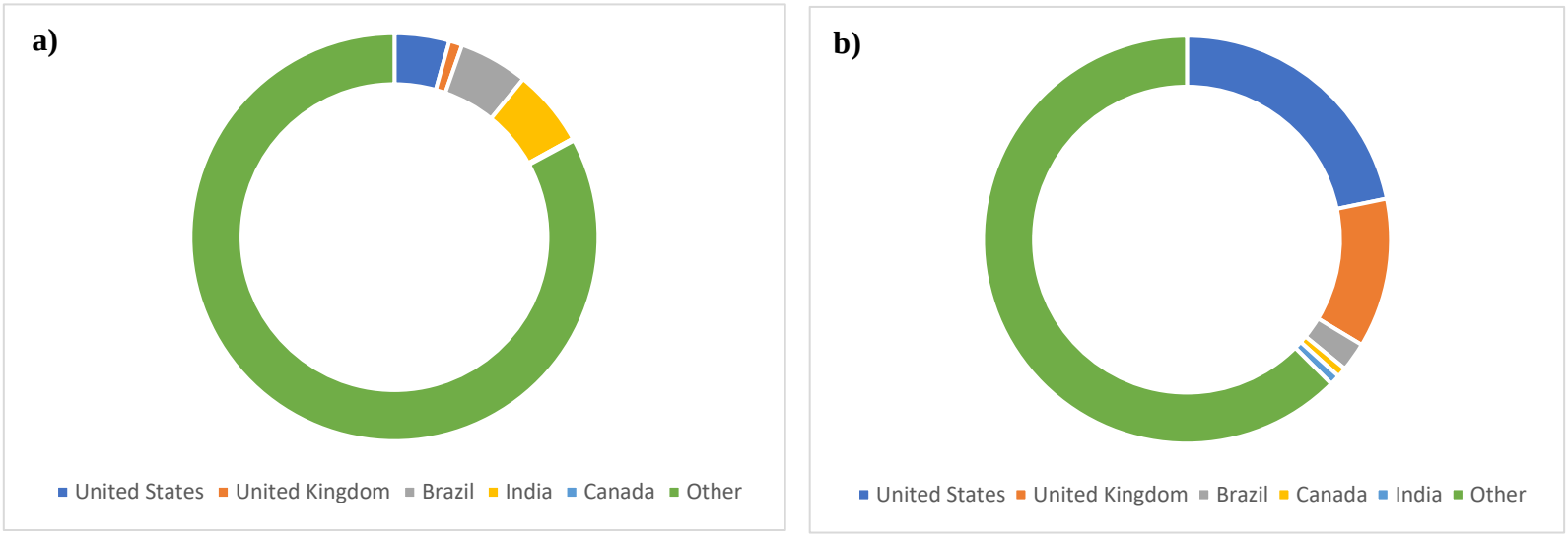

Figure 6: Doughnut chart of the number of sequences to be obtained from each country using weighted sampling to obtain a) 25,000 sequences from February 1st, 2020 to October 31st, or b) 1576 sequences from March 2020. These proportions represent the ideal number of sequences to be obtained from each country when using the weighting sampling strategy.

With the use of the weighted sampling strategy, there was significant differences in the number of sequences obtained from each region compared to if a random sequencing strategy was implemented (Figure 7). This difference was most prominent in countries such as the USA and the UK which had significant sequencing contributions. In the UK for instance, only 626 overall sequences out of 25,000 sequences were obtained in the weighted sampling strategy as opposed to 12,313 sequences if random sampling was implemented to obtain 25,00 sequences (Figure 7a). This marks a 1967% difference in sampling rates for the UK when random or weighted sampling is employed. These differences in sampling rates are also seen when individual months were isolated but in different proportions. For instance, in March 2020, 187 sequences (out of 25,000 sequences) would be sampled from the UK using weighted sampling versus 390 sequences if random sampling was employed (Figure 7b). Contrary, this only marks a 209% difference in sampling rates for the UK when random or weighted sampling is employed in March 2020.

