Science Translational Medicine
AAAS

stm.sciencemag.org/cgi/content/full/scitranslmed.abf0202/DC1

Supplementary Materials for

Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington State

Nicola F. Müller*, Cassia Wagner, Chris D. Frazar, Pavitra Roychoudhury, Jover Lee, Louise H. Moncla, Benjamin Pelle, Matthew Richardson, Erica Ryke, Hong Xie, Lasata Shrestha, Amin Addetia, Victoria M. Rachleff, Nicole A. P. Lieberman, Meei-Li Huang, Romesh Gautom, Geoff Melly, Brian Hiatt, Philip Dykema, Amanda Adler, Elisabeth Brandstetter, Peter D. Han, Kairsten Fay, Misja Ilcisin, Kirsten Lacombe, Thomas R. Sibley, Melissa Truong, Caitlin R. Wolf, Michael Boeckh, Janet A. Englund, Michael Famulare, Barry R. Lutz, Mark J. Rieder, Matthew Thompson, Jeffrey S. Duchin, Lea M. Starita, Helen Y. Chu, Jay Shendure, Keith R. Jerome, Scott Lindquist, Alexander L. Greninger, Deborah A. Nickerson, Trevor Bedford*

*Corresponding author. Email: nicola.felix.mueller@gmail.com; tbedford@fredhutch.org

Published First Release 3 May 2021, *Sci. Transl. Med.*
DOI: 10.1126/scitranslmed.abf0202

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at stm.sciencemag.org/cgi/content/full/scitranslmed.abf0202/DC1)

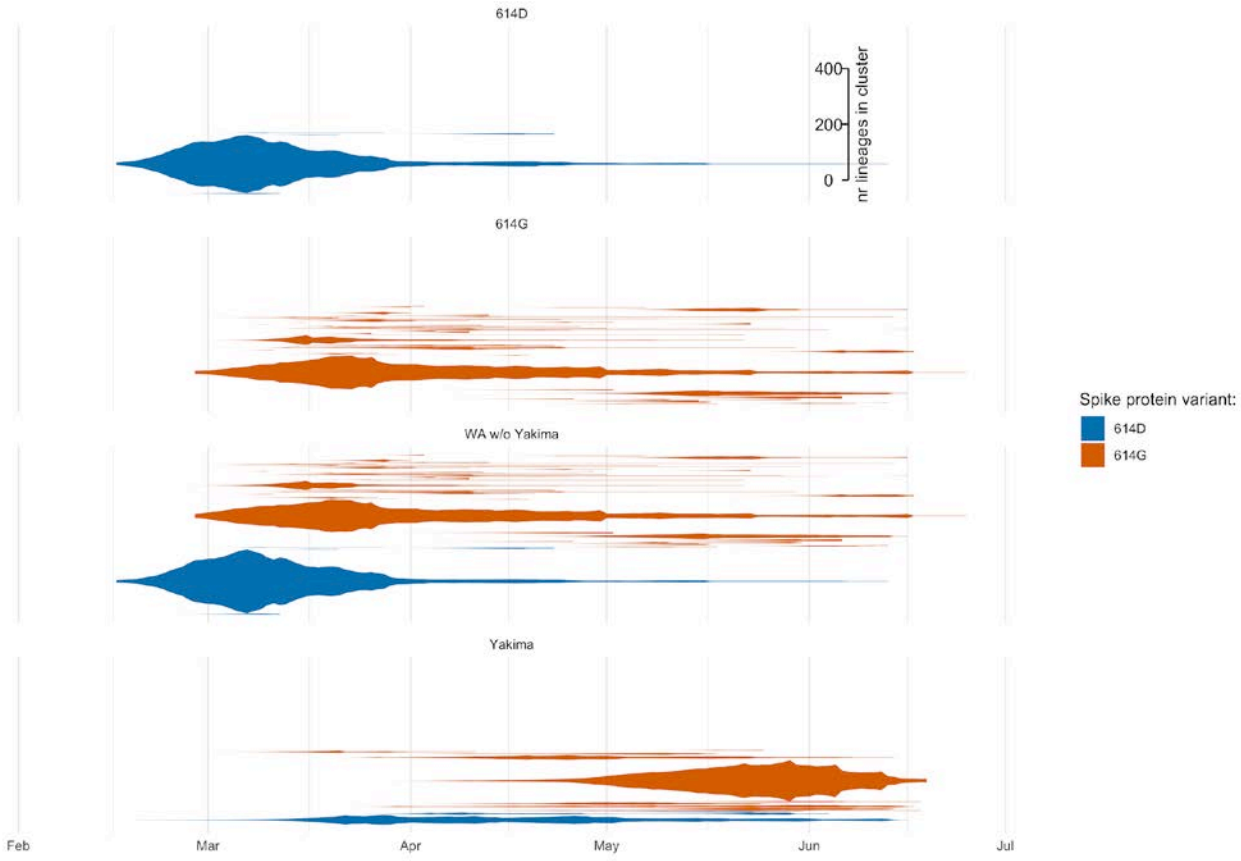


Fig. S1. Number of lineages through time for different local transmission clusters. Here we show the number of lineages in each local transmission cluster (y-axis) over time (x-axis). The different plots show the lineage through time plots for the different datasets analyses here.

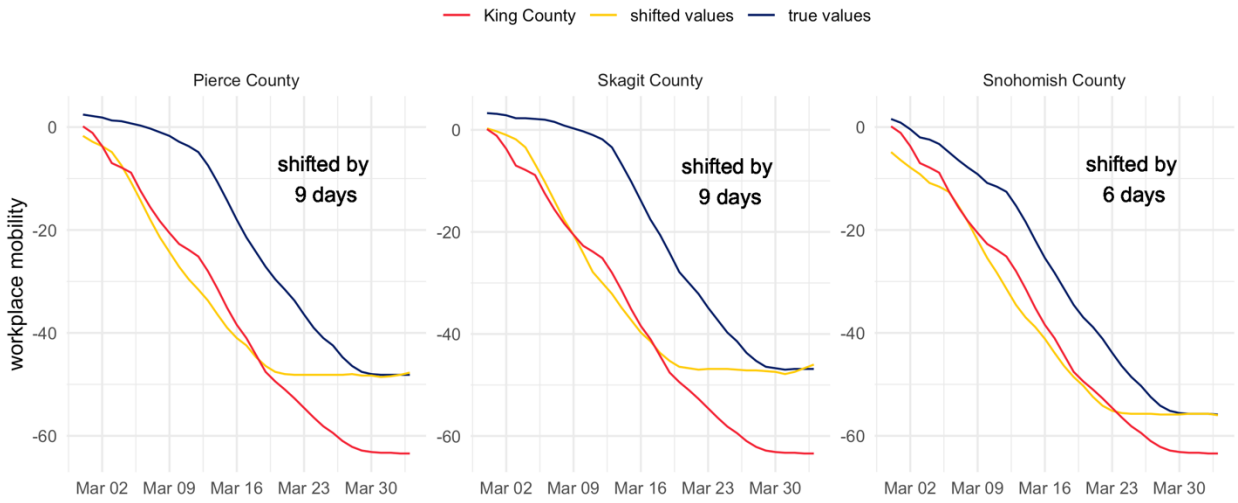


Fig. S2. Workplace mobility trends of different counties in Washington State compared to King County. Each plot shows the workplace mobility trend of King County and compares it to either Pierce County, Skagit County, or Snohomish County (in blue). The yellow line shows the mobility trend of a county shifted to match the trends in King County (in red). The number of days that the trend line is shifted by is shown in each subplot.

