CORONAVIRUS

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

Nicola F. Müller[1]*†, Cassia Wagner[1,2†], Chris D. Frazar[2†], Pavitra Roychoudhury[1,3], Jover Lee[1], Louise H. Moncla[1], Benjamin Pelle[2], Matthew Richardson[2], Erica Ryke[2], Hong Xie[3], Lasata Shrestha[3], Amin Addetia[3], Victoria M. Rachleff[1,3], Nicole A. P. Lieberman[3], Meei-Li Huang[3], Romesh Gautom[4], Geoff Melly[4], Brian Hiatt[4], Philip Dykema[4], Amanda Adler[5], Elisabeth Brandstetter[6], Peter D. Han[2], Kairsten Fay[1], Misja Ilcisin[1], Kirsten Lacombe[5], Thomas R. Sibley[1], Melissa Truong[2], Caitlin R. Wolf[6], Michael Boeckh[1,6,7], Janet A. Englund[5,8], Michael Famulare[9], Barry R. Lutz[7,10], Mark J. Rieder[7], Matthew Thompson[11], Jeffrey S. Duchin[12,13], Lea M. Starita[2,7], Helen Y. Chu[12,7], Jay Shendure[2,7,14], Keith R. Jerome[1,3], Scott Lindquist[4], Alexander L. Greninger[1,3‡], Deborah A. Nickerson[2,7‡], Trevor Bedford[1,2,7]*‡

The rapid spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has gravely affected societies around the world. Outbreaks in different parts of the globe have been shaped by repeated introductions of new viral lineages and subsequent local transmission of those lineages. Here, we sequenced 3940 SARS-CoV-2 viral genomes from Washington State (USA) to characterize how the spread of SARS-CoV-2 in Washington State in early 2020 was shaped by differences in timing of mitigation strategies across counties and by repeated introductions of viral lineages into the state. In addition, we show that the increase in frequency of a potentially more transmissible viral variant (614G) over time can potentially be explained by regional mobility differences and multiple introductions of 614G but not the other variant (614D) into the state. At an individual level, we observed evidence of higher viral loads in patients infected with the 614G variant. However, using clinical records data, we did not find any evidence that the 614G variant affects clinical severity or patient outcomes. Overall, this suggests that with regard to D614G, the behavior of individuals has been more important in shaping the course of the pandemic in Washington State than this variant of the virus.

INTRODUCTION

After its emergence near the end of November or beginning of December 2019 in Wuhan, China, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) rapidly spread around the world (*1*). In the United States, the first reported case of coronavirus disease 2019 (COVID-19), the disease caused by SARS-CoV-2, was found in Washington State on 19 January 2020 in a traveler who had returned from China 4 days earlier. Until the end of February, no additional cases of COVID-19 were reported in Washington State.

At the end of February, however, a case of COVID-19 was reported in Snohomish County, the same county where the initial case was reported. This case had no known travel history and constitutes the first reported case of community transmission in Washington State (*2*). Although genetically closely related to the initial case, the later sequenced cases share a common ancestor in early February and have been reported to likely be due to an independent introduction of the virus (*2*).

After these initial introductions, SARS-CoV-2 has been introduced repeatedly into Washington State from different parts of the globe. Viruses introduced later differed genetically from those introduced earlier, most notably in one amino acid in the spike protein that facilitates viral entry and includes the receptor-binding domain. Since its first occurrence, this amino acid substitution from aspartate (D) to glycine (G) at position 614 of the Spike protein increased in relative frequency around the world (visible at https://nextstrain.org/ncov/global?c=gt-S_614) and now represents the vast majority of all new cases of COVID-19 (*3*–*5*). This increase in relative frequency of the 614G variant has been proposed to be due to higher transmissibility of the 614G variant over the 614D variant (*4*, *6*). A modest increase in viral load has been observed in patients infected with the 614G variant (*4*, *7*). Recently, multiple in vitro studies in human cell lines found a three- to ninefold increase in infectivity of the 614G variant (*5*, *8*, *9*). However, it remains unclear whether these population-level trends are due to higher transmissibility of the virus or simply due to founder effects owing to strong bottlenecks when SARS-CoV-2 spread globally, as the D614G variant was introduced early on in the European COVID-19 epidemic and spread from Europe to the rest of the world.