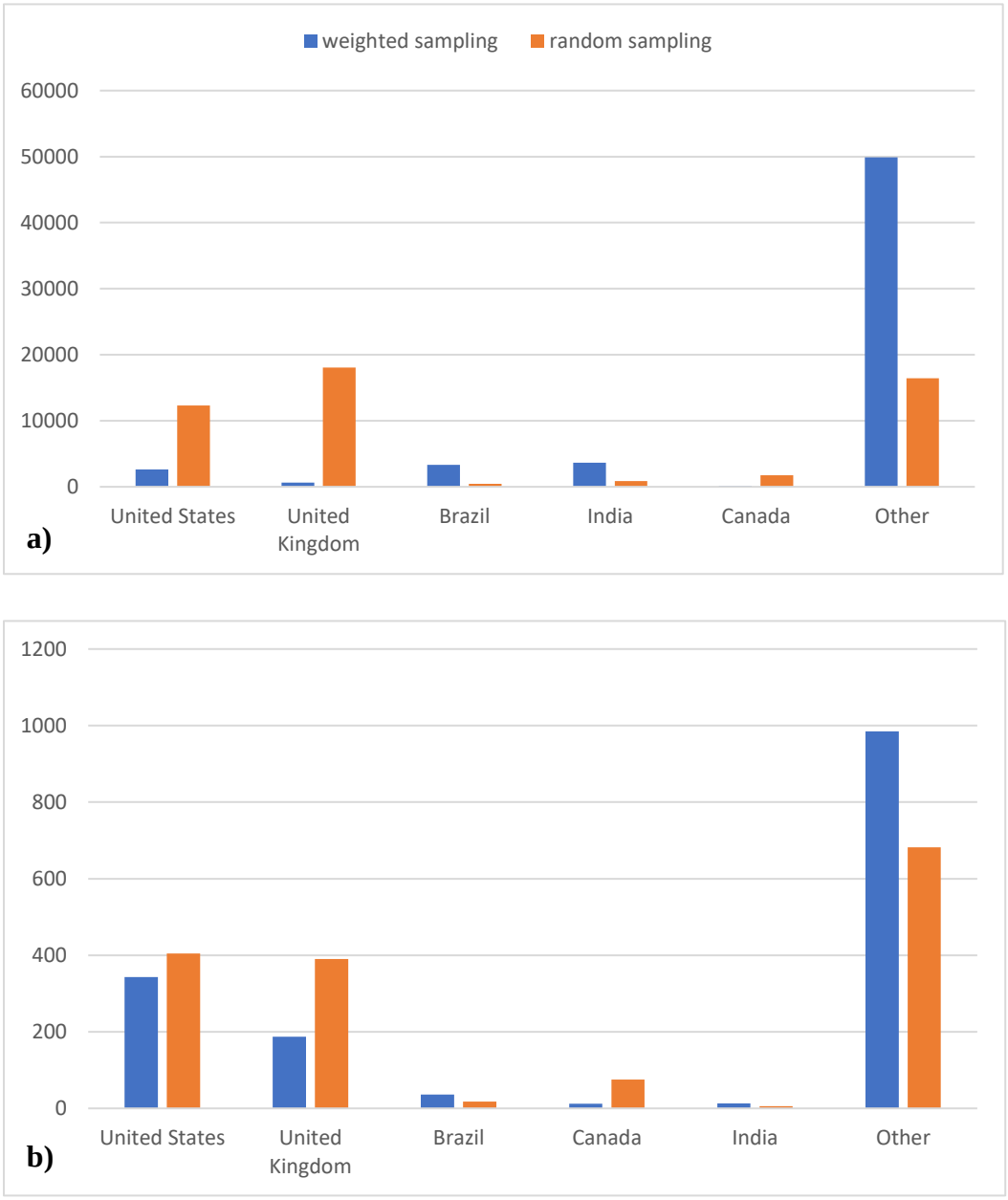

Figure 7: The number of sequences to be obtained from each country when using the random versus weighted sampling strategy. a) Depicts the number of sequences obtained from each country if 25,000 sequences were obtained from February 1st, 2020 to October 31st, 2020. b) Shows the number of sequences to be obtained from each country if 1576 sequences from March 2020 was obtained.