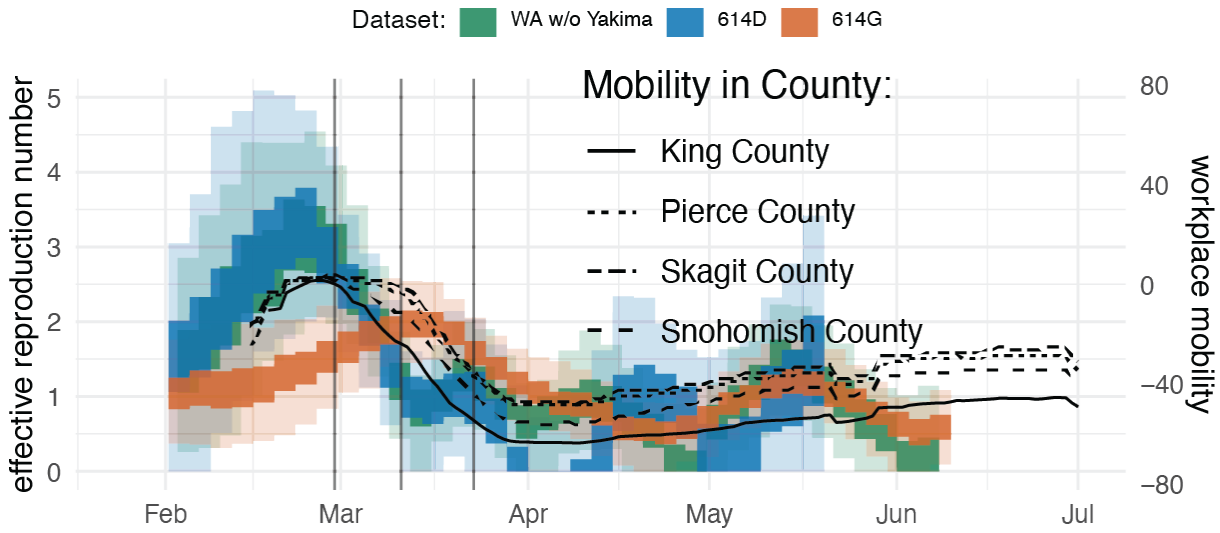


Fig. S3. R_e estimates using the coalescent skygrowth model compared to Google mobility data.

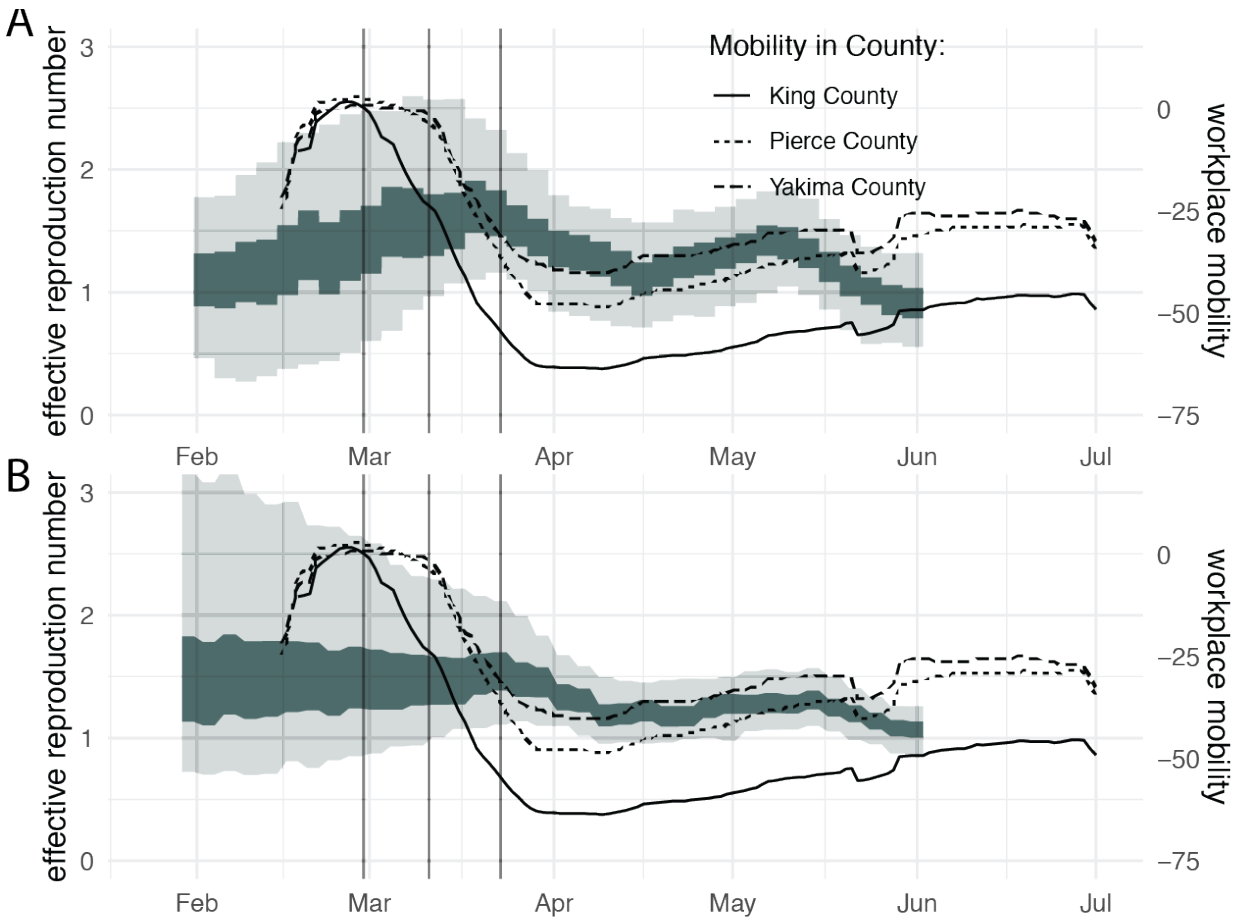
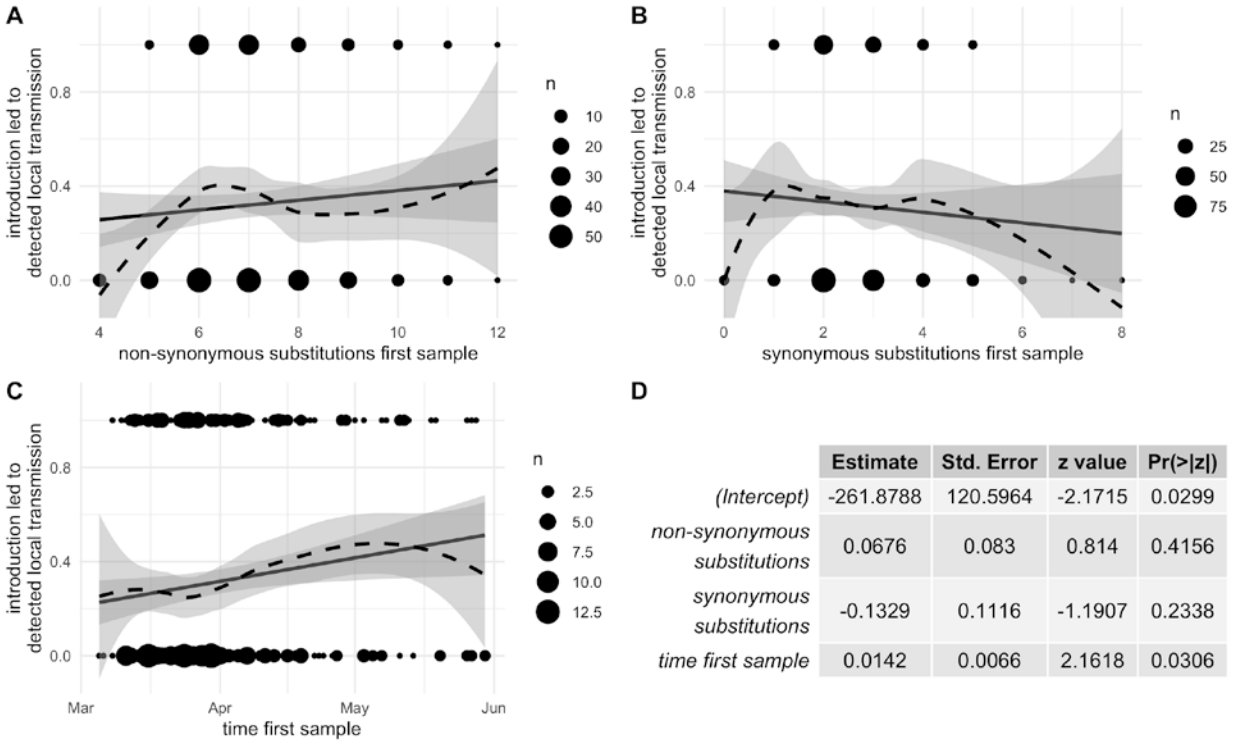


