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

David H Yang
Bioinformatics RIP Seminar (May)
Dr. Paul Gordon & Dr. Quan Long & Dr. Michael Hynes

May 19th, 2021

UNIVERSITY OF CALGARY

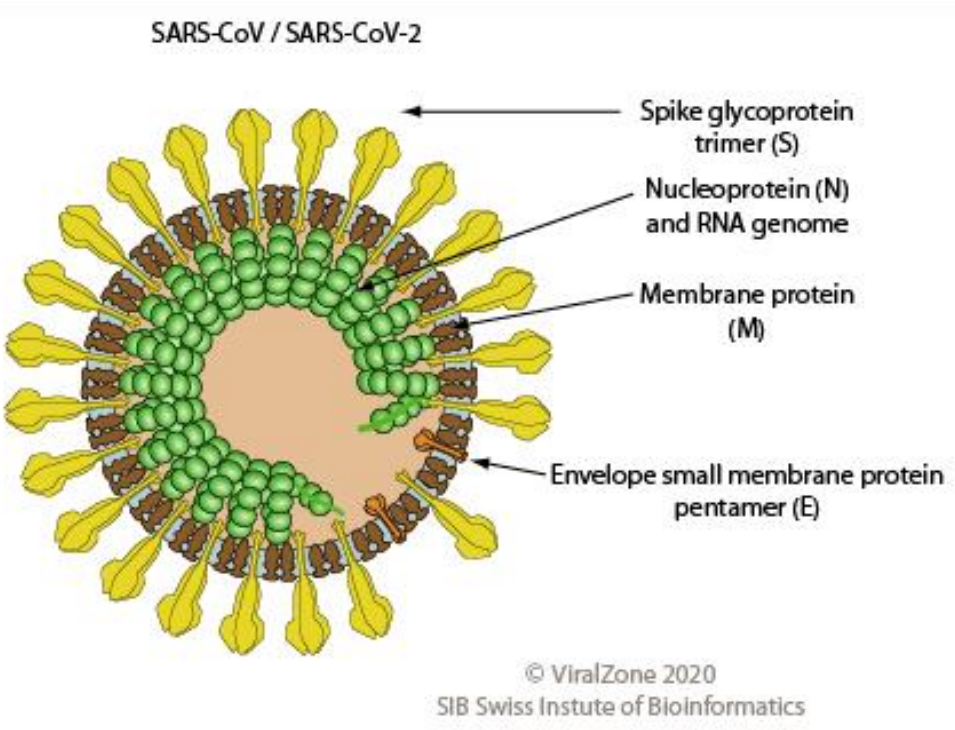
SARS-CoV-2

Coronavirus Cases:

164,669,268 X 10.5

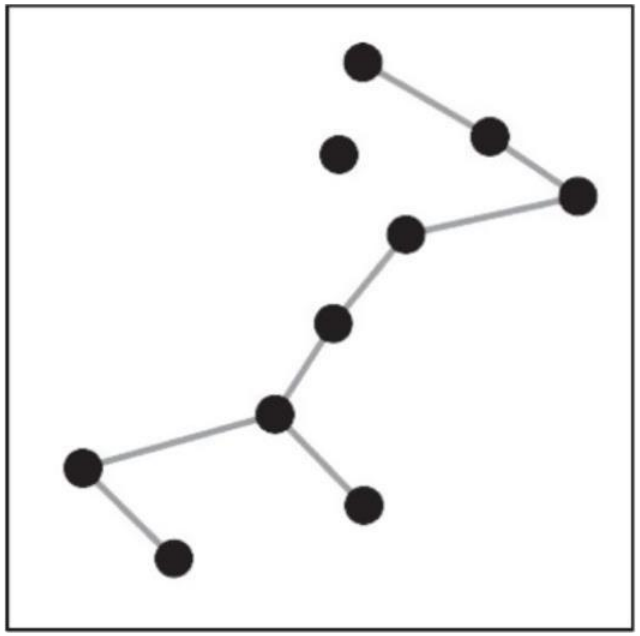
Deaths:

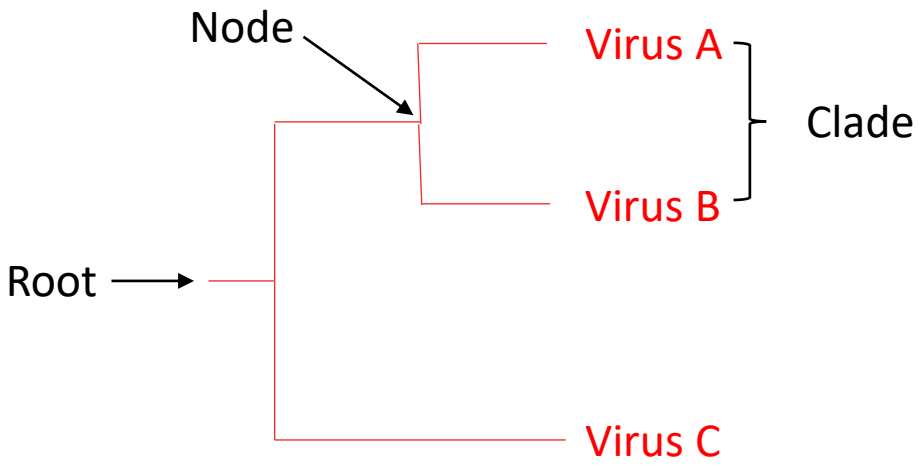
3,412,136 X 1.47



UNIVERSITY OF CALGARY

Phylogenetic Analysis





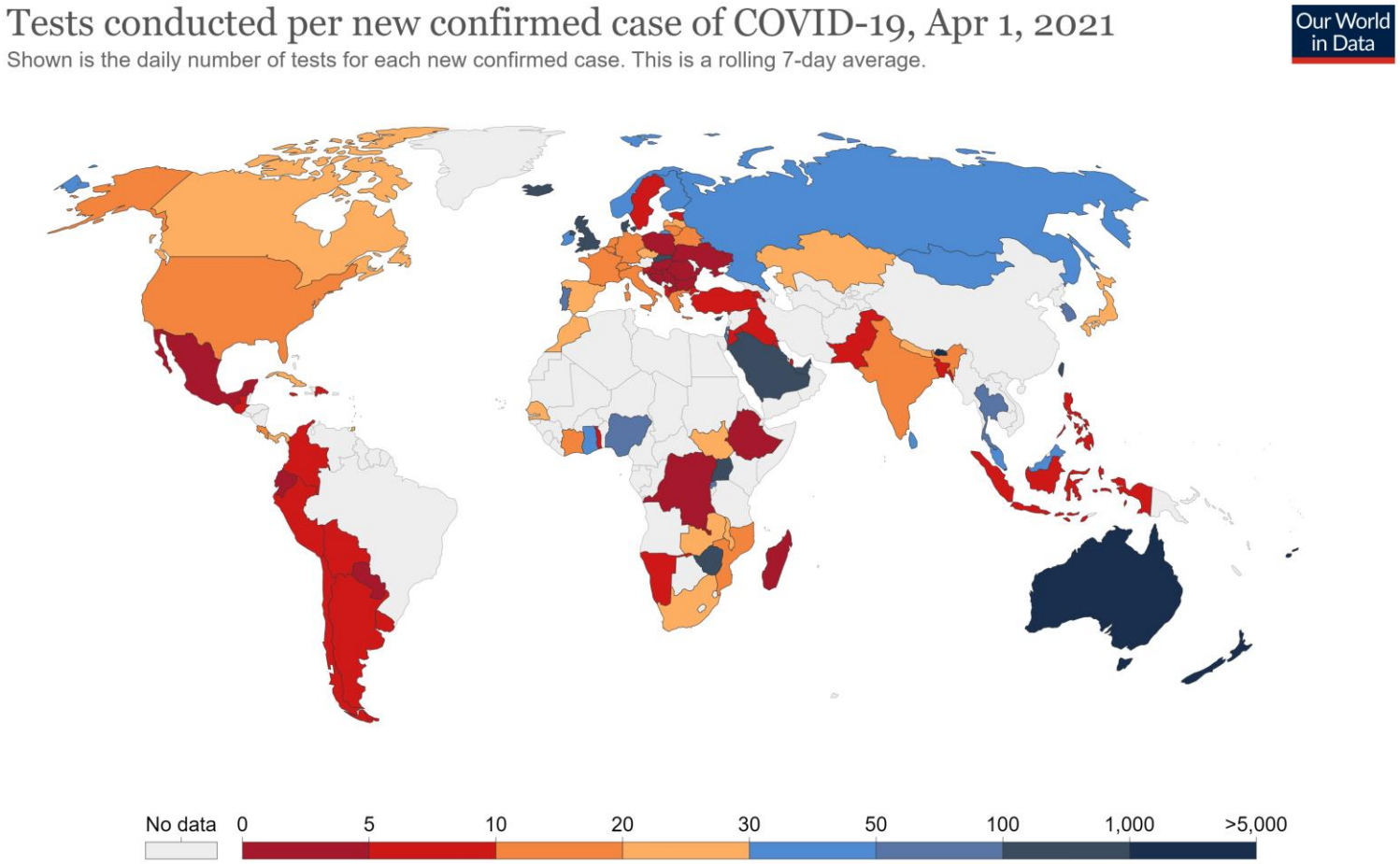
UNIVERSITY OF CALGARY

Source of Sampling Bias

Table 1: The total number of cases, cases per 1 million population, tests per 1 million population and genome sequences from Canada, China, India, USA, and the United Kingdoms. All statistics were obtained from Worldometer and GISAID on October 7th, 2020

Country:	Cases	Cases/1 million population	Tests/1 million population	Genome sequences on GISAID
Canada	196,321	5,188	226,302	2,714
China	85,659	60	111,163	994
India	7,492,727	5,414	67,381	3,092
USA	8,334,763	25,137	376,047	34,577
United Kingdom	705,428	10,375	426,333	61,441

Difficulties Controlling Sampling Bias



UNIVERSITY OF CALGARY

SARS-CoV-2 Prevalence Metrics

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

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

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

Serology test: Detection of previous infections via presence of anti-SARS-CoV-2 antibodies

Aims

1) Investigate the **true prevalence** of SARS-CoV-2 in each **region** around the world and the region's corresponding **sequencing contribution** to **public datasets**
2) Devise a **weighted sampling strategy** to create sequence subsamples that are **representative** of SARS-CoV-2 **prevalence** in **regions** around the world and different **months**
3) Generate and compare the accuracy of **phylogenetic trees** produced through **weighted sampling** and **random sampling**

Aim 1

Investigate the **true prevalence** of SARS-CoV-2 in each **region** around the world and the region's corresponding **sequencing contribution** to **public datasets**

Methodology

Data Collection

Obtain and calculate seroprevalence metrics

Data Manipulation

Group statistics by month and region

Data Analysis

Compare prevalence estimates from each metric

Our World in Data

GISAID

SeroTracker

UNIVERSITY OF CALGARY

Prevalence

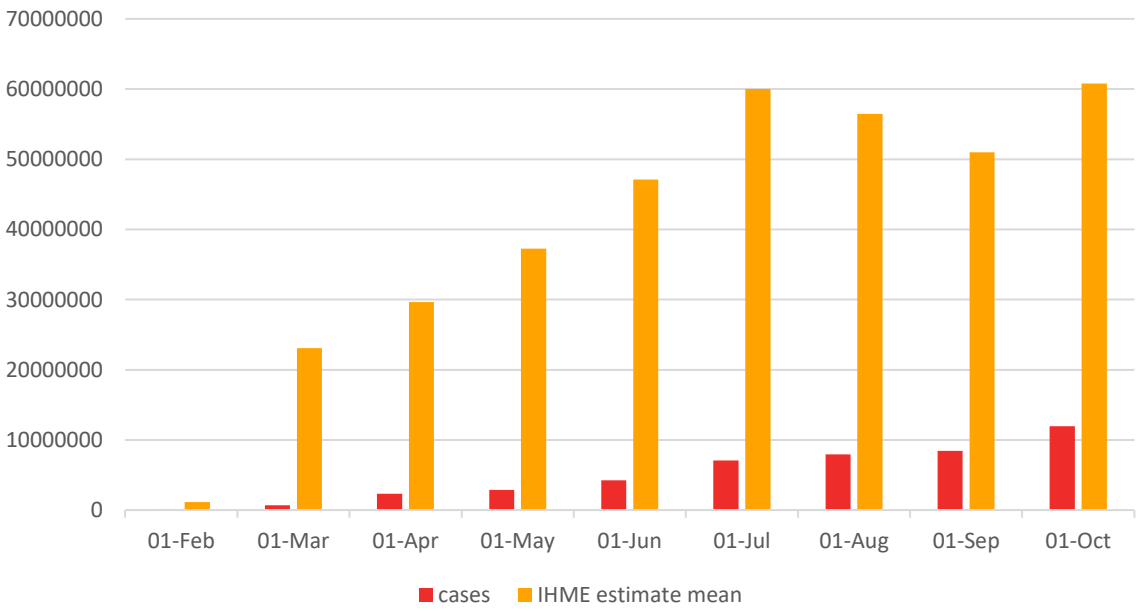


Figure 1: Number of cases and the Institute for Health Metrics and Evaluation's mean estimate of cases in the world from February 1st, 2020 to October 31st, 2020.

UNIVERSITY OF CALGARY

Seroprevalence

Table 2: SARS-CoV-2 prevalence metrics obtained for the five countries and months with the highest seroprevalence findings.