To perform phylogenetic analyses, random sampling was later performed on the 25,000 sequence sample to produce multiple ~1000 sequence subsamples that could be processed by the downstream phylogenetic analysis software, BEAST2. A comparison of the 1000 sequence subsamples shows that even after performing random sampling on the 25,000 sequence sample created from weighted sampling, the proportion of sequences obtained from each country shares resemblance with the proportion of infections (IHME) from each country (Figure 4a & 8a). On the other hand, the subsample obtained from random sampling resembles the sequence proportions in GISAID instead (Figure 4c & 8b).

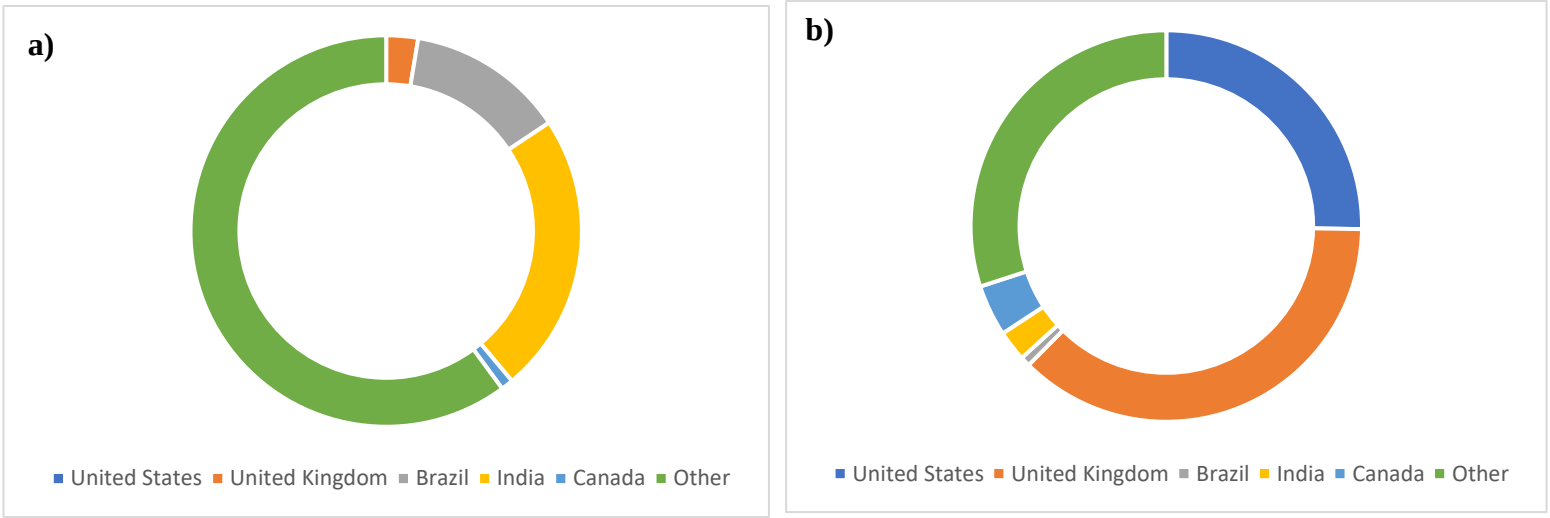

Figure 8: Doughnut chart of the number of sequences from each country in the sequence subsamples used to perform phylogenetic analyses in Figure 9 and Figure 10. a) represents the sequences obtained when using random sampling to collect 990 sequences from the pool of 25,000 sequences obtained from weighted sampling. b) represent the sequences obtained when random sampling is used to obtain 990 sequences total. All sequences selected were within the date range February 1st, 2020 to October 31st, 2020.