Fig. S4. Effective reproduction number and workplace mobility in Yakima County. Here, we show the effective reproduction number estimates over time in Yakima County using the birth-death skyline model (A) and the coalescent skygrowth model (B). The inner band shows the 50% highest posterior density (HPD) interval and the outer band, the 95% HPD interval. Additionally, we compared those estimates to mobility trends in Yakima County and (as a reference) King and Pierce County. The mobility trends are shown as a 7-day moving average.



D

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-261.8788	120.5964	-2.1715	0.0299
non-synonymous substitutions	0.0676	0.083	0.814	0.4156
synonymous substitutions	-0.1329	0.1116	-1.1907	0.2338
time first sample	0.0142	0.0066	2.1618	0.0306

Fig. S5. Substitutions and success of a SARS-CoV-2 introduction. Here, we look at whether there is a relationship between the number of RNA (A) or amino acid (B) substitutions or the timing of the introduction (C) on whether an introduction leads to detectable local spread. Detectable local transmission is defined as a local outbreak cluster with more than 1 sequenced sample in it. The lines denote linear (solid line) and loess (dashed line) regressions. In (D), we test if the time of the first sample in each local outbreak cluster, the number of synonymous or non-synonymous substitutions are significant predictors of an introduction having lead to detectable local transmission using a generalized linear model.

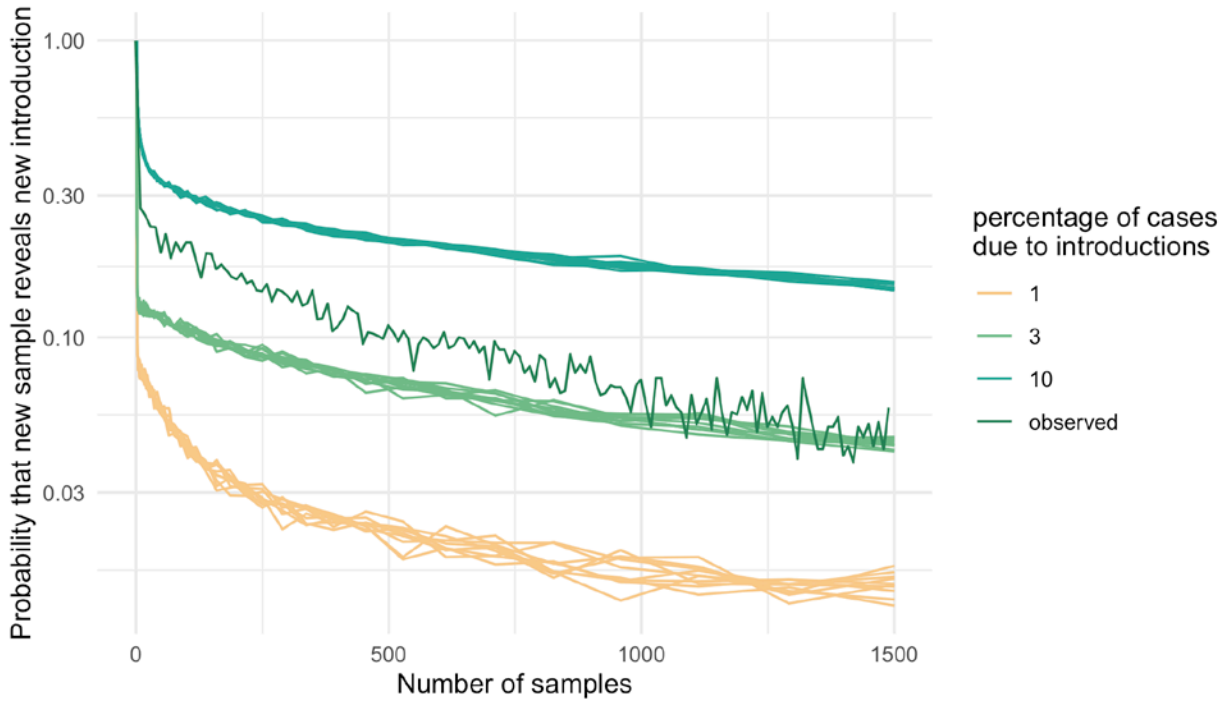


Fig. S6. Probability that a newly sampled case reveals a new introduction. Here we compared the probability that adding a new sequence to a dataset reveals a new introductions between what we observed empirically and when we simulated clusters using different percentages of introductions. To do so, we randomly chose n samples (x-axis) and then added one additional sample. We then estimated the probability that this additional sample revealed a new introduction (y-axis). We repeated the procedure for simulated clusters with different percentages of introductions in overall cases.