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

Sequencing contribution: march

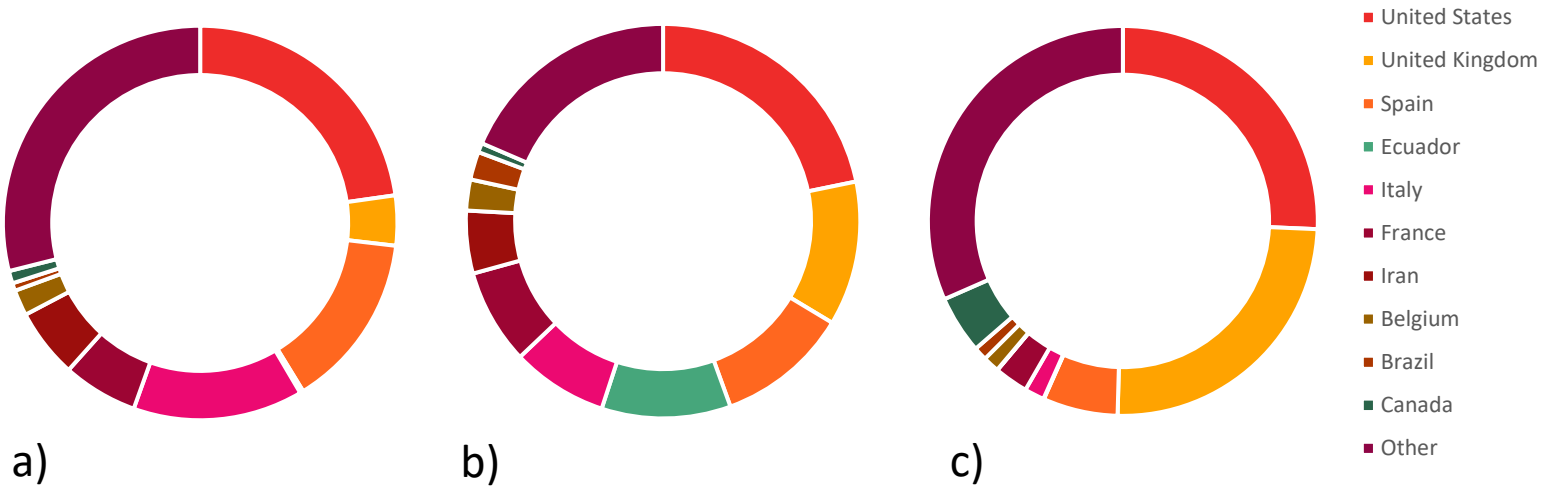


Figure 2: Doughnut charts for countries in March 2020. a) depicts the number of reported cases in each country, b) depicts the IHME estimate of the number of cases in each country, and c) depicts the number of sequences sequenced in each country.

Sequencing contribution: globally

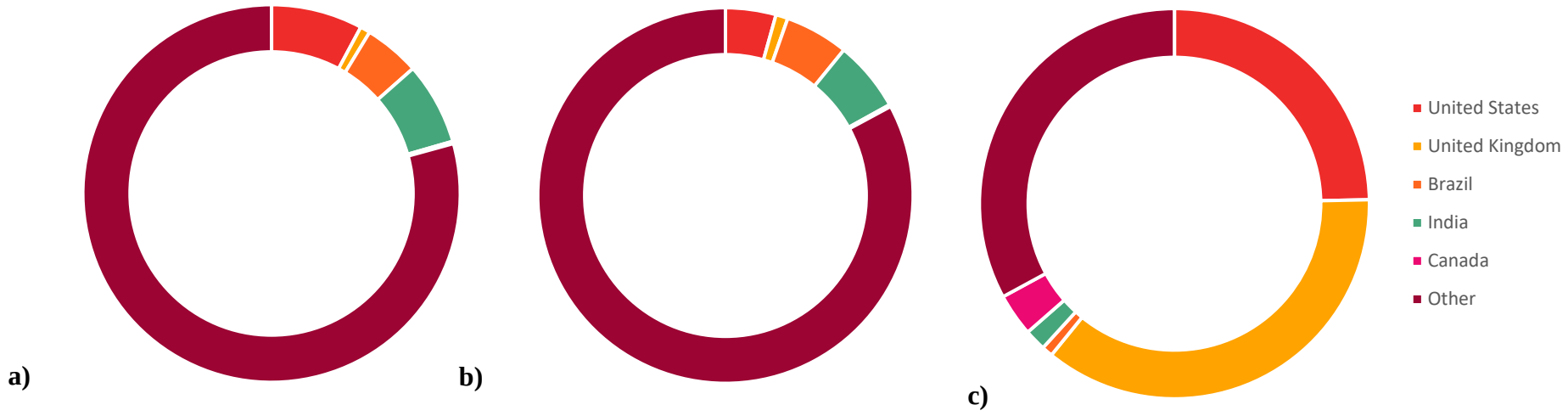


Figure 2: Doughnut charts for countries between February 1st 2020 to October 31st, 2020. a) depicts the number of reported cases in each country, b) depicts the IHME estimate of the number of cases in each country, and c) depicts the number of sequences sequenced in each country.

UNIVERSITY OF CALGARY

Aim 2

Devise a **weighted sampling strategy** to create sequence subsamples that are **representative** of SARS-CoV-2 **prevalence** in **regions** around the world and different **months**

Methodology

Random sampling

- Randomly select *N* sequences from all sequence data available

Weighted sampling

- Select *N* sequences from each country in each month based on SARS-CoV-2 prevalence
- Prevalence estimated with IHME mean estimates