Phylogenetic Analyses

To compare the effects of weighted sampling and random sampling on minimizing sampling bias, phylogenetic trees for subsamples (~1000 sequences) obtained from weighted and random

sampling were produced using BEAST2. The phylogenetic trees produced from random sampling and weighted sampling are shown in Figure 9 and Figure 10, respectively. Each respective phylogenetic tree was collapsed accordingly to the clades in Nextstrain.

In the phylogenetic tree produced from random sampling, lineage 20A diverged from the root of the tree. Moreover, lineage 20A is also seen to diverge directly into lineage 20B, 20C, 20E, 19A, 19B and 20D (Figure 9). Moreover, lineage 20G is seen to diverge from lineage 20C, and lineage 20D is also seen diverging from lineage 20B. Overall, in the phylogenetic tree produced from random sampling, sequences from the Nextstrain clades 20G, 20F, 20E, 20D, 20C, 20B, 20A, 19B, and 19A were observed.

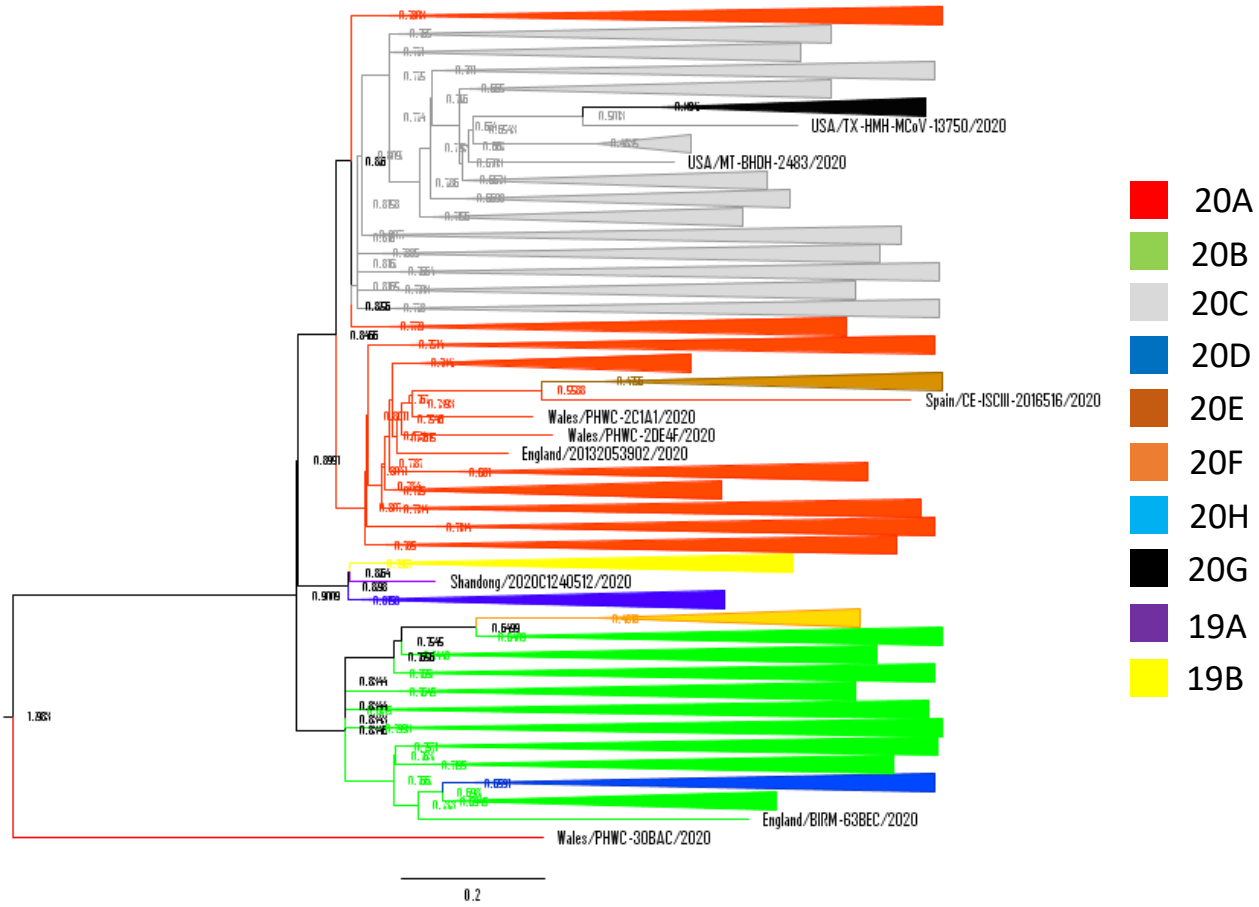

Figure 9: General clade (NextStrain) structure of phylogenetic tree produced from random sampling using BEAST2. The sequence sample used to produce this phylogenetic tree contained 990 sequences from February 1st, 2020 to October 31st, 2020.

Similarly, in the phylogenetic tree produced from weighted sampling, lineages 20A, 19A and 19B were found to diverge from the root of the tree (Figure 10). Moreover, in the phylogenetic tree, lineage 20A also diverges directly into lineage 20E, 20C and 20B. However, lineages 19A and 19B also diverges from lineage 20A. Overall, in the phylogenetic tree produced from weighted sampling the Nextstrain clades 19A, 19B, 20A, 20B, 20C, 20D, 20E, 20F, and 20H were observed.