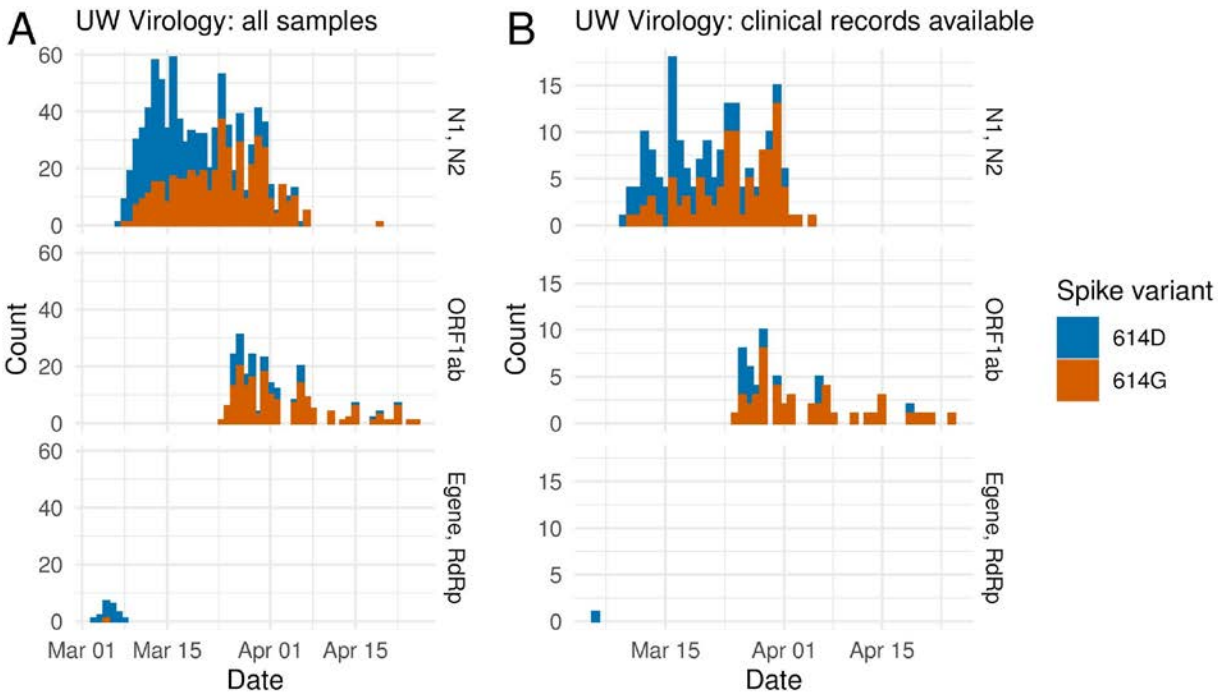
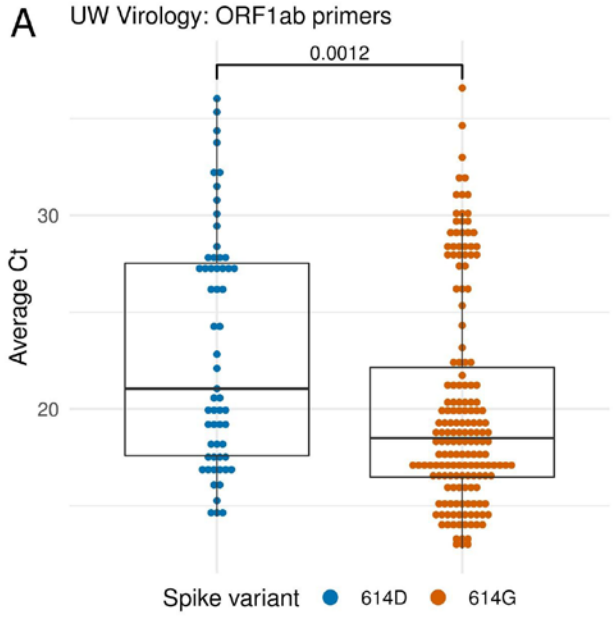


Fig. S7. Histogram of primers used by UW Virology across time. A All UW Virology samples. **B** Only UW Virology samples with clinical records available.



B

UW Virology: ORF1ab primers

Variable	Coeff. est.	Std. Error	p–value
Intercept	18.02	1.53	<2e–16***
614G	–1.50	1.01	0.143
Age	0.031	0.021	0.138
Days after symptom onset	0.35	0.079	2.5e–05***

C

SCAN

Variable	Coeff. est.	Std. Error	p–value
Intercept	26.10	1.96	<2e–16***
614G	–3.02	1.22	0.017*
Age	–0.05	0.040	0.23
Days after symptom onset	0.34	0.10	0.0026**

Fig. S8. Comparison of cycle threshold (Ct) across SARS-CoV-2 Spike variant. (A) Boxplot of Ct with ORF1ab primers by amino acid at Spike 614.GLM analysis of Ct values from ORF1ab (**B**) and SCAN (**C**) primers using several different predictors.