Problems with using serology tests

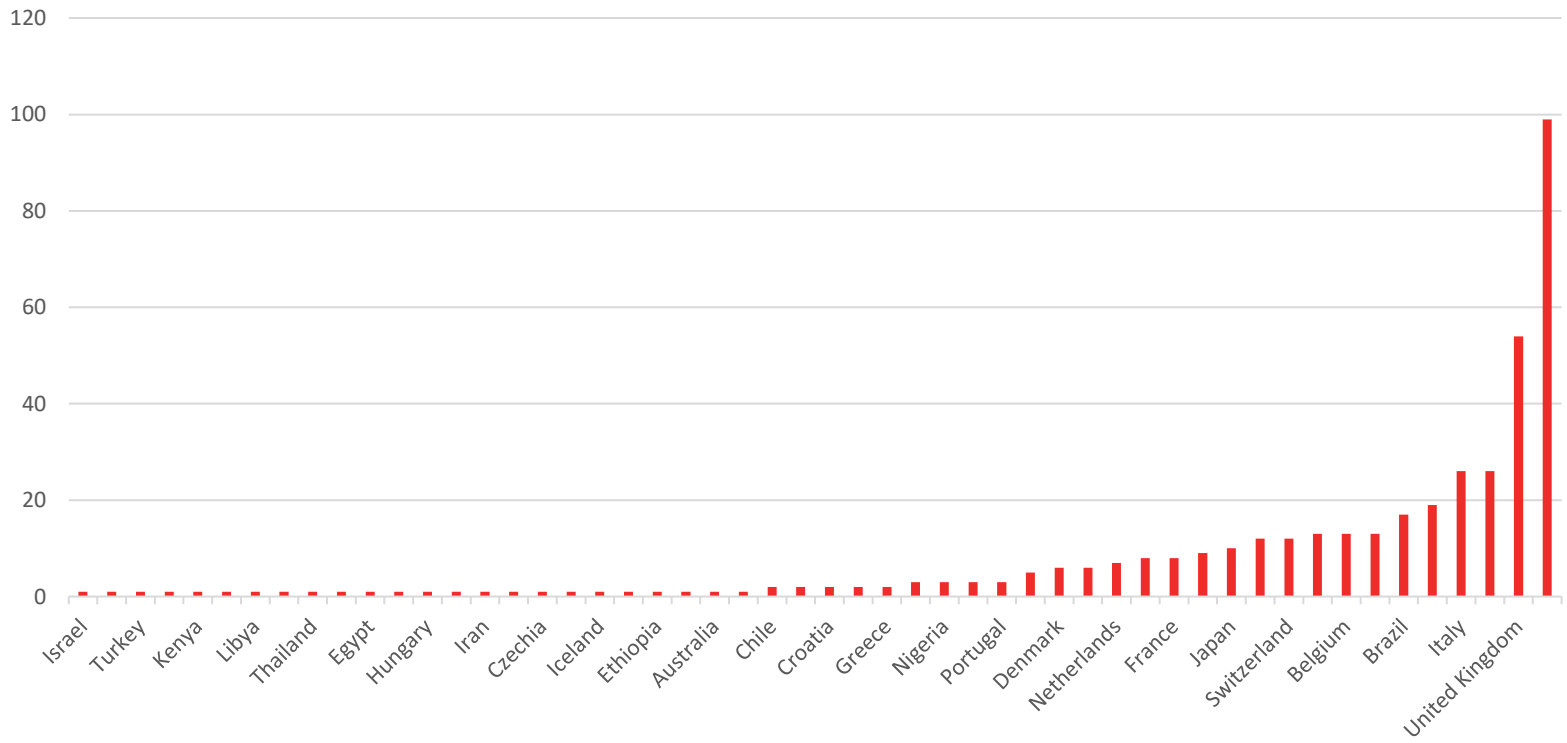
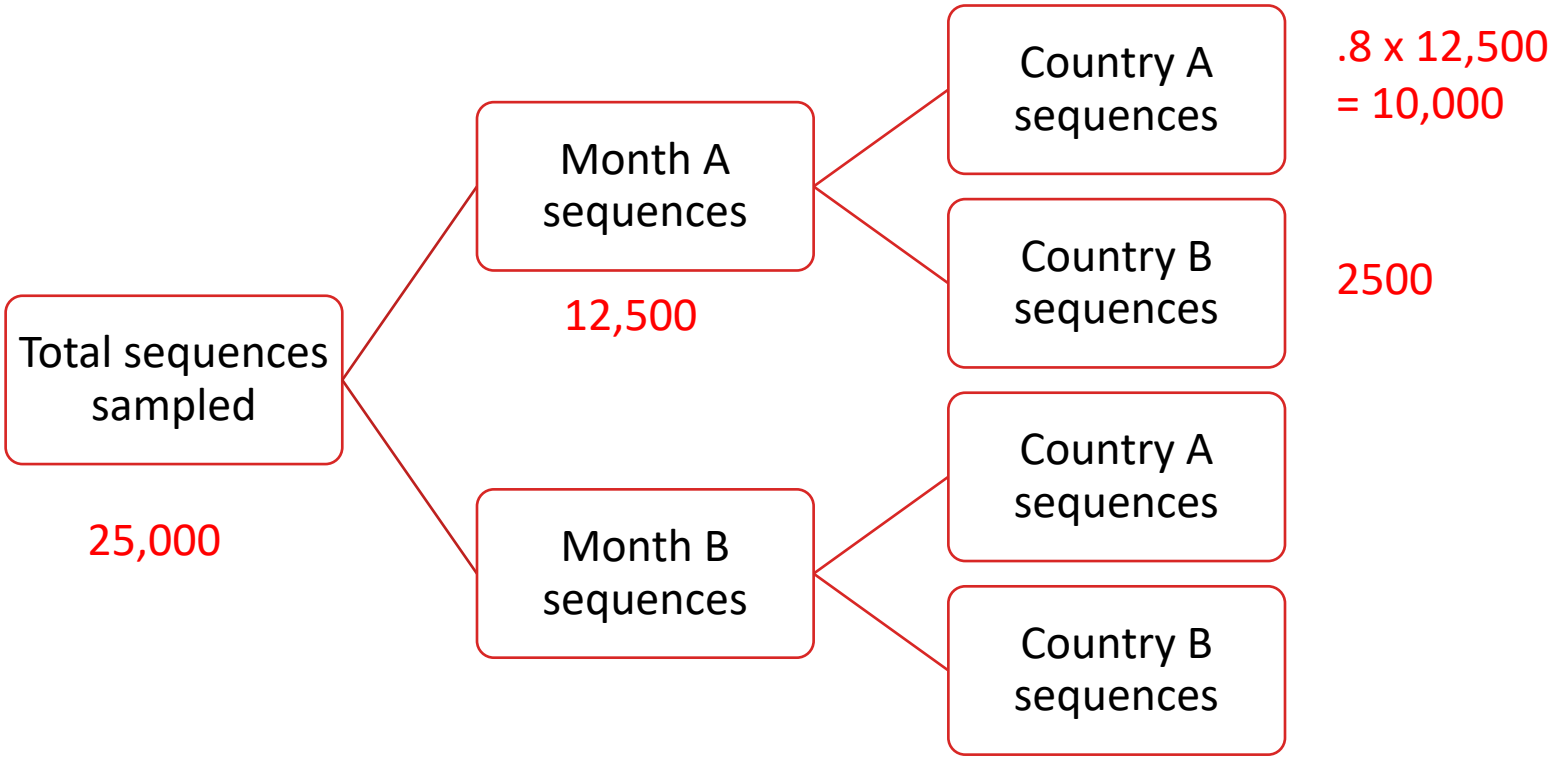


Figure 3: A tally of the number of seroprevalence studies across the world collected on SeroTracker on December 9th, 2020. The last bar and third last bar from the right on the chart represents the total number of studies from the USA and China respectively