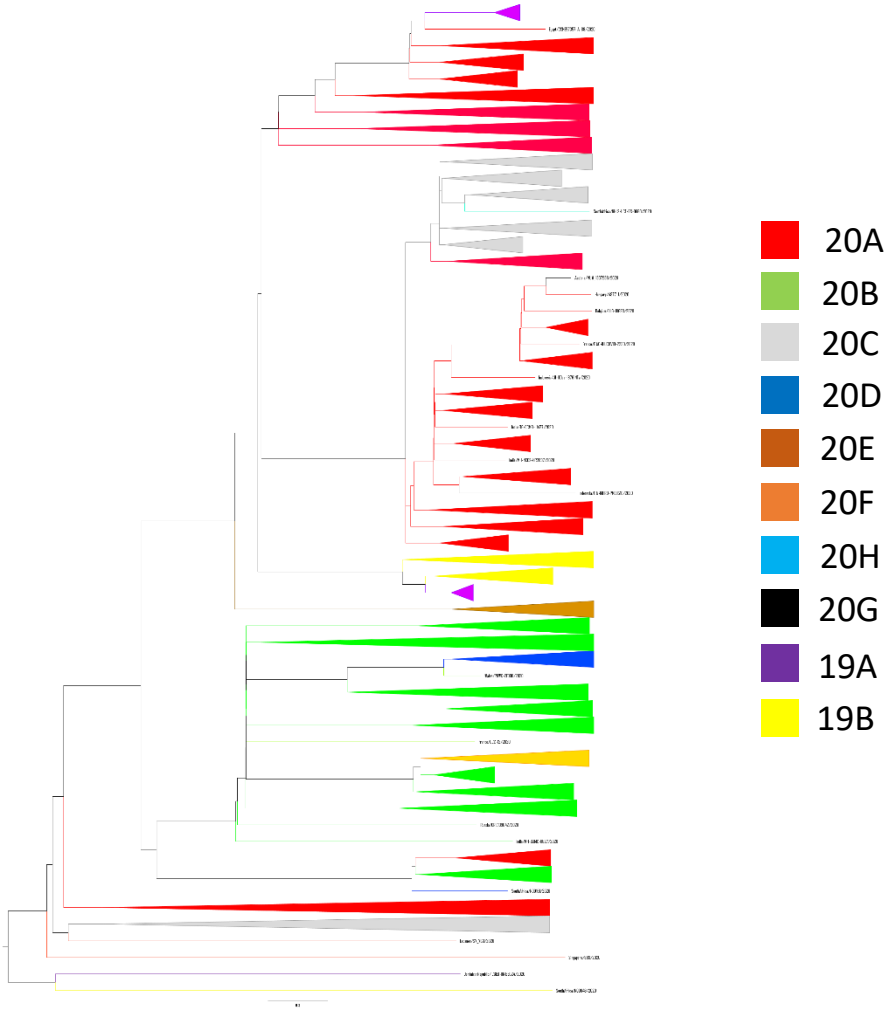

Figure 10: General clade (NextStrain) structure of phylogenetic tree produced from weighted sampling using BEAST2. The sequence sample used to produce this phylogenetic tree contained 990 sequences from February 1st, 2020 to October 31st, 2020.

Discussion:

SARS-CoV-2 Prevalence

With the progression of the COVID-19 pandemic in 2020, our results showed that SARS-CoV-2 prevalence did indeed rapidly increase from February 2020 to October 2020 (Figure 2). However, in the early stages of the pandemic (e.g. March), the IHME estimate which indicates the true prevalence of SARS-CoV-2, was found to be much greater (32X) than the reported case numbers compared to the latter months. This difference in case numbers was probably due to a lack of testing during the early stages of the pandemic to catch most cases. This likely contributed to the rapid transmission of the virus during the early stages of the pandemic. However, our results suggest that as the pandemic progressed, more cases were caught through testing, and consequently the difference between the IHME estimates and the reported case amounts decreased. Eventually, in October 2020, the IHME estimate was only around 5 times greater than the reported case amounts. These findings aligned with the increased global effort to perform more testing as the pandemic progressed, and the decreasing percent positive trend that was found for many regions such as the United States. Nonetheless, our results still suggest that more testing should be done as almost half the countries in the world have a percent positive ratio above the 5% standard provided by WHO. However, we found that the percent positive ratio could not necessarily serve as the gold standard to determine whether more testing should be performed. Take for the instance, in Italy, May 2020, the percent positive ratio indicated that enough testing was being performed despite reported case counts trailing seroprevalence estimates and IHME estimates by more than 10-fold (Table 1).