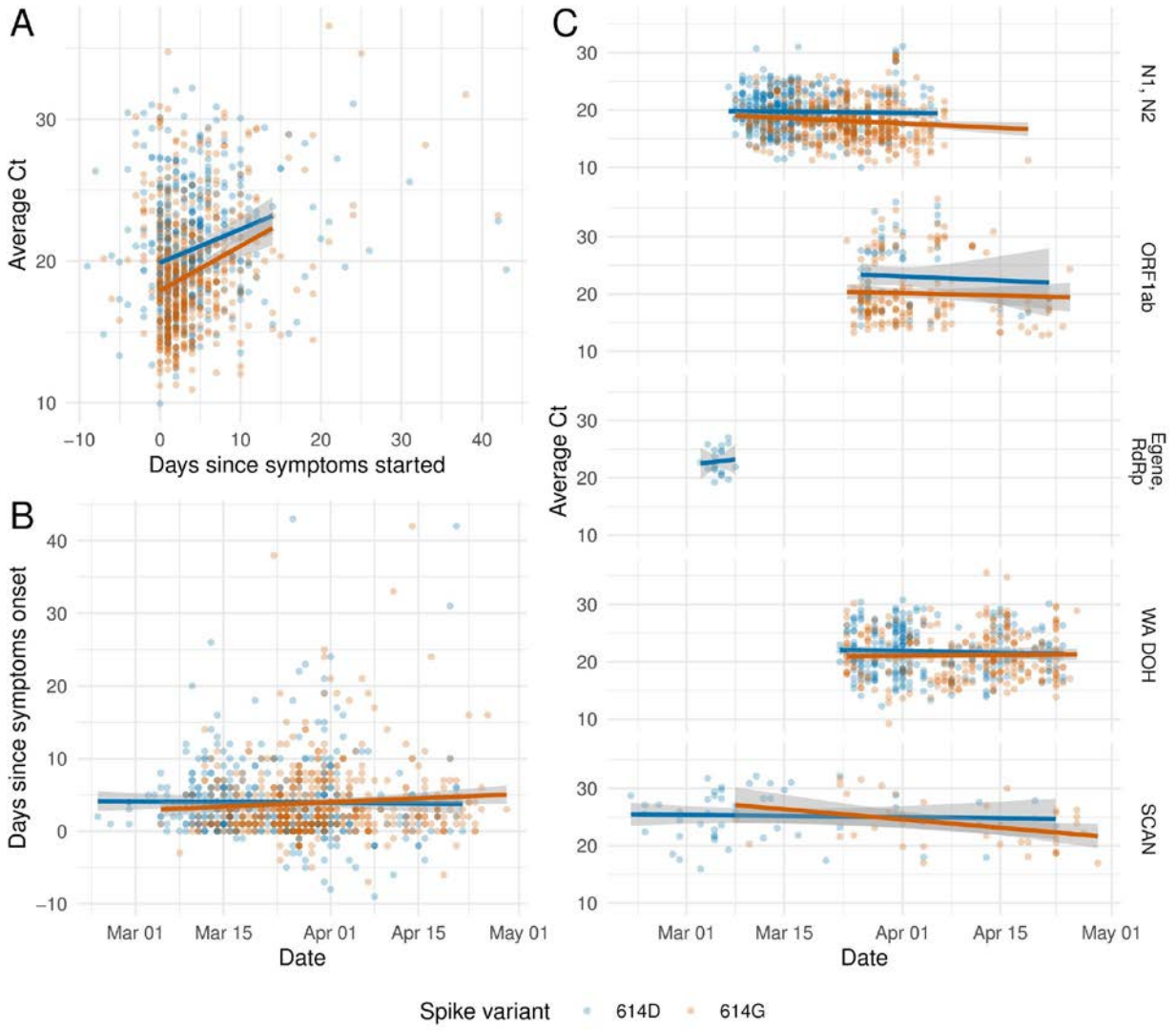
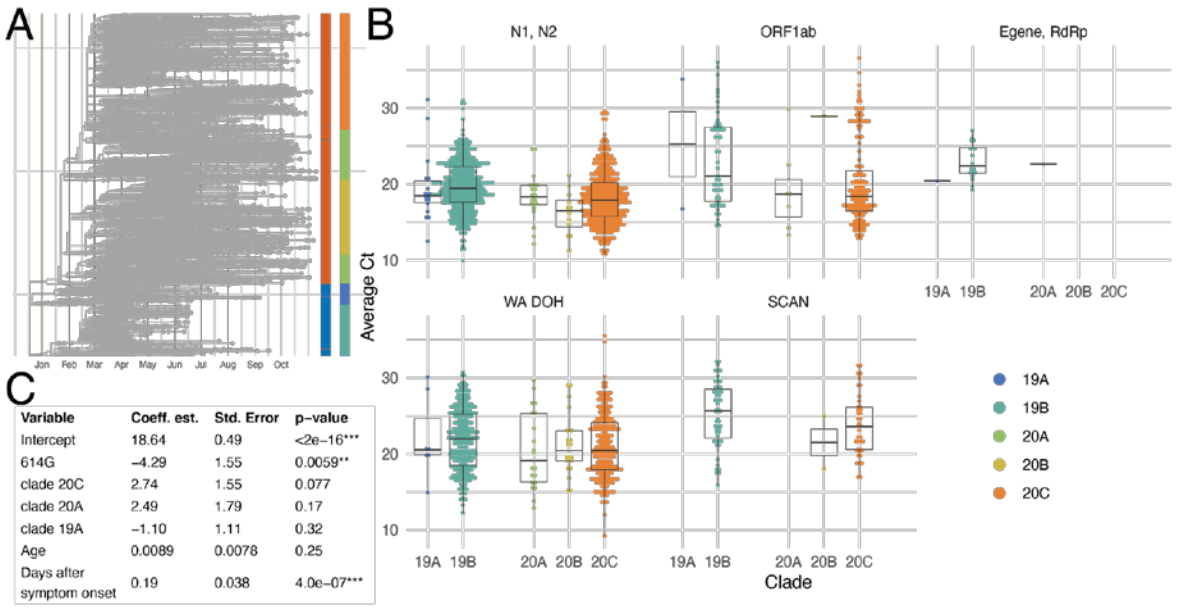


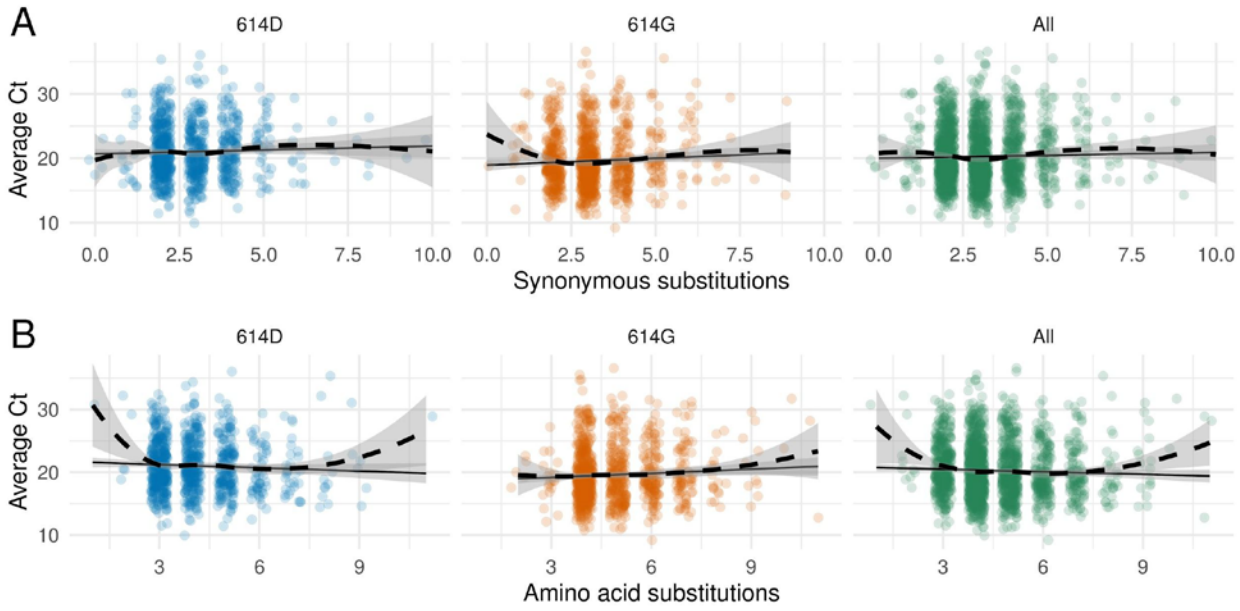
Fig. S9. Symptom and cycle threshold (Ct) values across time. A Scatterplot of Ct versus days since symptom onset by spike variant. **B** Scatterplot of days since symptom onset by date. **C** Average Ct by date split by primer set**.** In A & B, data is shown for all samples with Ct and symptom onset available (n=977); in C, data is shown for all samples (n=1743).



