Detecting the emergence of SARS-COV-2 and its sensitivity to data sharing policies

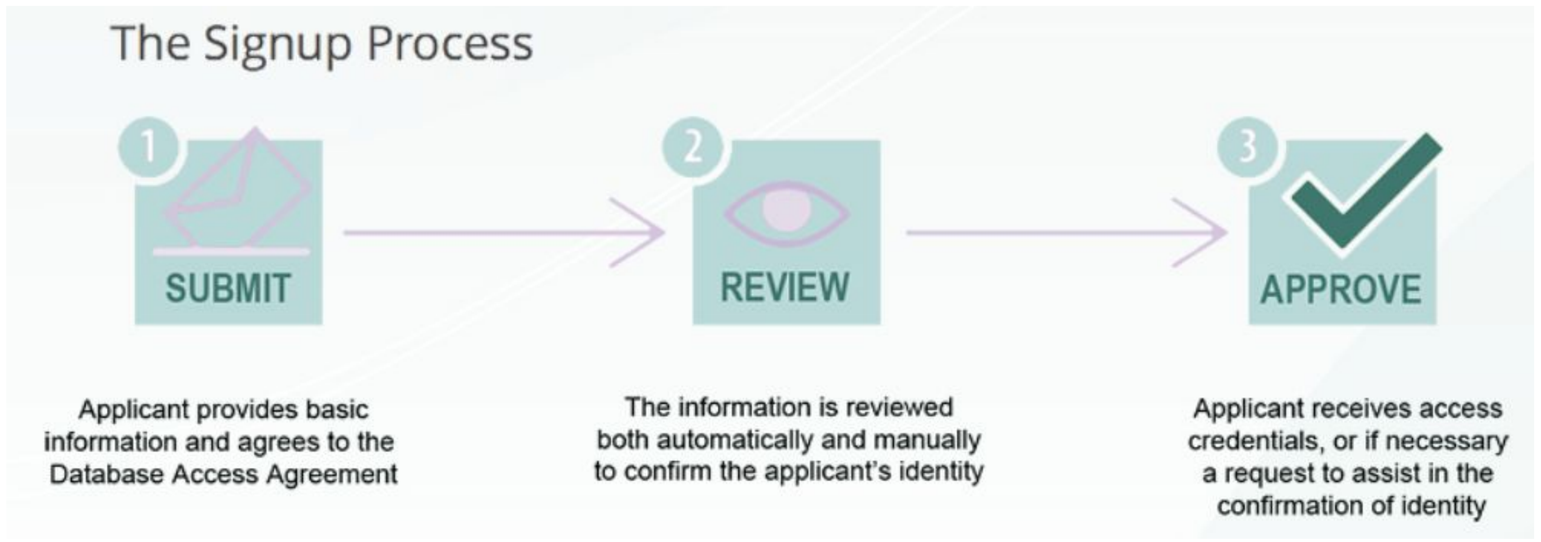
Western

Pathology and Laboratory Medicine

Background

GISAID

- World’s largest repository of SARS-CoV-2 sequences (> 10 million)

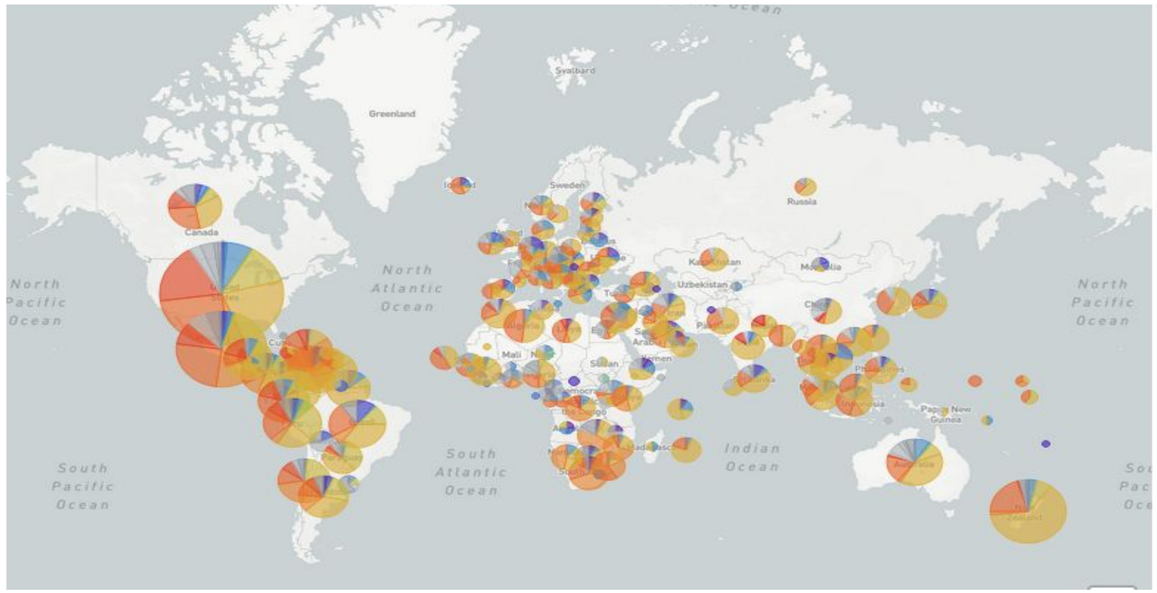


https://gisaid.org/register/

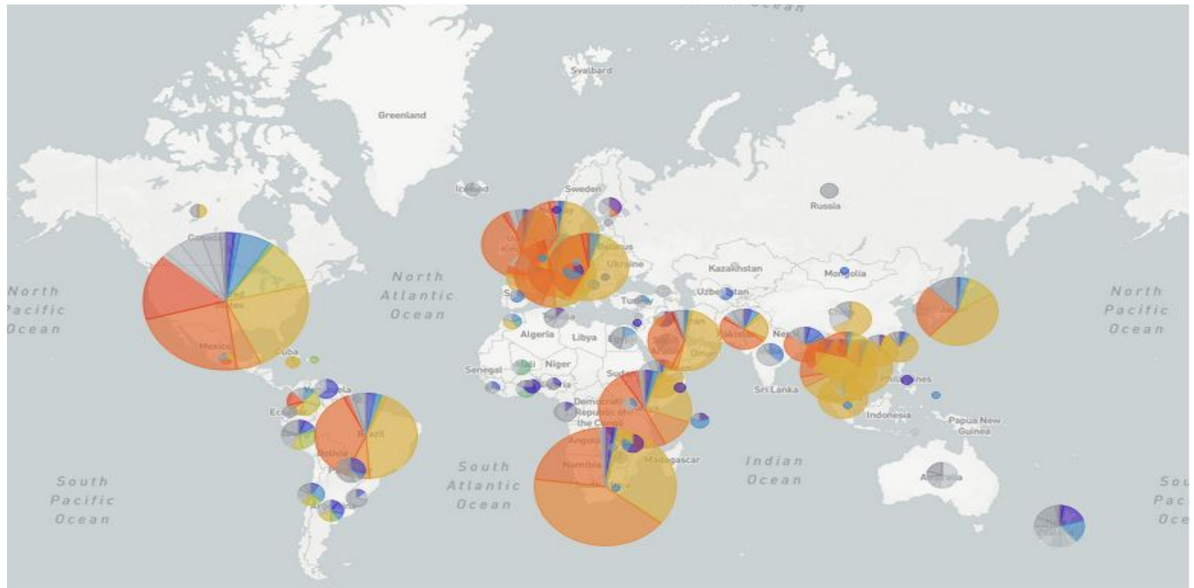
Nextstrain

- Open-source repository

GISAID



Nextstrain



https://nextstrain.org

Research Question

Would the identification of SARS-CoV-2 variants of concern (alpha, beta, gamma, delta, omicron) have been significantly delayed if the investigators were limited to open data feeds?