Washington State differs regionally, from more densely populated areas at the coast to more sparsely populated areas inland. We here focused on differences between the spread on lineages of 614D and 614G in the context of regional differences within Washington State.

[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA. [2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA. [3]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA 98195, USA. [4]Washington State Department of Health, Shoreline, WA 98155, USA. [5]Seattle Children's Research Institute, Seattle, WA 98101, USA. [6]Department of Medicine, Division of Allergy and Infectious Diseases, University of Washington, Seattle, WA 98195, USA. [7]Brotman Baty Institute for Precision Medicine, Seattle, WA 98195, USA. [8]Department of Pediatrics, University of Washington, Seattle, WA 98105, USA. [9]Institute for Disease Modeling, Bellevue, WA 98105, USA. [10]Department of Bioengineering, University of Washington, Seattle, WA 98105, USA. [11]Department of Global Health, University of Washington, Seattle, WA 98195, USA. [12]Department of Medicine, Division of Allergy and Infectious Diseases, University of Washington, Seattle, WA 98195, USA. [13]Public Health - Seattle & King County, Seattle, WA98121, USA. [14]Howard Hughes Medical Institute, Seattle, WA 98195, USA.
*Corresponding author. Emails: nicola.felix.mueller@gmail.com (N.F.M.); tbedford@fredhutch.org (T.B.)
†These authors contributed equally.
‡Co-senior authors.

Extensive local spread of SARS-CoV-2 was first detected in King County, which includes the city of Seattle. King County was also the first region in the state to take action to curb the spread of SARS-CoV-2, including several large companies in the area mandating work from home in early March 2020 (*10*). After a statewide lockdown, new cases began to fall in the whole state, except for Yakima County, where cases peaked substantially later than in the rest of the state.

Using viral genetic sequence data isolated from patients in Washington State between February and July 2020, we tested the impact of temporal differences in county level workplace mobility trends, as well as the role of introductions from outside the state in driving case loads. We additionally investigated potential transmissibility differences between the two spike variants by comparing viral loads using cycle thresholds for viral quantification. Last, we investigated whether the D614G amino acid substitution led to more severe disease in patients infected with SARS-CoV-2.

RESULTS

The Washington State outbreak was caused by repeated introductions and shaped by temporal differences in mobility reductions

We sequenced 3940 viruses from Washington State collected between February and July 2020 and used these sequences alongside other publicly available sequences from elsewhere in the world to characterize transmission dynamics. We observed that SARS-CoV-2 entered Washington State from different parts of the world and subsequently spread locally, evident as clusters of genetically similar Washington State viruses in the global phylogeny (Fig. 1A). In early February, an introduction of a 614D variant (*2*, *11*) fueled much of the early outbreak in March and April, but this lineage was supplanted through multiple introductions of 614G, and past April, the majority of viruses were 614G (Fig. 1).

To analyze the introduction and local spread of SARS-CoV-2 in Washington State, we first split these sequences into different local transmission clusters, which we defined as groups of sequences that originated from a single introduction into Washington State. To do so, we use a parsimony-based clustering approach, considering Washington State and everything outside Washington State as the two possible locations for parsimony clustering. The local transmission clusters obtained are shown at https://nextstrain.org/groups/blab/ncov/wa-phylodynamics?c=cluster_size, and their size distribution and D614G makeup are shown in fig. S1. We then used these local transmission clusters to analyze the spread of SARS-CoV-2 in the state using two phylodynamic approaches. First, we estimated the effective reproduction number (R_e) using a birth-death approach (*12*), where we treated each individual local transmission cluster as independent observation of the same underlying population process (*13*). Next, we estimated effective population sizes over time and the degree of introductions using a coalescent skyline approach (*14*). To do so, we assumed that all sequences that clustered together were the result of local transmission and each individual cluster was the result of one introduction into Washington State. We then modeled the whole process as a structured coalescent process (*15*, *16*), where we assumed the migration history on the basis of the previous clustering (see Materials and Methods for details). In contrast to the birth-death model, the coalescent is conditioned on sampling, meaning that the information about population-level trends comes from the phylogenetic tree itself and not from the number of sequences through time.