C

Variable	Coeff. est.	Std. Error	p-value
Intercept	18.64	0.49	<2e-16***
614G	-4.29	1.55	0.0059**
clade 20C	2.74	1.55	0.077
clade 20A	2.49	1.79	0.17
clade 19A	-1.10	1.11	0.32
Age	0.0089	0.0078	0.25
Days after symptom onset	0.19	0.038	4.0e-07***

Fig. S10. Comparing cycle threshold (Ct) by viral clade. A Phylogenetic tree showing distribution of 614D (blue) vs. 614G (orange) variants in the first column and spread across viral clades 19A, 19B, 20A, 20B, and 20C in the second column. **B** Comparison between cycle threshold across viral clade for each primer type. **C** GLM analysis of Ct values using samples amplified with N1, N2 primers considering clade, 614G variant, age, and days since symptom onset as predictors.



C

Spike 614D

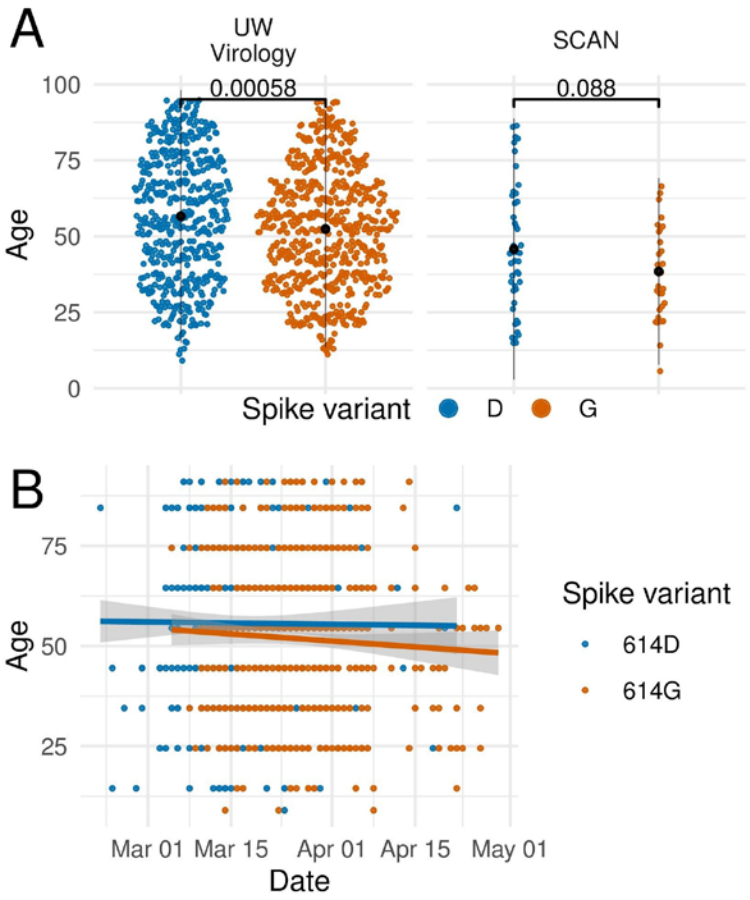
Variable	Coeff. est.	Std. Error	p–value
Intercept	20.14	0.97	<2e–16***
Amino acid substitutions	–0.42	0.16	0.011*
Synonymous substitutions	0.22	0.16	0.18
Days after symptom onset	0.12	0.034	0.00031***
Week since start of WA epidemic	0.057	0.16	0.73
ORF1ab primers	1.39	0.77	0.073
Egene, RdRp primers	1.42	1.39	0.31
WA DOH primers	2.18	0.59	0.00024***
SCAN primers	5.20	0.79	1.1e–10***

D

Spike 614G

Variable	Coeff. est.	Std. Error	p–value
Intercept	18.82	1.09	<2e–16**
Amino acid substitutions	0.061	0.14	0.66
Synonymous substitutions	0.20	0.16	0.21
Days after symptom onset	0.23	0.034	4.0e–11***
Week since start of WA epidemic	–0.43	0.14	0.0026**
ORF1ab primers	1.88	0.49	0.00013***
Egene, RdRp primers	4.33	3.91	0.27
WA DOH primers	3.96	0.54	6.4e–13***
SCAN primers	5.21	0.96	7.9e–08***

Fig. S11. Cycle threshold by number of substitutions. Number of synonymous (**A**) and amino acid (**B**) substitutions versus Ct by D614G variant. GLM analysis of Ct values with amino acid substitutions, synonymous substitutions, and other known predictors for 614D (**C**) and 614G (**D**) variants.



C

Variable	Coeff. est.	Std. Error	p–value
Intercept	57.18	2.26	<2e–16***
614G	–3.52	1.32	0.0077**
Week	–0.31	0.44	0.48

Fig. S12. Age of infected individuals by 614D or 614G variant over time. (A) Age of infected individuals in UW Virology and SCAN samples according to D614G variant. Mean age and two standard deviations are shown in black. (**B)** Age of infected individuals over time partitioned by D614G variant. (**C)** GLM of patient age predicted by D614G variant and sampling week.