Methods

Publication delay

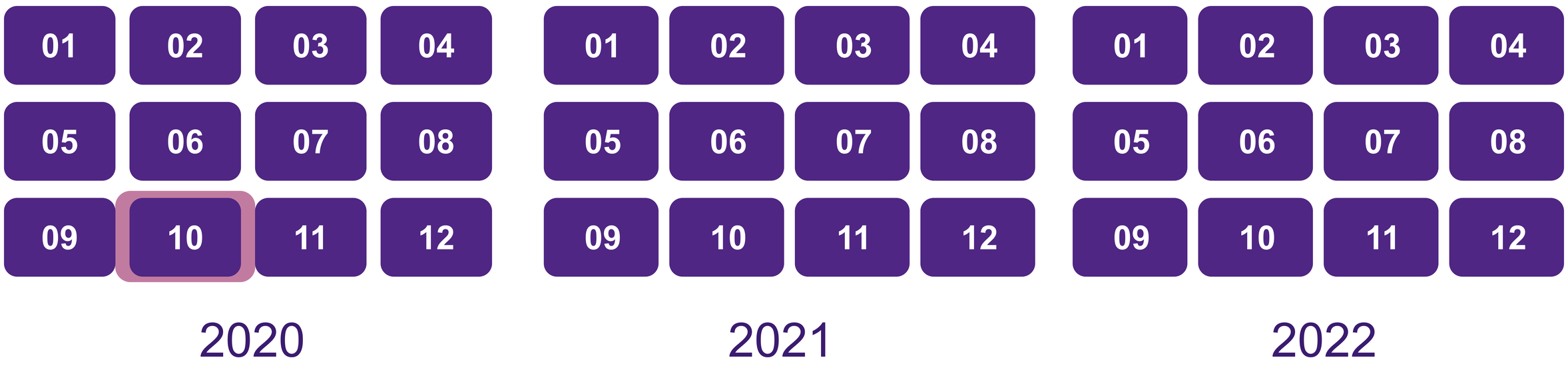
- Comparison of the time delay from genome collection to genome submission
- Random sample of n = 500

Lineage	Most common countries	Earliest date	# designated	# assigned	Description	WHO Name
B.1.1.7	United Kingdom 24.0%, United States of America 20.0%, Germany 9.0%, Sweden 6.0%, Denmark 6.0%	2020-09-01	69636	1101636	UK lineage of concern, associated with the N501Y mutation. More information can be found at cov-lineages.org/global_report.html	Alpha

https://cov-lineages.org/

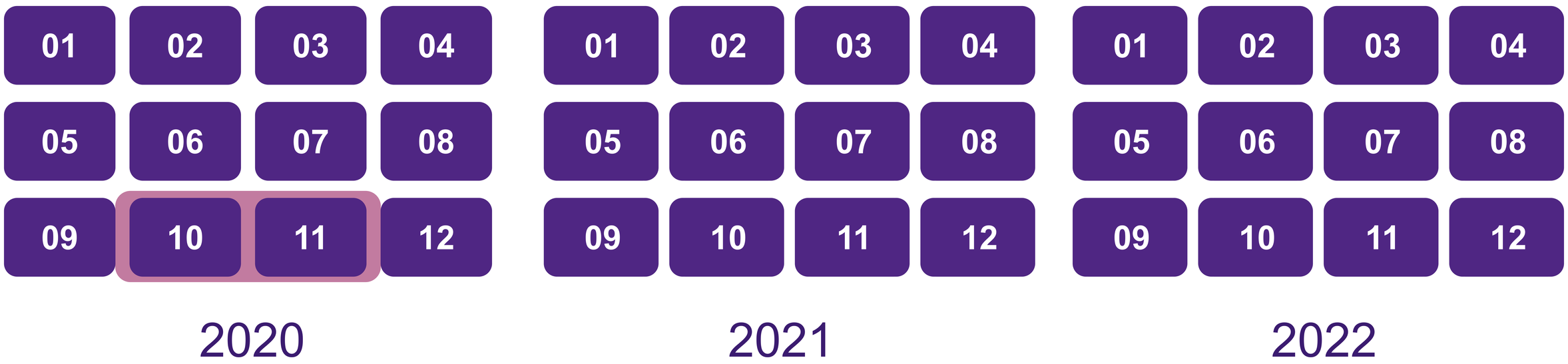
Detecting VoCs

- Monthly Submission Date cutoffs (October 2020 — May 2022)



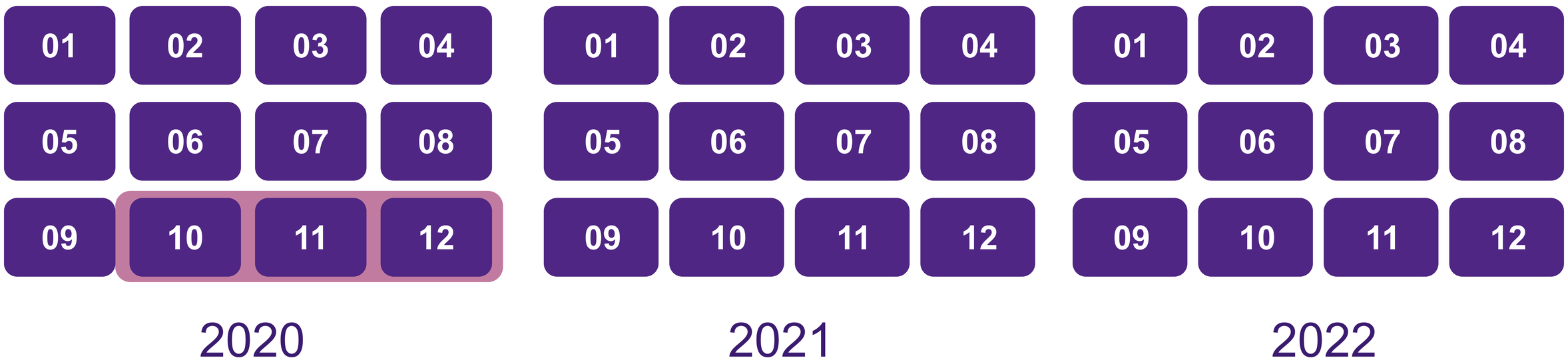
Detecting VoCs

- Monthly Submission Date cutoffs (October 2020 — May 2022)