The most striking trend we found in our results was that most CFR ratios were smaller than the age-stratified IFRs provided by literature[19]. This was shocking as reported case numbers were

much lower than the actual prevalence of SARS-CoV-2, and that estimated mortality rates were predicted to be greater than the actual mortality rate of SARS-CoV-2 due to overwhelming of healthcare systems in many countries such as Italy. We believe, this trend was likely observed as the age-stratified IFRs obtained from our source of literature were much higher the IFR estimate range of 0.00% to 1.63% provided by the WHO[20]. In terms of seroprevalence results, we found in our quick survey that most serosurveys were performed in the USA or the UK, and that not enough or no serology surveys were being performed in other countries (Figure 3). Ultimately, in our study we decided to use the IHME estimate as an indication of SARS-CoV-2 prevalence due to the unreliability or lack of coverage provided by the other metrics.

GISAID Sequencing Contribution

From our findings, we found that SARS-CoV-2 sequences from the USA and the UK made up more than half the sequences found in GISAID (Figure 4c). This was especially problematic as the USA and the UK do not even make up a quarter of the SARS-CoV-2 cases or IHME estimates worldwide. On the other hand, we found that specific countries such as India and Brazil had large amounts of COVID-19 cases but relatively few sequences on GISAID. We believe that part of the issue is that many of the researchers in these countries have not uploaded their sequences onto GISAID. This is shown in the example where 139 more SARS-CoV-2 sequences were uploaded for Brazil August 2020 after December 2020. Undoubtedly, we believe that these reasons produce a bias on GISAID where countries are not being represented accordingly to the SARS-CoV-2 prevalence in each nation. These results aligned with our preliminary research and elucidated the exact sources that caused the biases in the GISAID database.