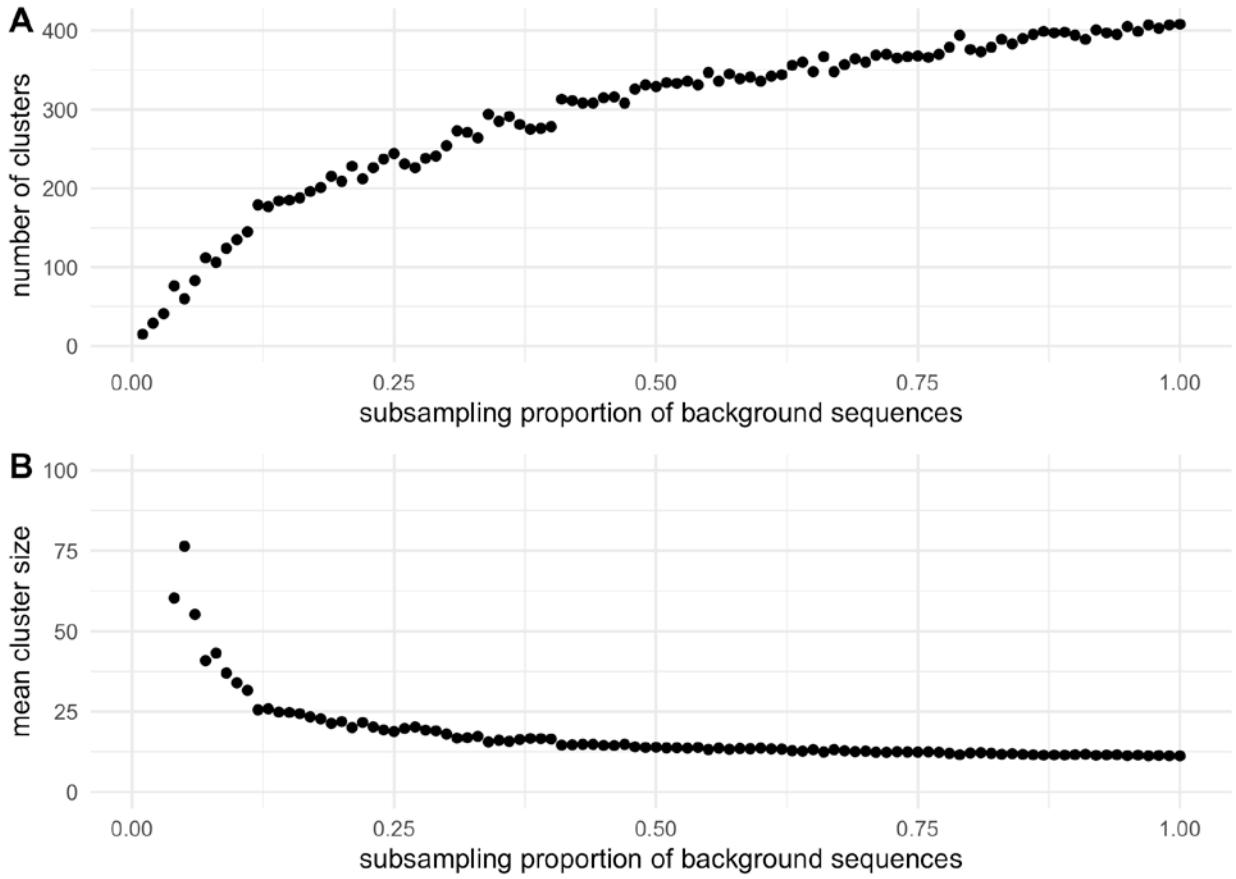


Fig. S13 Dependence of the local outbreak clusters and the number of background sequences used. Here, we estimated the number (**A)** of mean size (**B**) of local outbreak clusters depending on the number of background sequences used during the clustering. To do so, we randomly subsampled different proportions of the background sequences (x-axis) and repeat the clustering. We then computed the number of clusters and the average sizes of cluster(y-axis) depending on the proportion of background samples used relative to the full dataset.

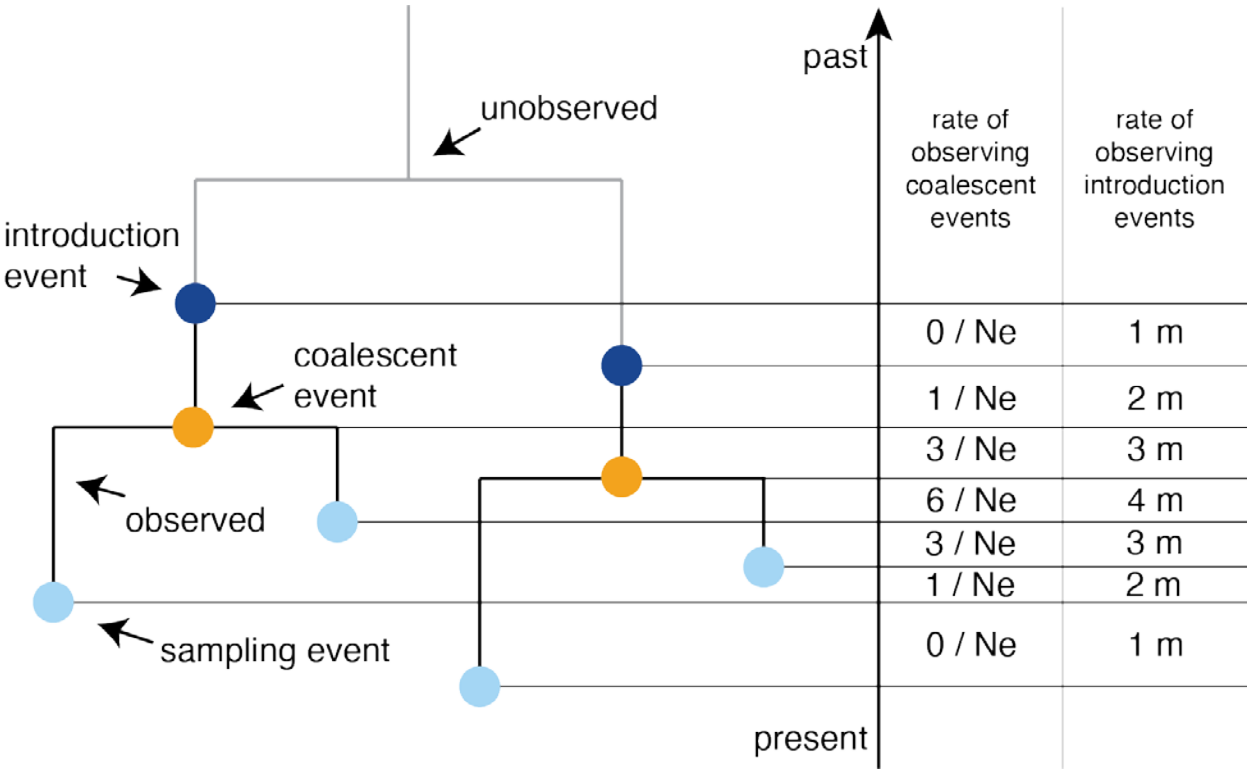


Fig. S14. Principle of the multi-tree coalescent. The tree above shows a full hypothetical phylogenetic tree with two independent introductions from an outside population (dark blue) and subsequent local spread. The black branches are observed parts of the phylogeny and denote branches of a local transmission tree. The grey branches are unobserved and denote part of the transmission history that happened outside of the population of interest. Within the population of interest, we can observe sampling events (light blue) and coalescent events (yellow). The rate of observing a coalescent event is equal to the number of pairs of co-existing (black) lineages in any local transmission cluster divided by 2 * Ne. The rate of observing an introduction event is given by the number of co-existing black lineages and the rate of introduction.