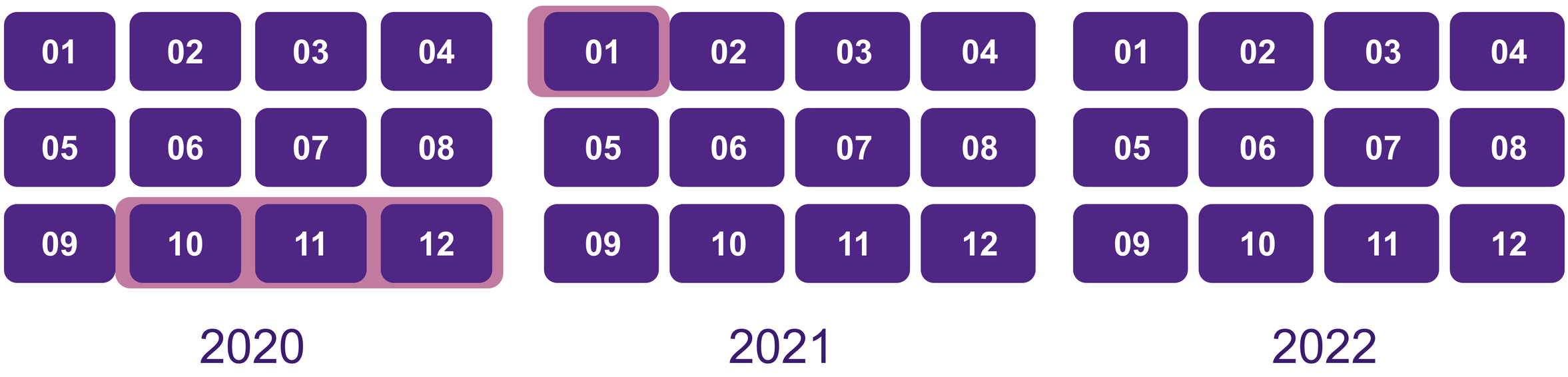
Detecting VoCs

- Monthly Submission Date cutoffs (October 2020 — May 2022)



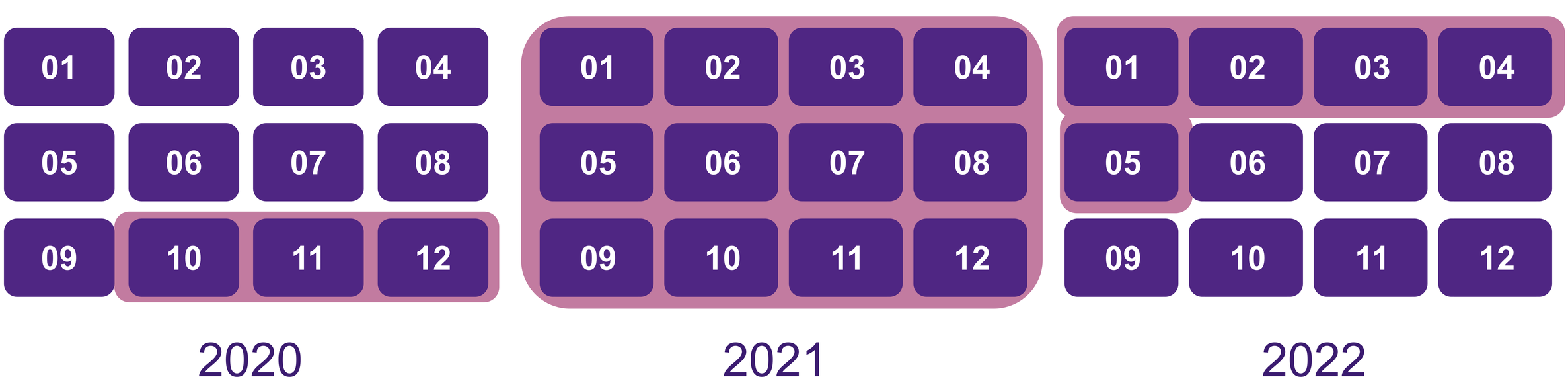
Detecting VoCs

- Monthly Submission Date cutoffs (October 2020 — May 2022)



Detecting VoCs

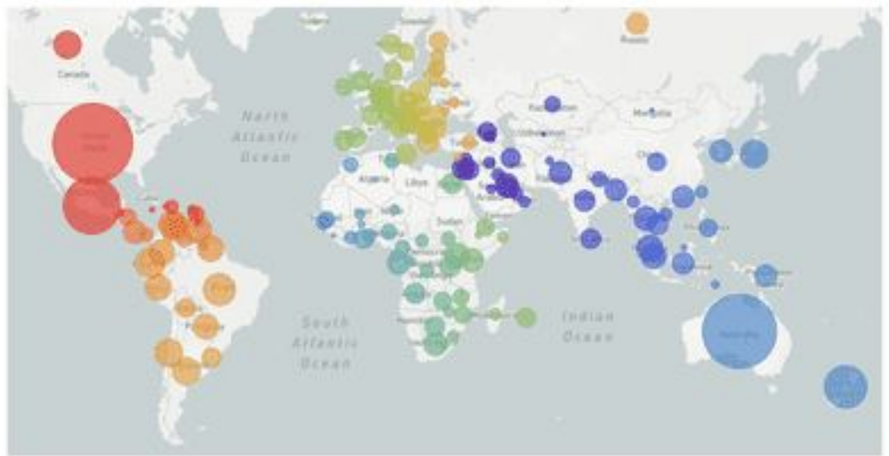
- Monthly Submission Date cutoffs (October 2020 — May 2022)