UNIVERSITY OF CALGARY

Weighted sampling strategy flowchart



UNIVERSITY OF CALGARY

Weighted sampling versus random sampling

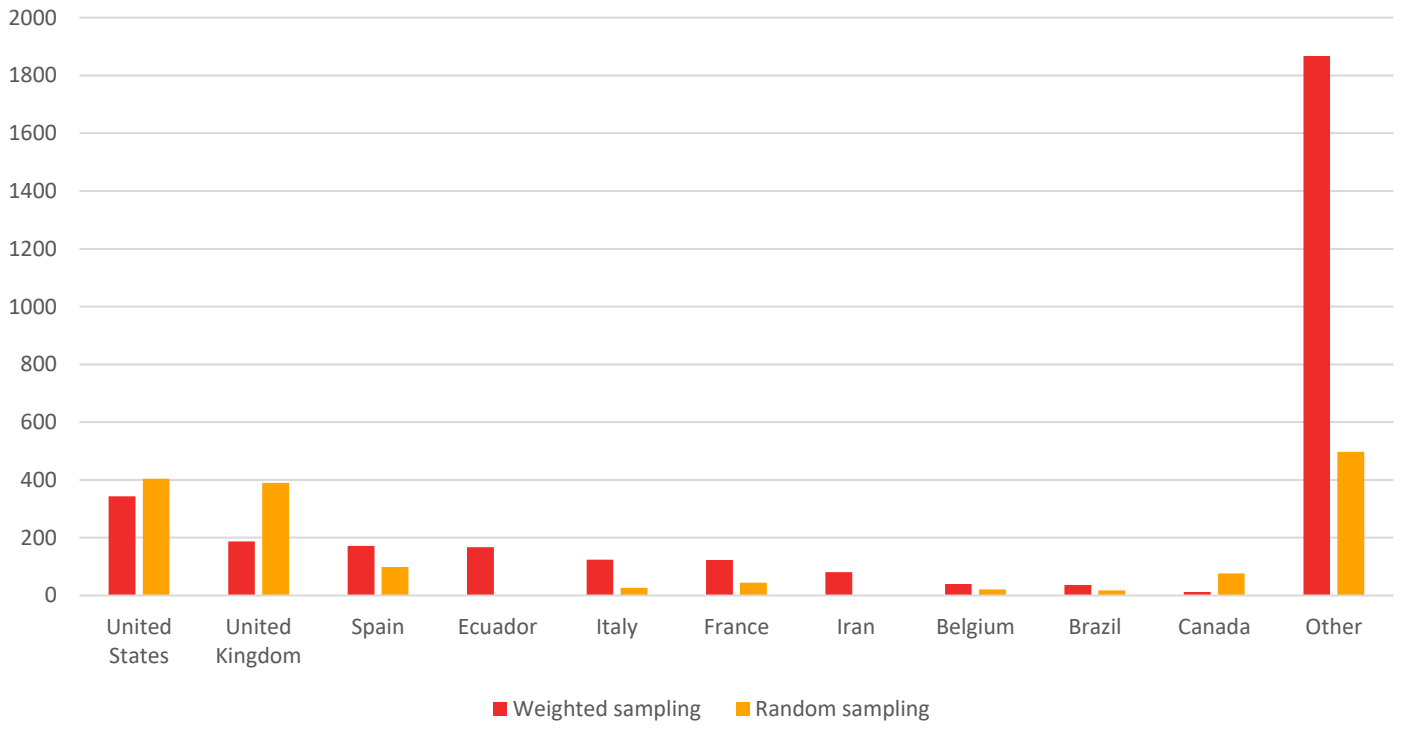


Figure 4: The number of sequences to be obtained from each country if 1576 sequences from March 2020 were to be obtained via weighted or random sampling.

UNIVERSITY OF CALGARY

Subsamples created

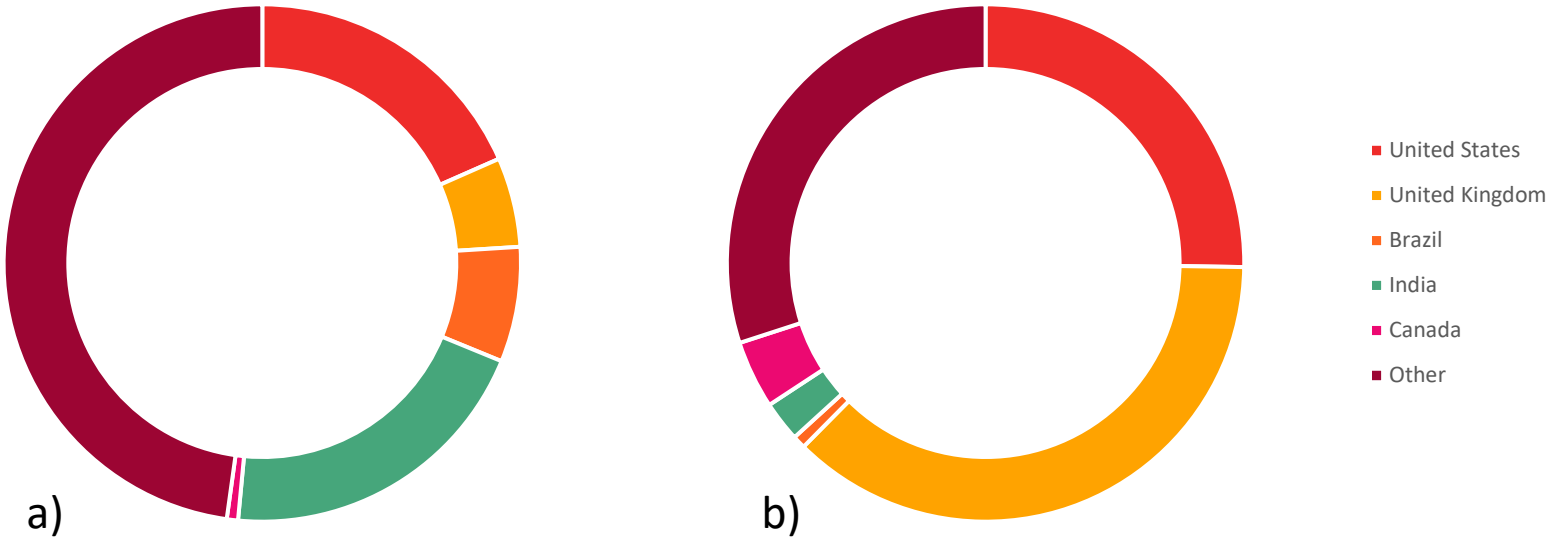


Figure 5: Doughnut charts for the sequencing location of subsamples obtained via a) our weighted sampling strategy versus b) the random sampling strategy.