We performed these phylodynamic analyses for a random subsample of 1500 samples from all Washington counties except for Yakima County as well as for the 614D (500 sequences) and 614G (1000 sequences) lineages separately. In addition, we performed the same analysis using 750 sequences from Yakima County only. After an initial introduction of SARS-CoV-2 (*2*), the number of cases grew rapidly (Fig. 2A). As expected, growth in confirmed cases was mirrored in phylodynamic estimates of viral effective population size (Fig. 2A). In addition, we observed maximal transmission intensity at the end of February 2020 when R_e was between 2 and 3 (Fig. 2B). This is consistent with other estimates

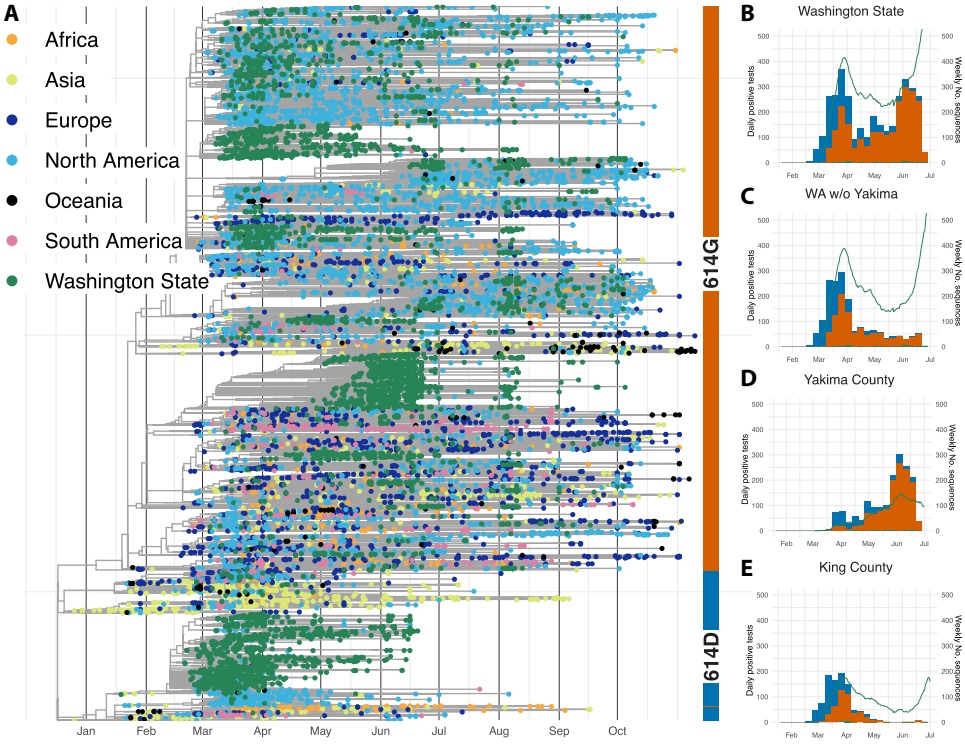

Fig. 1. SARS-CoV-2 phylogeny highlighting D614G split and cases through time in Washington State. (**A**) Phylogenetic tree of 13,900 sequences from Washington State and around the world. Tips are colored on the basis of sampling location. This is a time-calibrated phylogeny with time shown in the *x* axis. The split between 614D sequences (blue) and 614G (orange) sequences is shown as a bar to the right of the phylogeny. (**B** to **E**) Confirmed cases and genetic makeup of SARS-CoV-2 across Washington State and individual counties. The green line shows a 7-day moving average of daily confirmed cases. The bar plots show weekly sequenced cases in our dataset. Cases due to the 614D variant are shown in blue, and cases due to the 614G variant are shown in orange. w/o, without.