Geographic distribution

- Top 10 submitting countries

GISAID data



Open Data

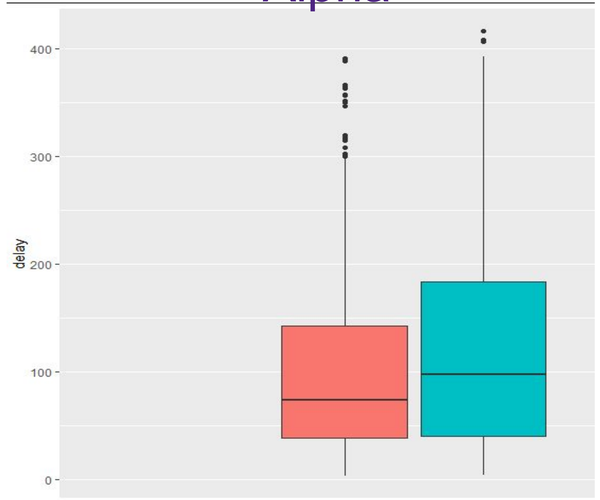
https://nextstrain.org

Results

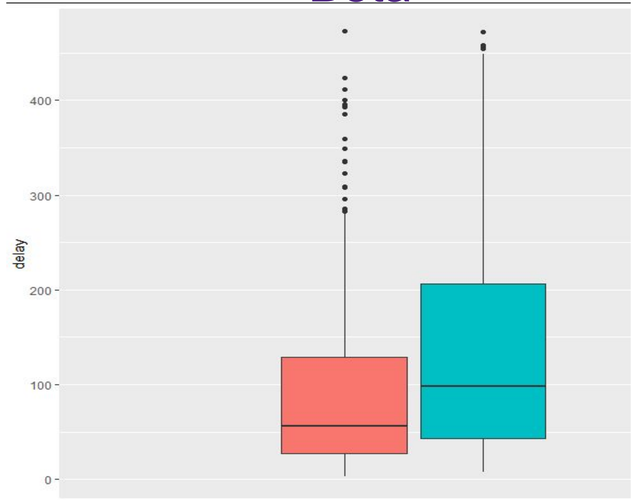
Publication Delay

Publication delay (days)

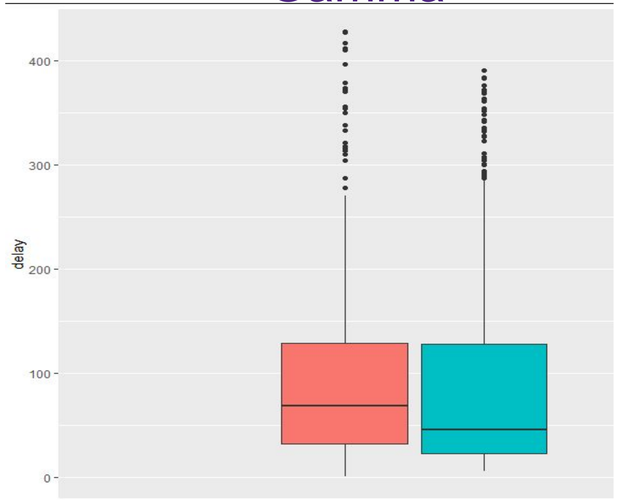
Alpha



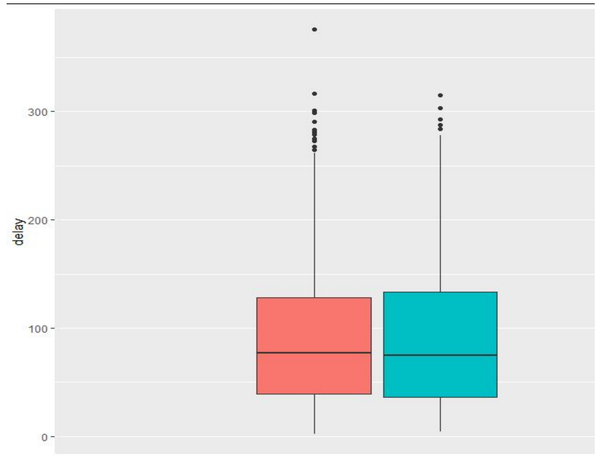
Beta



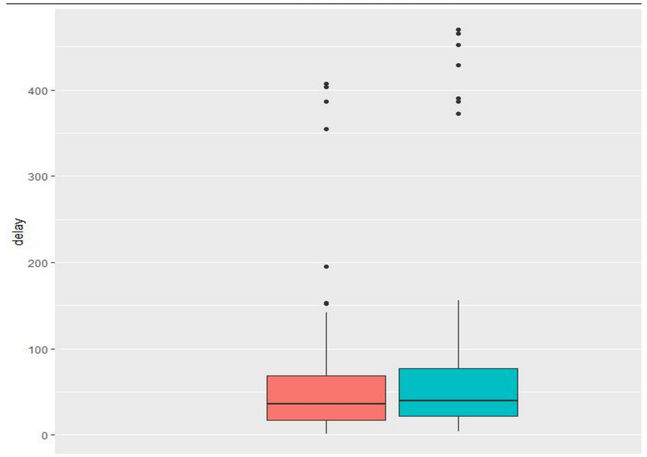
Gamma

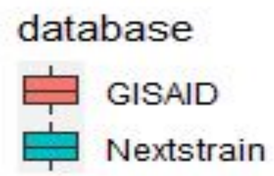


Delta



Omicron





Detecting VoCs

	GISAID	Nextstrain
Alpha	12-2020	02-2021
Beta	12-2020	03-2021
Gamma	02-2021	03-2021
Delta	04-2021	10-2021
Omicron	12-2021	03-2022