Another trend that we found was that the sequencing rates of each country varied during each month of 2020. We found that in certain months, countries would contribute more sequences than

other months. For instance, we found that in March 2020, the sequences provided by the UK made up only 25% of the sequences found that month as opposed to 36% of the sequences overall (Figure 4c & 5c). This was expected as the UK did not unveil their COVID-19 Genomics UK Consortium until March 23rd, 2020. These results suggested that there was another source of sampling bias due to differences in sequencing rates at different time frames which also needed to be accounted for during weighted sampling.

Sampling strategies

Correspondingly, the weighted sampling strategy we devised accounted for both sampling bias due to location and time. With our sampling strategy, we were able to obtain 22,238/25,000 full length sequences (~89%) from our desired locations and time. In our pool of 25,000 sequences, the sequences contributed by each location and time was correlated with the prevalence of SARS-CoV-2 in each of those locations during those times (Figure 6). Even after randomly sampling 990 sequences from the pool of 25,000 sequences, the subsample still resembles the IHME case proportions in each country more than the subsamples produced from random sampling (Figure 8). Interestingly, in the subsample produced from weighted sampling in Figure 8a, no sequences were obtained from the USA. Obviously, this marks a limitation in the strategy of performing random sampling after weighted sampling as it introduces the possibility of sampling bias due to chance. However, the greater resemblance of the subsample produced from weighted sampling still demonstrates the merits of weighted sampling. In future runs, a larger subsample size (i.e. 2000) that can be ran with BEAST2 should be aimed for to minimize the occurrence of such chance-driven potential sources of error. Ultimately, in our weighted sampling strategy, we found that the changes in sampling rates were most prominent for countries that sequenced significantly more than, or less than peers. This included countries such the UK and the USA which contributed more

sequences, as well as countries such as Brazil and India which contributed relatively few sequences. Overall, our sampling strategy was most limited when it involved countries that contributed relatively few sequences. Though sequences filled by uncalled bases (N) were used to in place of missing sequences, this strategy brings obvious limitations when trying to infer evolutionary relationships based on genetic relatedness. Ultimately, the 25,000 sequence target in our weighted sampling strategy was chosen intentionally. This number was both large enough to allow multiple random sampling runs afterwards as well as small enough so that not too many sequences are required of from undersequenced countries such as Brazil and India.

Phylogenetic Analyses

In the phylogenetic trees produced in the study, the advantages of the weighted sampling strategy were not particularly obvious. Both in the phylogenetic tree produced from the weighted and random sampling subsamples, most of the Nextstrain clades are observed. Moreover, in the phylogenetic tree produced from the weighted sampling strategy (Figure 10), the lineages that evolved from clade 20A did not align as accurately with literature as the phylogenetic tree produced from random sampling (Figure 9). However, several key characteristics in the phylogenetic tree produced from weighted sampling support the weighted sampling strategy as opposed to the random sampling strategy. First, unlike the tree produced from random sampling, the phylogenetic tree produced from weighted sampling was rooted by a sequence that evolves into the three clades 20A, 19A and 19B. This is significant as clades 19A and 19B dominated the early outbreak which was not sequenced as less emphasis was put towards sequencing at the time. On the other hand, in the tree from random sampling, the clades 19A and 19B appeared during the same time as clades 20B and 20C even though 19A and 19B arose before 20B and 20C in literature. Second, in the phylogenetic tree produced from random sampling, most sequences were obtained

from the later stages during the span of February 1st, 2020 to October 31st, 2020. On the other hand, the sequences in the tree produced from weighted sampling, show a greater spread in terms of the date they were collected. This is significant once again as during the early stages of the pandemic, less sequencing was performed compared to the later stages of the pandemic in 2020. Thus, it appears that the weighted sampling strategy minimized the amount temporal based sampling bias in the sequences. Finally, in the phylogenetic tree produced from weighted sampling, Nextstrain clades such as 20H which emerged from South Africa during the end of 2020, was only observed in the phylogenetic tree produced from weighted sampling. This shows that the weighted sampling strategy was able to extract sequences found in low quantities in GISAID. However, it is important note that clade 20G was only observed in the tree produced from random sampling and not in the tree produced from weighted sampling. Fortunately, this only occurred because strain 20G emerged in the USA, and coincidentally the subsample obtained from weighted sampling did not contain any sequences from the USA. Ultimately, the absence of clade 20G shows that in practice the samples obtained via weighted sampling can still deviate from the ideal sequencing rates because of chance. However, nonetheless, the subsample obtained from the weighted sampling strategy still resembles the IHME case proportions in the world more than the subsample obtained from random sampling (Figure 8). Future replicates using weighted sampling will likely produce phylogenetic trees that do contain sequences in clade 20G.