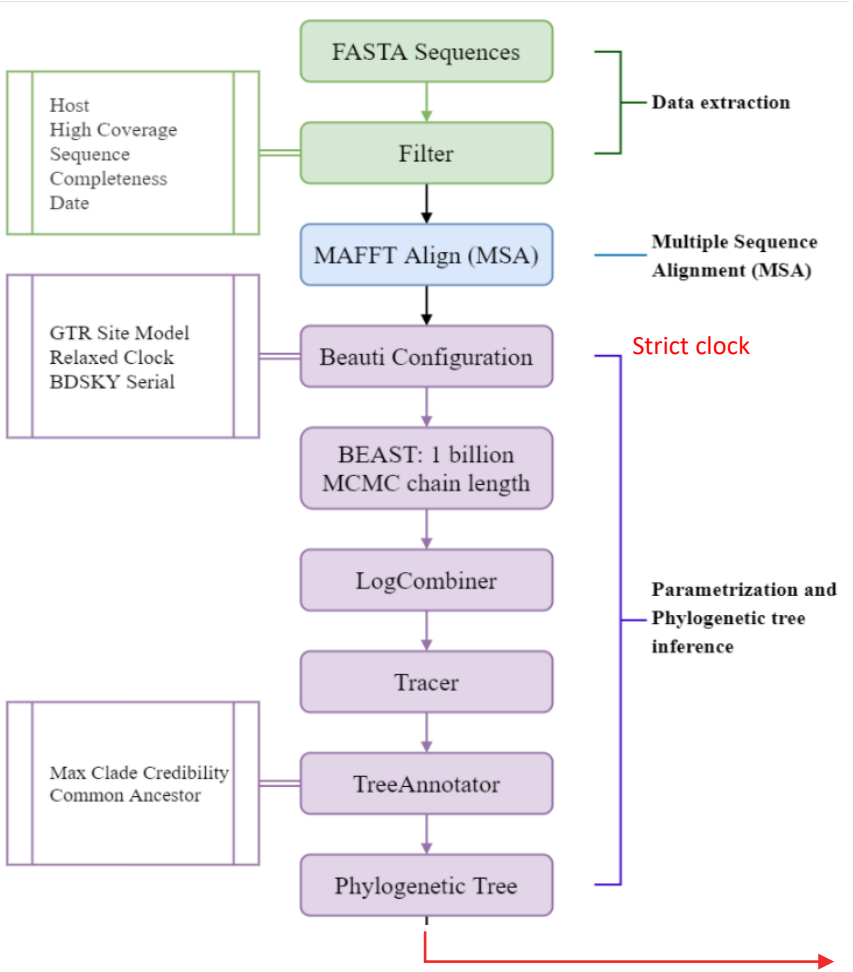
Aim 3

Generate and compare the accuracy of **phylogenetic trees** produced through **weighted sampling** and **random sampling**

Methodology

Samples from weighted and random sampling

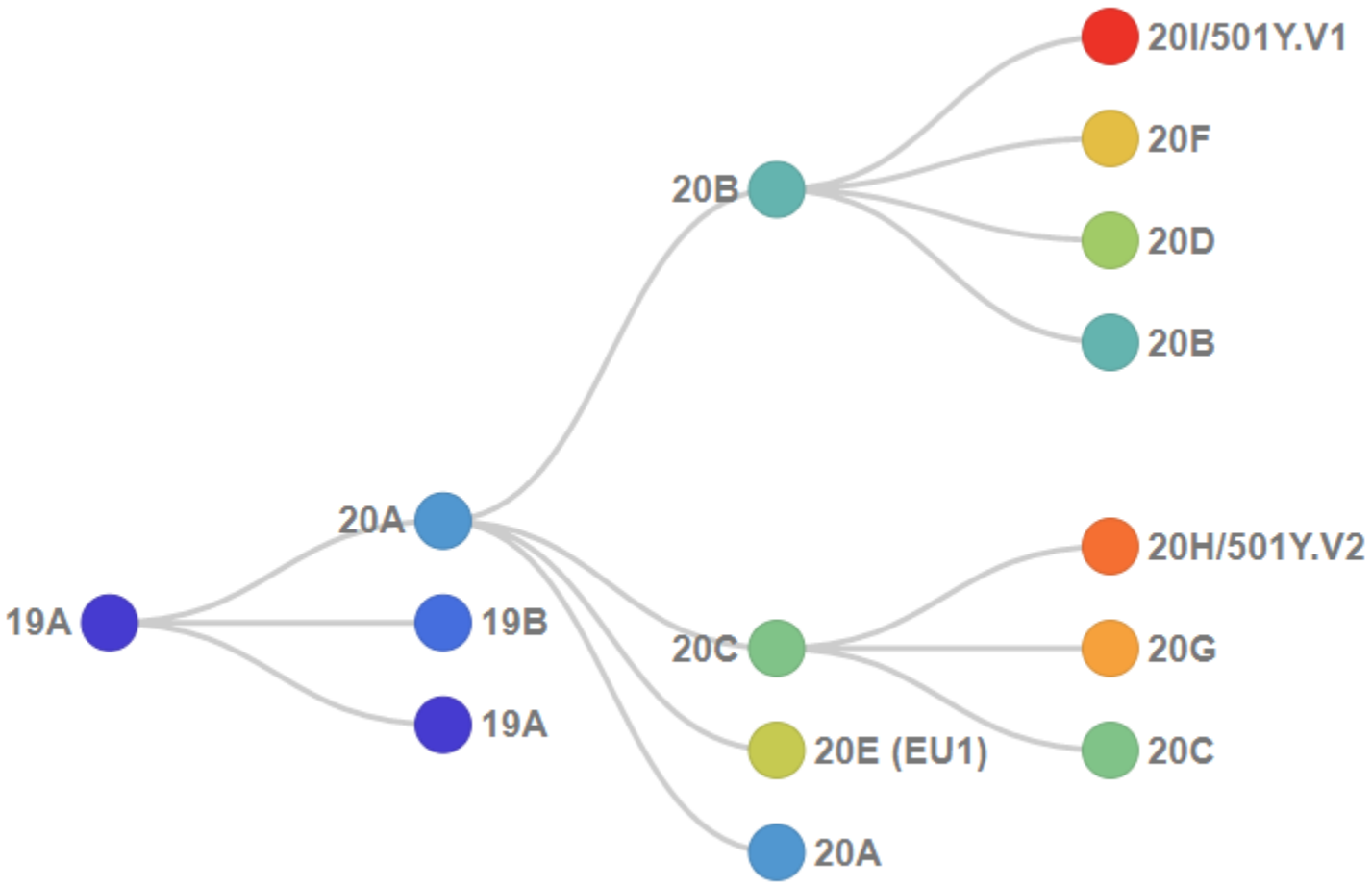
TransCOVID pipeline (Long Lab)



Compare to Nextstrain global analysis & dedicated builds

UNIVERSITY OF CALGARY

Nextstrain Clades



(Nextstrain)