Earlier Submission Cut-Off date

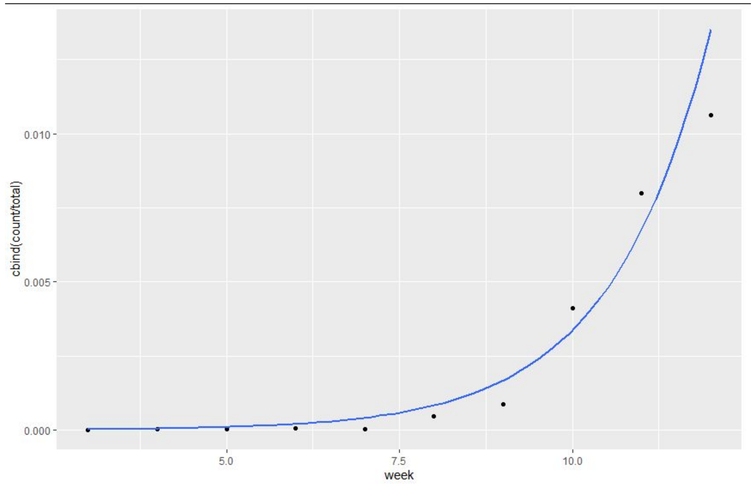
Detecting VoCs

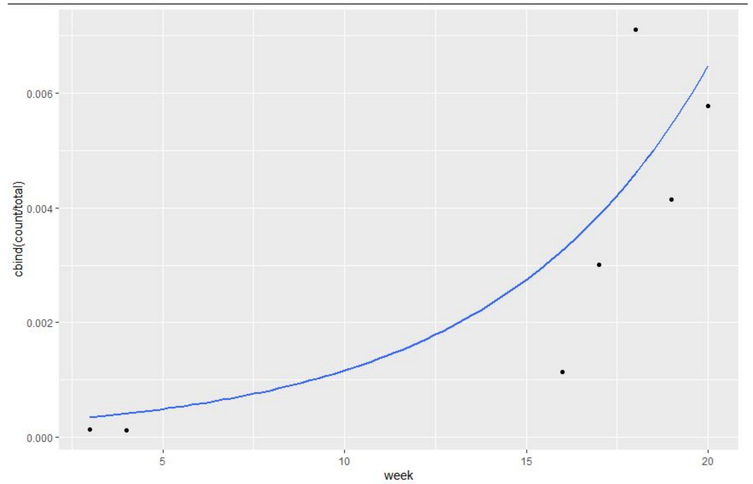
Relative Frequency of VoCs/week

GISAID

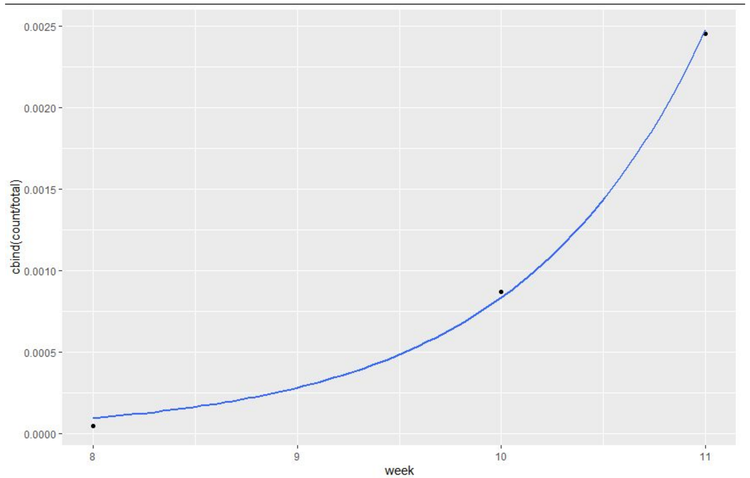
Nextstrain

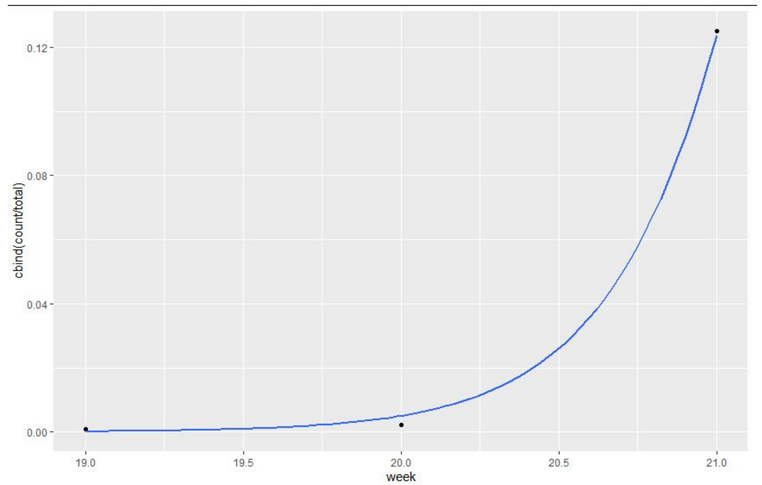
Alpha





Beta





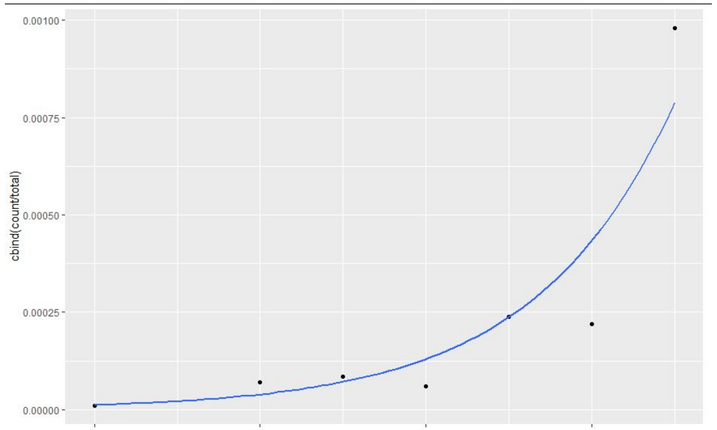
Detecting VoCs

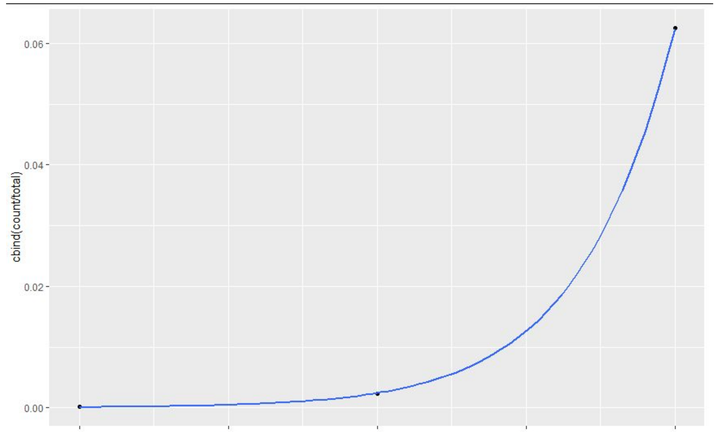
Relative Frequency of VoCs/week

GISAID

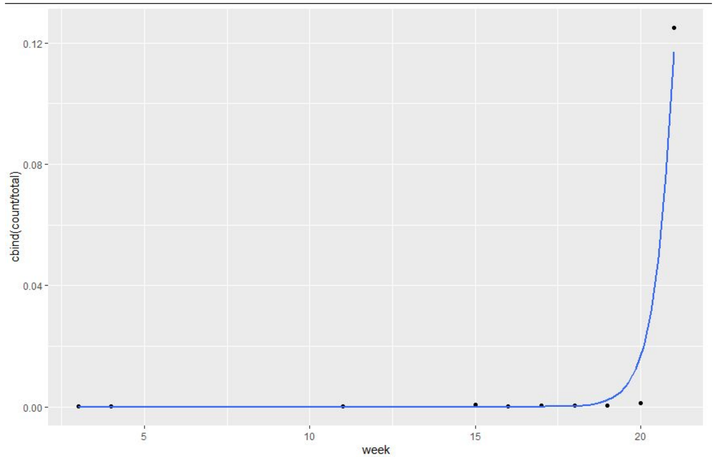
Nextstrain

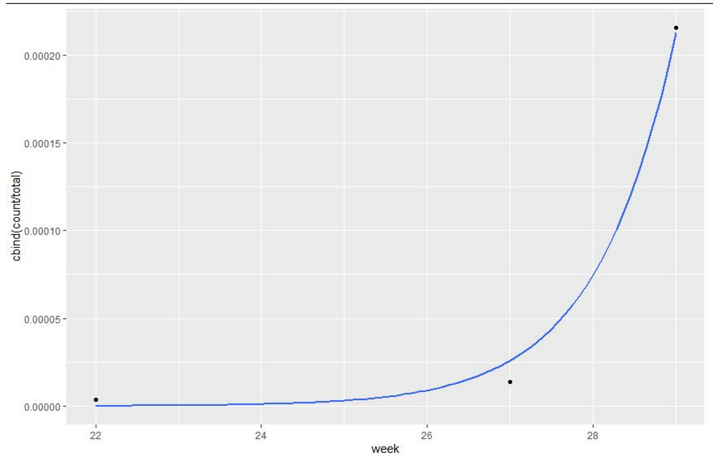
Gamma





Delta



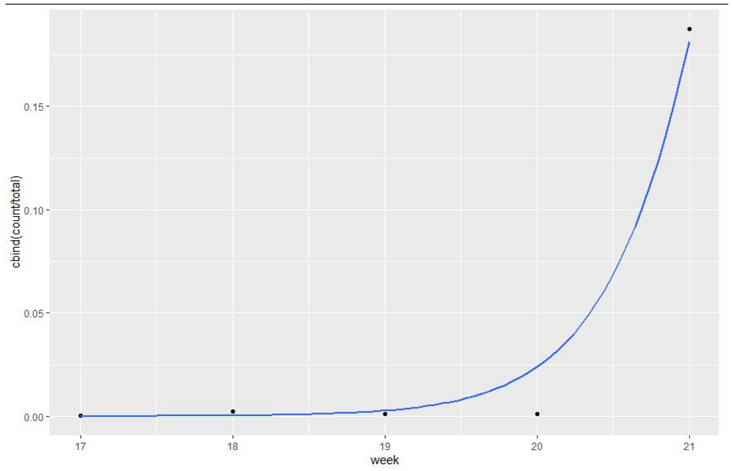


Detecting VoCs

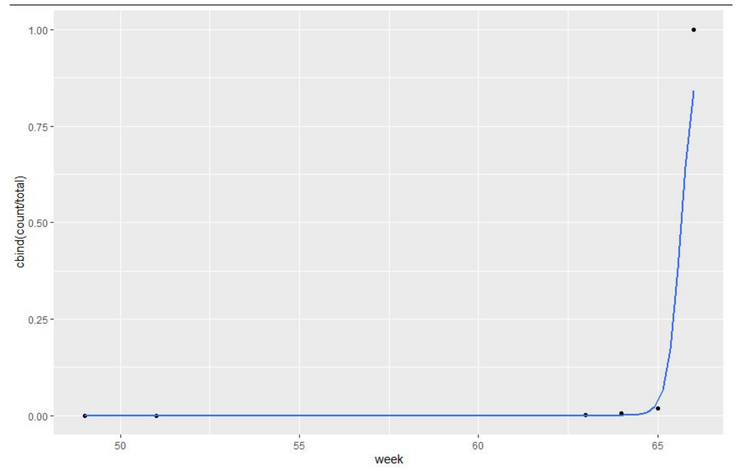
Relative Frequency of VoCs/week