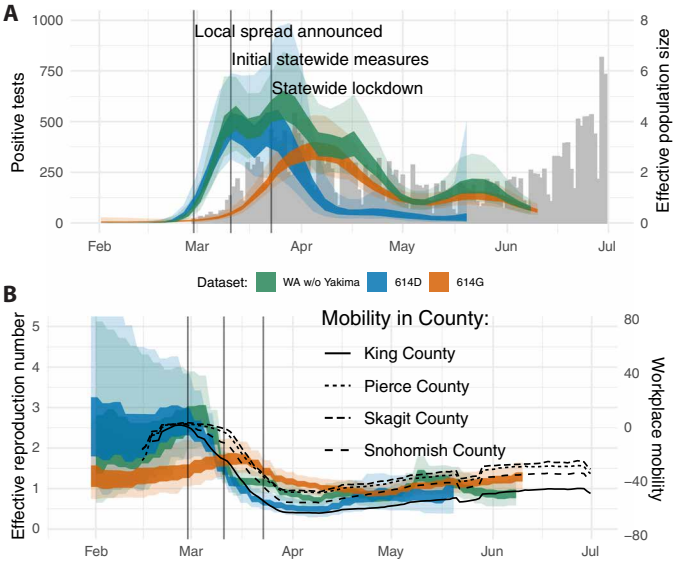

Fig. 2. Regional dynamics of SARS-CoV-2 in Washington State inferred from confirmed cases and pathogen genomes. (**A**) Estimates of effective population sizes for the outbreak in Washington State (green interval) as well as for 614D (blue interval) and 614G (orange interval) individually as compared to confirmed cases in the state (gray bars). The inner band denotes the 50% highest posterior density (HPD) interval and the outer band denotes the 95% HPD interval. (**B**) R_e estimates using a birth-death approach for the same groups as in (A). The R_e estimates are compared to Google workplace mobility data for King, Pierce, Skagit, and Snohomish Counties shown as black solid and dashed lines. Workplace mobility is represented as a 7-day moving average.

of the effective reproduction number of SARS-CoV-2 during early phases of an epidemic when control measures are not in place (*17*–*19*).

Around the time when community spread in King County was announced on 29 February 2020, we observed decreased occupancy of workplaces according to Google mobility data (fig. S2) (*20*). This reduction in workplace mobility occurred relatively early in King County compared to other regions of the state that had little or no reported cases at the time (fig. S2). This is consistent with several businesses starting to institute measures, such as work-from-home policies, at the beginning of March (*10*). This reduction in mobility in King County coincided with a reduction in the effective reproduction number of 614D cases in the state (Fig. 2B). By the time initial statewide measures were implemented on the 11th of March, cases of 614D had almost peaked and were starting to decline, whereas overall cases were about constant or still increasing (Fig. 2A).

Cases of 614G were still increasing and peaked a little over a week later than cases of 614D (Figs. 1 and 2A). This was around the time when the statewide lockdown order came into effect on 24 March 2020. Whereas cases of 614D were initially mostly located around Seattle, cases of 614G were more widespread throughout the state. Viruses sampled from cases in Pierce County and in the counties north of King County mostly harbored the 614G variant (Fig. 1C). Changes in the effective reproduction number of 614G coincided with changes in mobility outside of King County (Fig. 2B). An alternative phylodynamic method using a coalescent approach yielded highly similar results (fig. S3).

Yakima County was the other county in the state besides King County with a large number of 614D cases later in the epidemic (Fig. 1D). The outbreak there happened later than the first large outbreak in King and neighboring counties. In addition, the trend in cases in Yakima County became increasingly decoupled from workplace mobility as measured by cellphone movement for reasons likely associated with a large population of essential workers in the agricultural sector and seasonal worker migration poorly captured in mobility metrics (fig. S4) (*21*, *22*).

To test whether amino acid substitutions beyond D614G affected the chance of SARS-CoV-2 of spreading locally, we next tested whether introductions of lineages with more amino acid substitutions were more successful in spreading locally. We computed the number of amino acid and nucleotide substitutions of the first sampled sequence of each local transmission cluster relative to Wuhan/Hu-1/2019 (*23*). We then estimated whether there was a relationship between the number of amino acid and nucleotide substitutions when a lineage was introduced into Washington State and whether that introduction was successful, which we defined as having led to detectable local transmission. Consistent with a previous publication (*24*), we did not find any substantial relationship between the number of amino acid substitutions and the success of an introduction (fig. S5).

Introductions of SARS-CoV-2 cases from different countries or different areas within a country have repeatedly been discussed as drivers of local outbreaks, particularly in the context of travel bans. We therefore investigated the importance of introductions in driving the outbreak in Washington State. We estimated the relative contribution of introductions compared to local transmission following the coalescent approach introduced above. In short, we used the estimated changes in effective population sizes over time and the estimated rates of introduction to compute the percentage of new cases in the state due to introductions (see Materials and Methods for details).