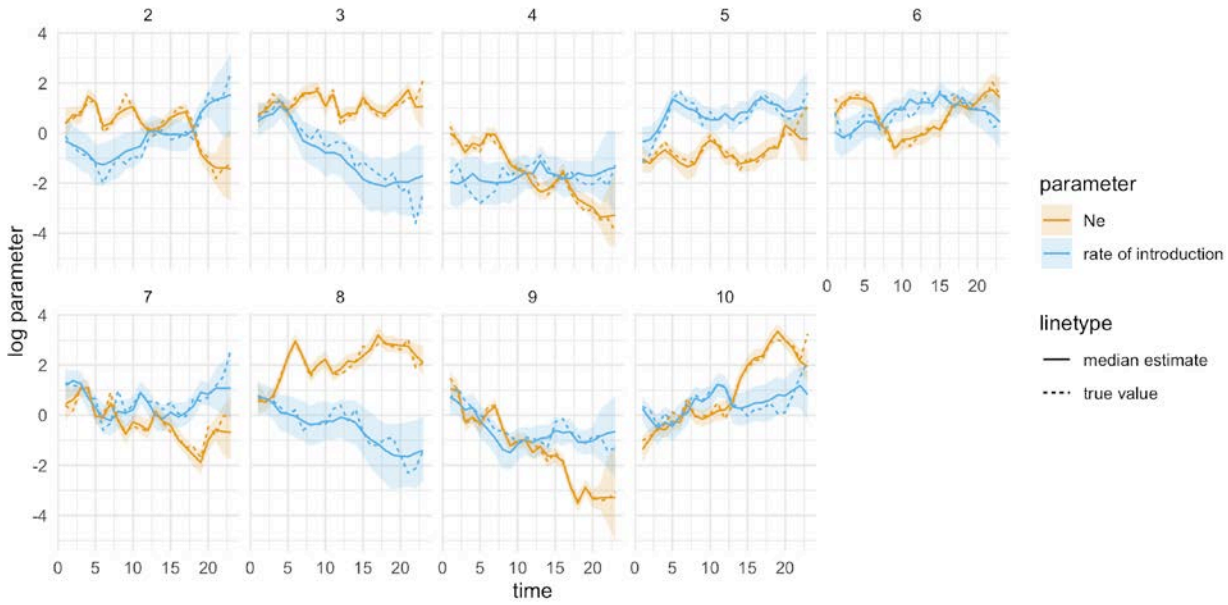


Fig. S15. Estimation of effective population sizes and rates of introductions from simulations. Here, we inferred effective population sizes and rates of introductions from phylogenetic trees, simulated under the structured coalescent when conditioning on observing a migration history. Of the ten runs, one was discarded due to bad convergence.

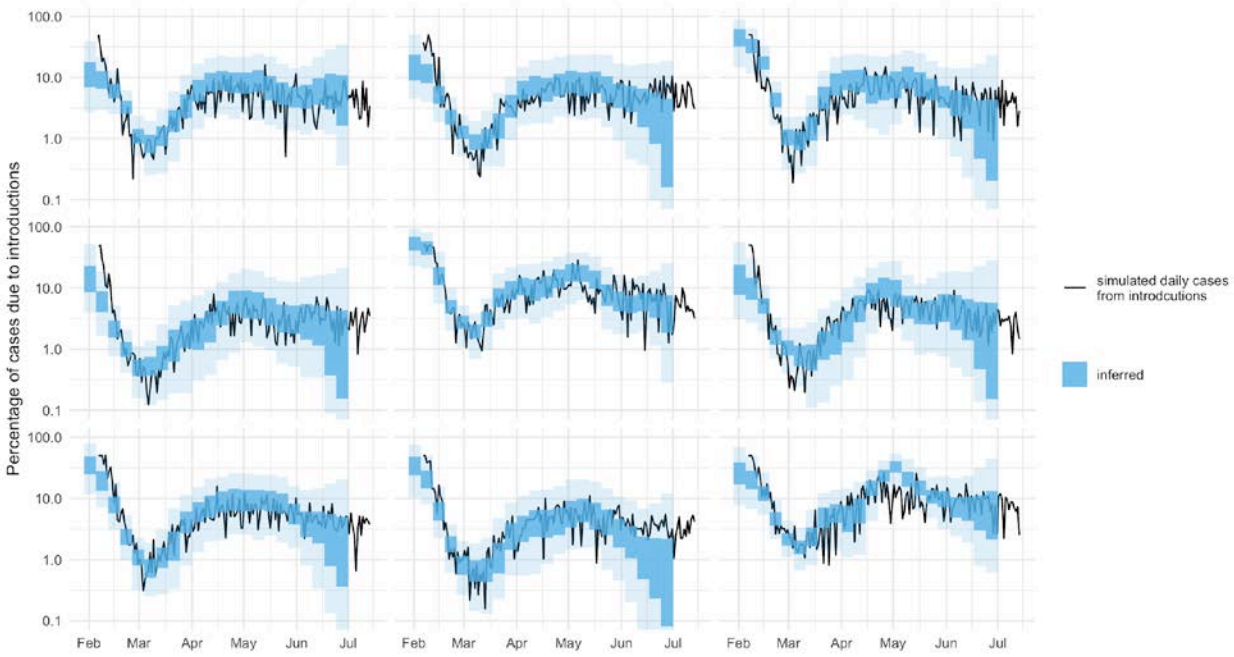


Fig. S16. Estimation of the percentage of new cases due to introductions from simulations. Here, we tested how well we can retrieve the percentage of new cases due to introductions over time from simulations. To do so, we simulated a local outbreak using a constant rate of introduction. We then simulated genetic sequences and then used the local transmission cluster to estimate the percentage of introductions in blue using the multi-tree coalescent.

Table S1. GLM of Ct with N1, N2 primers in patients at UW affiliates

Variable	Coefficient estimate	Std. Error	p-value
Intercept	17.45	0.81	<2e-16***
614G	-1.04	0.48	0.032*
Male	1.09	0.48	0.024*
Age	0.015	0.013	0.28
Active cancer or immunocompromised	-0.17	0.77	0.83
Hospitalized	1.02	0.75	0.18
Critical care or deceased	-0.52	0.97	0.60

Table S2. GLM of Ct with ORF1ab primers in patients at UW affiliates

Variable	Coefficient estimate	Std. Error	p-value
Intercept	17.32	3.14	9.7e-07***
614G	0.35	1.79	0.84
Male	1.34	1.67	0.43
Age	0.029	0.041	0.48
Active cancer or immunocompromised	1.63	2.88	0.57
Hospitalized	4.89	1.97	0.016*
Critical care or deceased	-2.00	2.85	0.48

The following data files are available in the online version of the supplement:

Data file S1. GISAID acknowledgment table (.tsv)

Data file S2. Cycle threshold values for isolates