Omicron

GISAID



Nextstrain

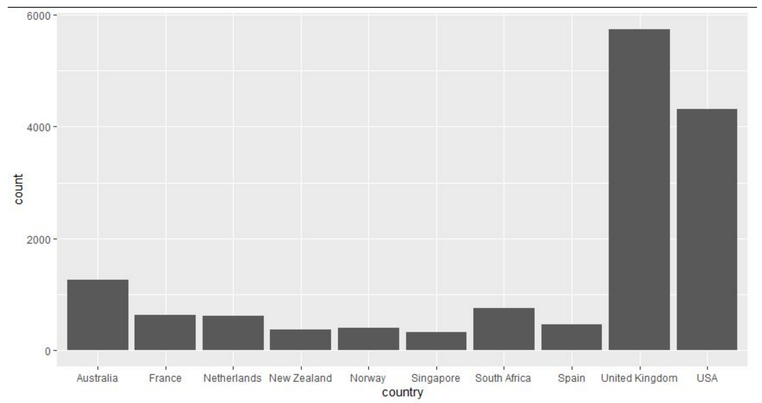


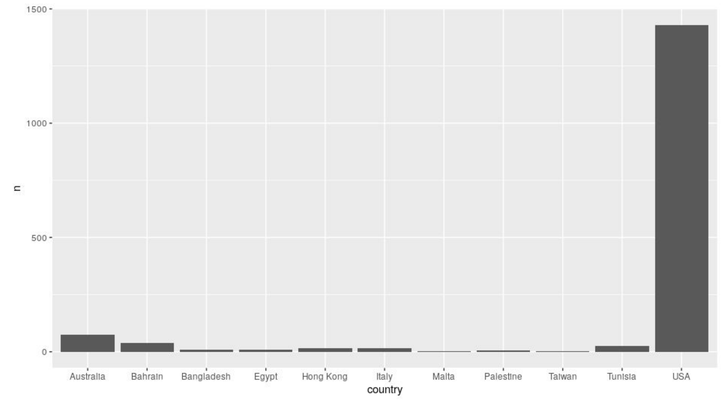
Geographic distribution

GISAID

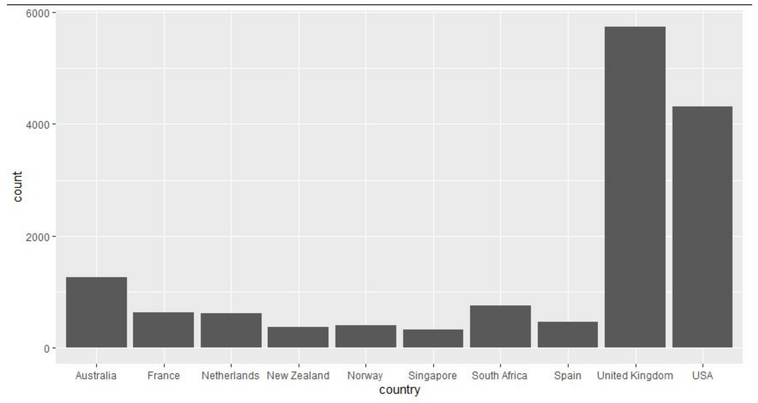
Nextstrain

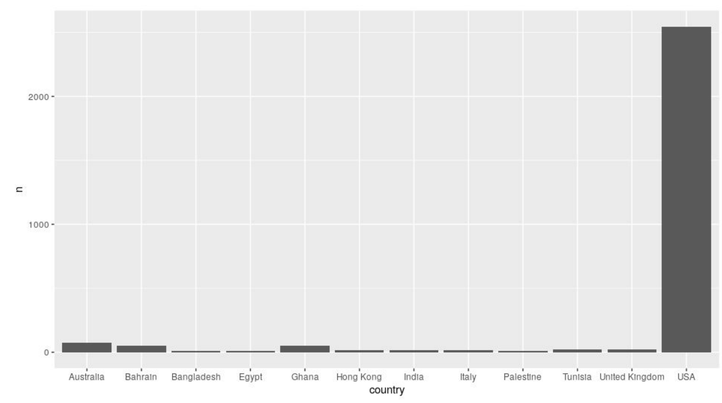
Alpha





Beta



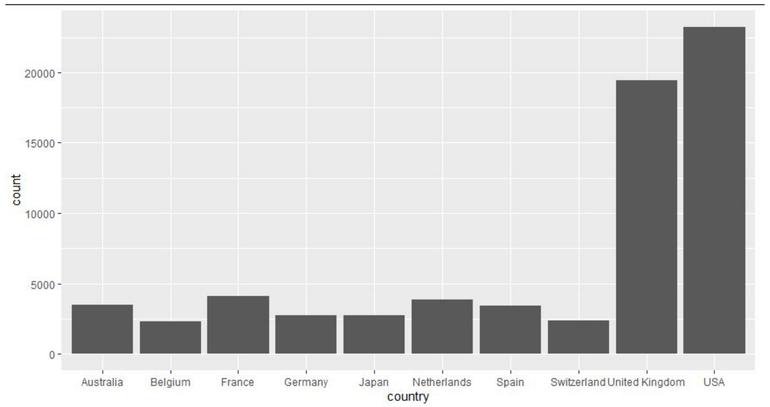


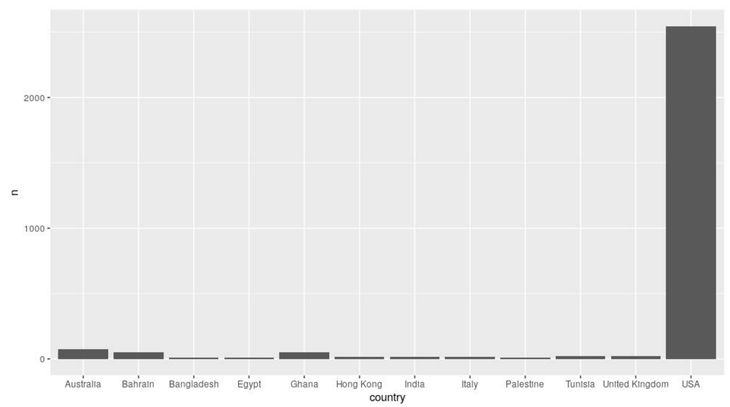
Geographic distribution

GISAID

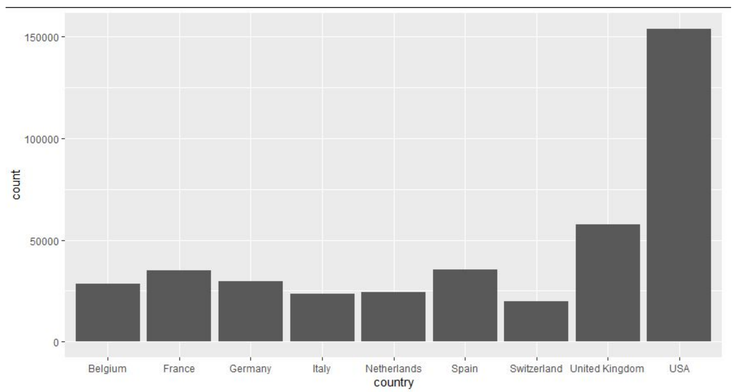
Nextstrain

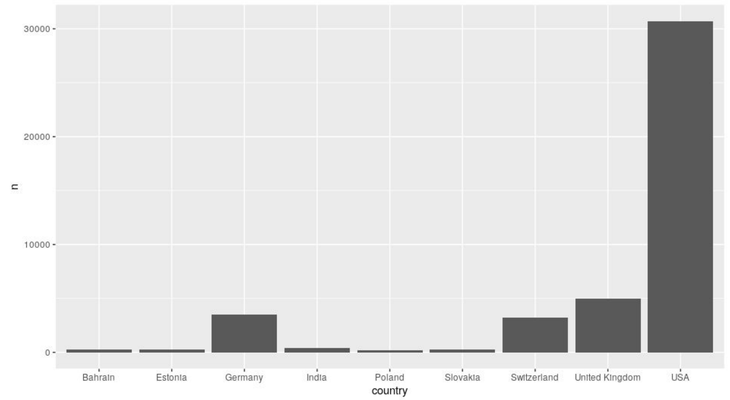
Gamma





Delta

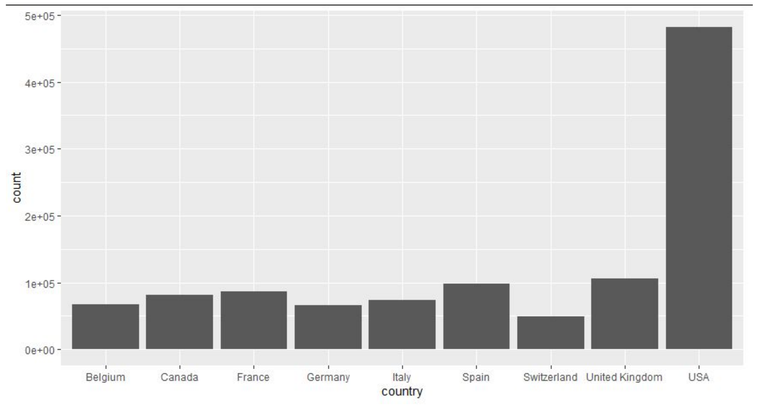




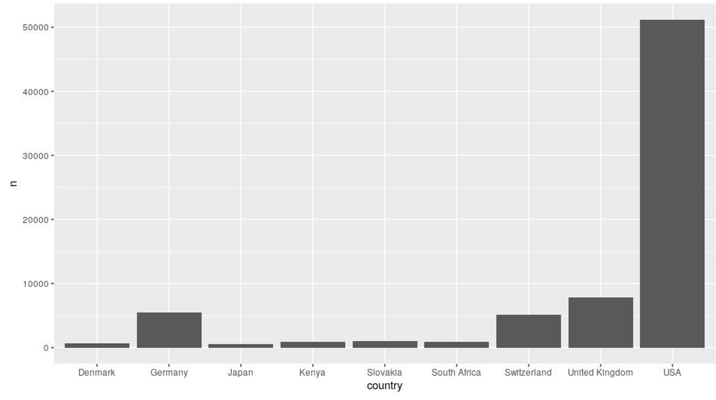
Geographic distribution

Omicron

GISAID



Nextstrain



Implications

Open data feeds significantly delays the detection of VoCs

- Smaller repository of SARS-CoV-2 sequences
- Publication delay for certain VoCs
- Limited geographic distribution—limits

GISAID

Nextstrain

References

McLaughlin, A., Montoya, V., Miller, R. L., Mordecai, G. J., Worobey, M., Poon, A., & Joy, J. B. (2021) Early and ongoing importations of SARS-CoV-2 in Canada. *MedRxiv*. https://doi.org/10.1101/2021.04.09.21255131.

Bedford, T., Hadfield, J., Hodcoft, E., Huddlestone. J., Neher, R., & Sibley, T. (2021, July 8) Extension of SARS-CoV-2 data processing to incorporate Open Data through GenBank. *Nextstrain*. https://nextstrain.org/blog/2021-07-08-ncov-open-announcement.