We estimated the percentage of new cases due to introductions in Washington State (excluding Yakima County) to be below 10% initially and to then have increased to about 10% by the middle of March through early April (Fig. 3). As a reference, the United States instituted a travel ban for nonresidents coming from China on 2 February 2020 and a travel ban from Europe effective 16 March 2020. Increases in the proportion of introductions of the overall cases can be driven by either a reduction in the local transmission rate or an increase in the rate of introduction.

The observed introductions were unevenly distributed across the different clades 614D and 614G (Fig. 3) (*6*, *25*). The proportion of introduced 614G cases was substantially greater than the proportion of introduced 614D cases. We estimated the percentage of introduced 614D cases to be below 3% during the whole outbreak. On the other hand, we inferred the percentage of introduced 614G cases to have been over 10% until the beginning of April. This means that a substantially higher fraction of 614G cases were caused by introductions than for 614D cases. This is expected, considering that cases of 614G were much more widespread outside of China (Fig. 1A), including in areas with relatively strong travel patterns to Washington State during the epidemic, such as New York State.

We next tested whether the percentage of new cases caused by introductions was reasonable given the number and size distribution of local transmission clusters. We simulated local transmission clusters where 0.1, 1, or 10% of all infections were caused by independent introductions. We found that the observed patterns in transmission cluster size distributions fell between the simulated patterns for 1 and 10% of all infections caused by recent introductions (fig. S6).

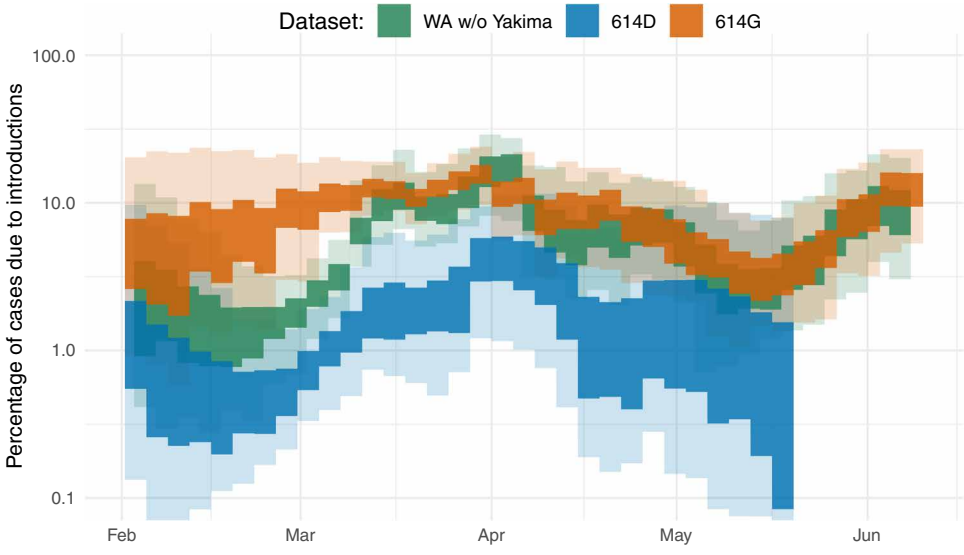

Fig. 3. Phylogenetic estimate of the percentage of introductions of the overall cases. Percentages were estimated as the relative contribution of introductions to the overall number of infections using the multitree coalescent. Percentages are shown for the 2020 outbreak in Washington State (green interval) as well as for 614D (blue interval) and 614G (orange interval). The inner area denotes the 50% HPD interval and the outer area denotes the 95% HPD interval.

Overall, it appears that population-level changes in Washington State in relative frequencies of the two lineages can be explained by differences in timing of measures to curb the spread of SARS-CoV-2 on a county level and by repeated introductions of 614G. Although a parsimonious explanation of observed dynamics, this does not preclude 614G having a higher transmission rate relative to 614D. In addition, these population-level trends are affected by many confounding factors that are not directly related to the virus itself. We therefore next moved to investigate whether we could observe differences between individuals infected with viruses from either lineage on an individual level.