UNIVERSITY OF CALGARY

Random sampling

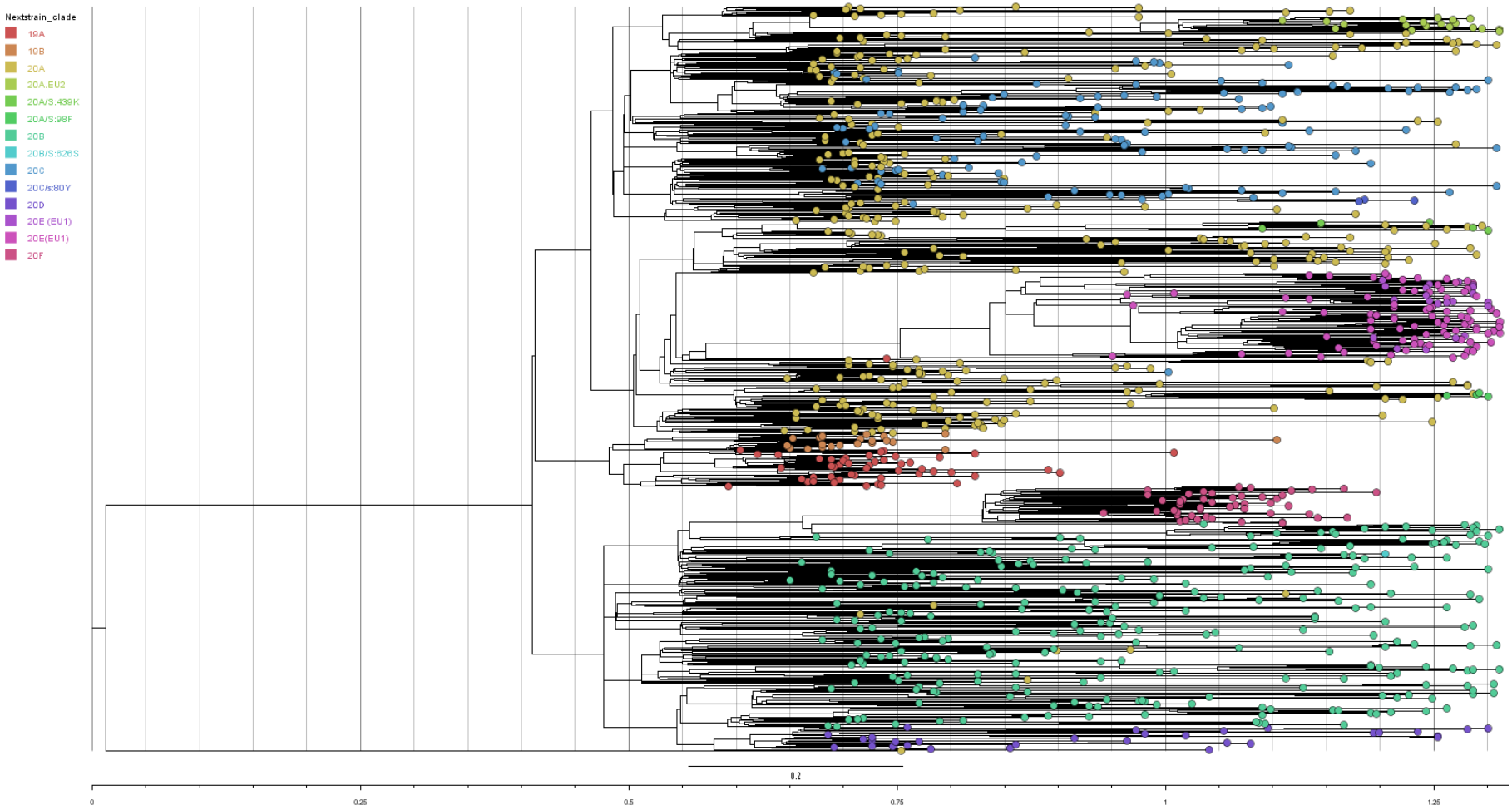


Figure 6: General clade (NextStrain) structure of phylogenetic tree produced from random sampling using BEAST2.

UNIVERSITY OF CALGARY

Weighted sampling

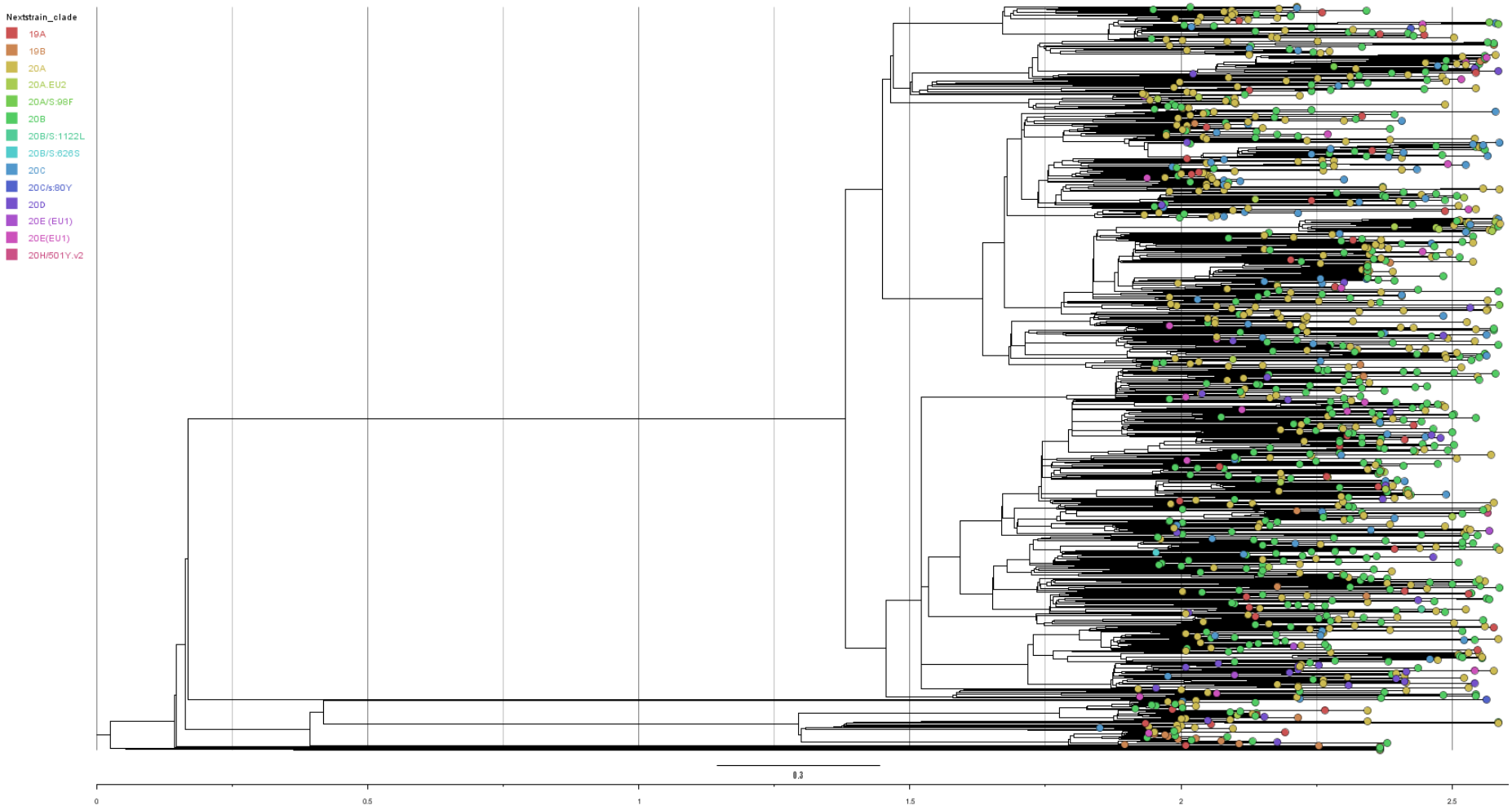


Figure 7: General clade (NextStrain) structure of phylogenetic tree produced from weighted sampling using BEAST2.