Ultimately, the effectiveness of the weighted sampling strategy in reducing sampling bias due to time and location was apparent from our findings. However, future experiments should produce more replicate and perform further downstream verification to confirm the advantages of the weighted sampling strategy over random sampling. All in all, we believe our sampling strategy can aid in the production of more reliable phylogenetic trees and transmission networks, and

correspondingly help aid in the containment of SARS-CoV-2 spread. Even with the rollout of vaccines for SARS-CoV-2, we believe that our strategy will remain beneficial as it can be applied to other possible future pandemics and epidemics.

Acknowledgements:

Supervisors: Dr. Paul Gordon, Dr. Quan Long, Dr. Michael Hynes

Lab members: Deshan Perera & the Long Lab

Special thanks to GISAID, Compute Canada, and Researching Computing Services at the University of Calgary

Literature Cited:

1. Benvenuto, D. *et al.* The global spread of 2019-nCoV: a molecular evolutionary analysis. *Pathog Glob Health* 1–4 (2020) doi:10.1080/20477724.2020.1725339.

2. Zhao, S. *et al.* Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases* **92**, 214–217 (2020).

3. Coronavirus Update (Live): 39,945,224 Cases and 1,114,578 Deaths from COVID-19 Virus Pandemic - Worldometer. https://www.worldometers.info/coronavirus/.

4. Rahmandad, H., Lim, T. Y. & Sterman, J. *Estimating COVID-19 Under-Reporting Across 86 Nations: Implications for Projections and Control*. https://papers.ssrn.com/abstract=3635047 (2020).

5. Tan, J. *et al.* Transmission and clinical characteristics of asymptomatic patients with SARS-CoV-2 infection. *Future Virol* doi:10.2217/fvl-2020-0087.
6. Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci USA* **117**, 9241 (2020).
7. Nextclade. https://clades.nextstrain.org.
8. GISAID - Initiative. https://www.gisaid.org/.
9. Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T. & Jacobsen, K. H. Complexity of the Basic Reproduction Number (R0). *Emerg Infect Dis* **25**, 1–4 (2019).
10. Giattino, C. How epidemiological models of COVID-19 help us estimate the true number of infections. (2020). https://ourworldindata.org/covid-models#institute-for-health-metrics-and-evaluation-ihme
11. Roser, M., Ritchie, H., Ortiz-Ospina, E., & Hasell, J. Coronavirus Pandemic (COVID-19). *(2020) https://ourworldindata.org/coronavirus*
12. Arora, R. K. *et al.* SeroTracker: a global SARS-CoV-2 seroprevalence dashboard. *The Lancet Infectious Diseases* **21**, e75–e76 (2021).
13. Huddleston, J. *et al.* Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. *Journal of Open Source Software* **6**, 2906 (2021).
14. Gordon, P. Nybbler. (2020) https://github.com/nodrogluap/nybbler
15. theLongLab. TransCOVID. (2021) https://github.com/theLongLab/TransCOVID

16. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* **20**, 1160–1166 (2019).
17. Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., et al. (2019) BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS computational biology, 15(4), e1006650.
18. FigTree. http://tree.bio.ed.ac.uk/software/figtree/.
19. Bevand, M. COVID-19 Age Stratified IFR. (2021) https://github.com/mbevand/covid19-age-stratified-ifr
20. Loannidis, J. Infection fatality rate of COVID-19. *Bullentin of the World Health Organization*, ID: BLT.20.265892 (2020).