Bernasconi, A., Canakoglu, A., Masseroli, M., Pinoli, P., & Ceri, S. (2021). A review on viral data sources and search systems for perspective mitigation of COVID-19. *Briefings in bioinformatics*, 22(2), 664–675. https://doi.org/10.1093/bib/bbaa359.

Cohen, J. (2020, January 11) Chinese researchers reveal draft genome of virus implicated in Wuhan pneumonia outbreak. *ScienceInsider*. https://www.science.org/content/article/chinese-researchers-reveal-draft-genome-virus-implicated-wuhan-pneumonia-outbreak.

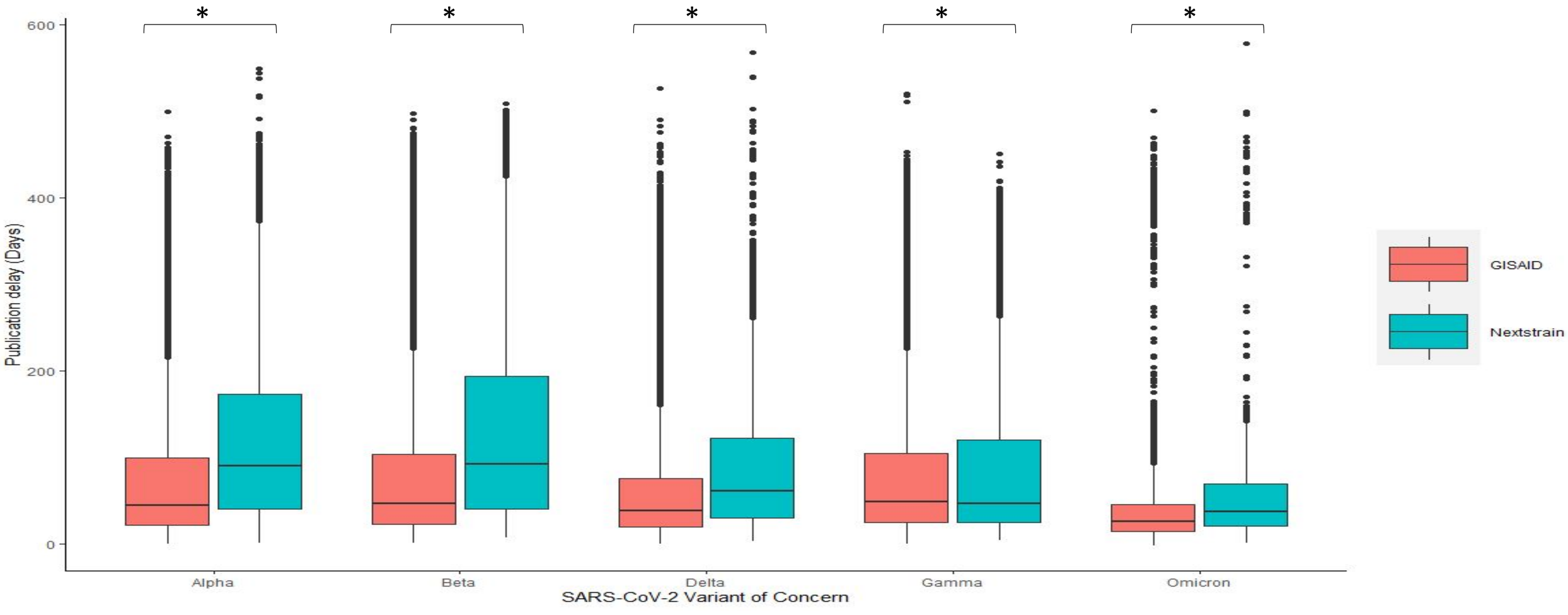
Maxem, A. (2021) Why some researchers oppose unrestricted sharing of coronavirus genome data. *Nature*, 593, 176-177. https://doi.org/10.1038/d41586-021-01194-6.

Ferreria, R.-C., Wong, E., Gugan, G., Wade, K., Liu, M., MuñozBaena, L., Chato, C., Lu, B., Olabode, A. S, Poon, A. F. Y. (2021) CoVizu: Rapid analysis and visualization of the global diversity of SARS-CoV-2 genomes. *Virus Evolution,* 7(2), 1-7.

Wadman, M. (2021, March 10). Critics decry access, transparency issues with key trove of coronavirus sequences. *ScienceInsider.* https://www.science.org/content/article/critics-decry-access-transparency-issues-key-trove-coronavirus-sequences.

	GISAID	Nextstrain
Alpha	December 2020	February 2021
Beta	December 2020	March 2021
Gamma	February 2021	March 2021
Delta	April 2021	October 2021
Omicron	December 2021	March 2022

VoC	Collection Date	Submission Date	Country
Alpha	...		
⋮			
Beta			
Gamma			
Delta			
Omicron			


*
*
*
*
*
600
400
200
0
Publication delay (Days)
Alpha
Beta
Delta
Gamma
Omicron
SARS-CoV-2 Variant of Concern
GISAID
Nextstrain

```
data:  combined$delay[combined$variant == "Alpha" & combined$database == "GISAID"] and combi
ned$delay[combined$variant == "Alpha" & combined$database == "Nextstrain"]
t = -55.856 days, df = 23235, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -46.69307 days -43.52711 days
sample estimates:
Time differences in days
mean of x mean of y
 75.74265 120.85274
```

```
data:  combined$delay[combined$variant == "Beta" & combined$database == "GISAID"] and combin
ed$delay[combined$variant == "Beta" & combined$database == "Nextstrain"]
t = -25.129 days, df = 3514, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -64.83710 days -55.45193 days
sample estimates:
Time differences in days
mean of x mean of y
 78.56992 138.71443
```

```
data:  combined$delay[combined$variant == "Delta" & combined$database == "GISAID"] and combi
ned$delay[combined$variant == "Delta" & combined$database == "Nextstrain"]
t = -63.335 days, df = 39939, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -28.35465 days -26.65236 days
sample estimates:
Time differences in days
mean of x mean of y
 57.76007  85.26357
```

```
data:  combined$delay[combined$variant == "Gamma" & combined$database == "GISAID"] and combi
ned$delay[combined$variant == "Gamma" & combined$database == "Nextstrain"]
t = -7.998 days, df = 6692.4, p-value = 1.479e-15 days
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13.907620 days  -8.432142 days
sample estimates:
Time differences in days
mean of x mean of y
 78.79891  89.96879
```

```
        Welch Two Sample t-test

data:  combined$delay[combined$variant == "Omicron" & combined$database == "GISAID"] and com
bined$delay[combined$variant == "Omicron" & combined$database == "Nextstrain"]
t = -36.475 days, df = 15967, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.70157 days -13.20208 days
sample estimates:
Time differences in days
mean of x mean of y
 34.94005  48.89187
```