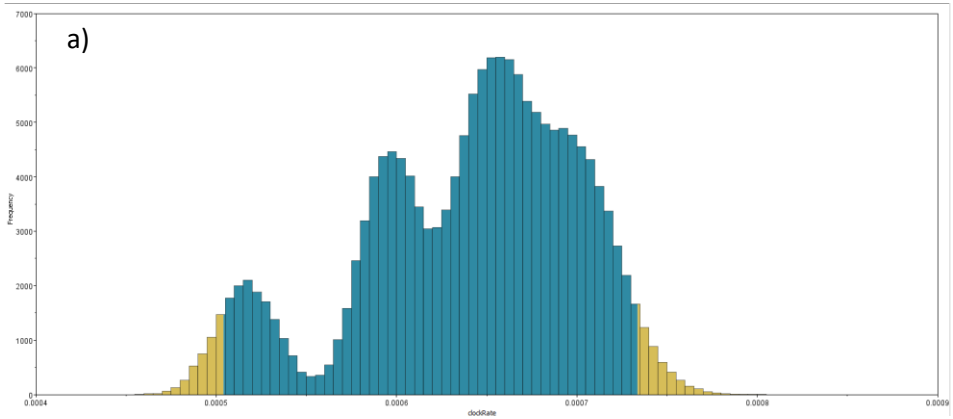
UNIVERSITY OF CALGARY

Comparison of Trees with Nextstrain

Table 3: Earliest dates sampled for various Nextstrain clades in each subsample. The global analysis and the dedicated builds were performed by Nextstrain.

Clade	Random	Weighted	Global Analysis	Dedicated Build
20A.EU2	2020-08-17	2020-08-20	2020-09-01	2020-06-10 (02-04 to 05-01)
20A/S:439K	2020-08-10	N/A	2020-08-24	2020-04-03 (03-14 to 04-13)
20A/S:98F	2020-10-12	2020-09-08	2020-08-28	2020-03-10 (01-23 to 02-19)
20B/S:1222L	N/A	2020-10-07	2020-09-07	2020-07-06 (04-15 to 06-29)
20B/S:626S	2020-09-21	2020-10-06	2020-12-22	2020-07-15 (05-31 to 07-10)
20C/S:80Y	2020-09-12	2020-06-29	2020-09-21	2020-07-16 (04-09 to 04-20)
20E (EU1)	2020-06-20	2020-07-10	2020-06-25	2020-04-30 (03-04 to 04-24)
20H/501Y.v2	N/A	2020-10-19	2020-11-17	2020-10-08 (07-02 to 09-20)

Phylogenetic Analysis: Clock rate



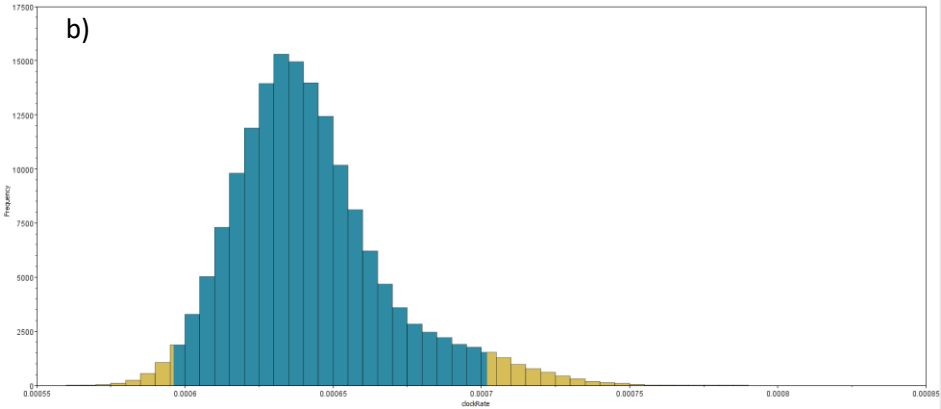
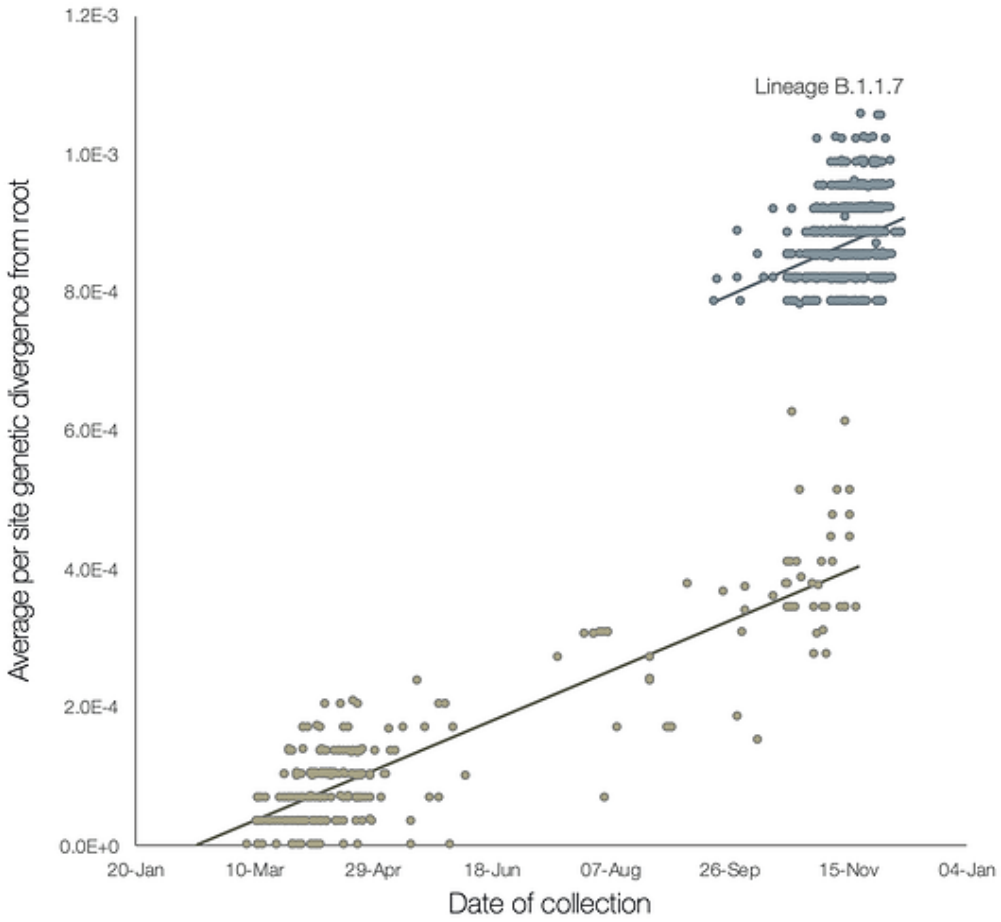


Figure 8: Clock rate trace files obtained from the analysis of the a) weighted sampled subsample and b) random sampled subsample.

UNIVERSITY OF CALGARY

Differences in Clock Rate



(Rambaut et al., 2020)

UNIVERSITY OF CALGARY

Acknowledgements

Supervisors

Dr. Paul Gordon

Dr. Quan Long

Dr. Michael Hynes

Special thanks to

Deshan Perera

Long lab

UNIVERSITY OF CALGARY

Thank you for listening!