D614G leads to higher viral load, without apparent effects on virulence

We tested for differences in viral loads between patients infected with either the 614D or the 614G viral variants by comparing cycle threshold (Ct) values. Ct values are inversely correlated with viral load, and differences in Ct values between these two variants have been reported previously (*4*, *7*). We analyzed 1743 sequenced SARS-CoV-2 samples from Washington State for which we had access to Ct values. We only used genomes sampled between February and April 2020, when both lineages were circulating in Washington State.

Of these 1743 genomes, 1128 genomes were from patients referred by a health care provider for nasopharyngeal swab testing to the University of Washington (UW) Virology laboratory. A total of 523 genomes were from samples collected by the Washington Department of Health (WA DOH), and 92 samples were from self-collected mid-turbinate nasal swabs mailed in for testing as part of the Seattle Coronavirus Assessment Network (SCAN). During this time period, UW Virology used multiple platforms for polymerase chain reaction (PCR) testing (fig. S7A). Because it is difficult to compare Ct values across primer sets and platforms (*26*), we mainly focused on samples amplified with the most common primer set: N1 and N2 ($n = 879$), although analyses using ORF1ab primers ($n = 229$) were also conducted.

We found that patients infected with viruses with the 614G substitution had lower Ct values (higher viral load) than those infected with 614D viruses in all three collection channels (Fig. 4A and fig. S8). This difference was significant by Wilcoxon rank sum test in samples from UW Virology (N1 and N2 primers: median $\Delta = 1.5$ cycles, $P = 1.5 \times 10^{-12}$; ORF1ab primers: median $\Delta = 2.5$ cycles, $P = 0.0012$) and WA DOH (median $\Delta = 1.4$ cycles, $P = 0.046$), but not in SCAN samples, where we had far fewer samples (median $\Delta = 2.1$ cycles, $P = 0.077$) (Fig. 4A and fig. S8).

We next tested whether factors other than the D614G variant predicted Ct values. We applied a generalized linear model (GLM) assuming normally distributed Ct values to the UW Virology and SCAN samples using variant, patient age, and days after symptom onset as potential predictors of Ct values given that we, like others, have found Ct to be positively correlated with time since symptom onset (fig. S9A) (*27*–*30*). We found that the D614G variant and days since symptom onset were significant ($P = 1.9 \times 10^{-7}$) predictors of Ct values. Variant 614G has a Ct value that is, on average, 1.6 cycles lower than the 614D variant (N1 and N2 primers) (Fig. 4B) when controlling for age and time since symptom onset. This difference in Ct translated to a 0.47 $\log_{10}$ increase in viral load [95% confidence interval (CI): 0.29 to 0.64 $\log_{10}$], assuming the standard curve is linear in this region. For each day after symptom onset, Ct value was predicted to increase by 0.2 cycles (N1 and N2 primers: $P = 3.3 \times 10^{-7}$), which is consistent with other work on Ct values and infection time course (*27*–*30*). In SCAN samples, we observed similar coefficients and significance in the GLM (fig. S8). With ORF1ab primers, D61G variant was not a significant predictor; however, the residuals were not normally distributed, suggesting the model fit poorly with ORF1ab primers (fig. S8).

We additionally looked for a difference in time of symptom onset and sampling date between the two variants—sampling date could be a confounding variable because the relative abundance of the 614G variant increased over time (Fig. 1B)—but did not find any (fig. S9B). Because Ct values were shown to vary with effective reproduction numbers (*31*), we tested whether Ct values changed over time after accounting for the two spike variants. There were also no clear differences in Ct across time when accounting for the spike variant (fig. S9C).

We next tested whether substitutions other than spike D614G contributed to observed Ct differences. First, we considered the genetic diversity defined by five viral clades using the Nextstrain nomenclature: 19A, 19B, 20A, 20B, and 20C (fig. S10A). Clades 19B and 20C differed significantly in their Ct values from the other clades (mean $\Delta = 1.5$ cycles, P adjusted $\leq 2 \times 10^{-8}$ Tukey's range test) (fig. S10B). However, when controlling for the 614G variant, clade membership was not predictive of Ct (fig. S10C). Most samples with available Ct fell into clades 19B and 20C, which primarily contained 614D and 614G variants, respectively, so there may not have been

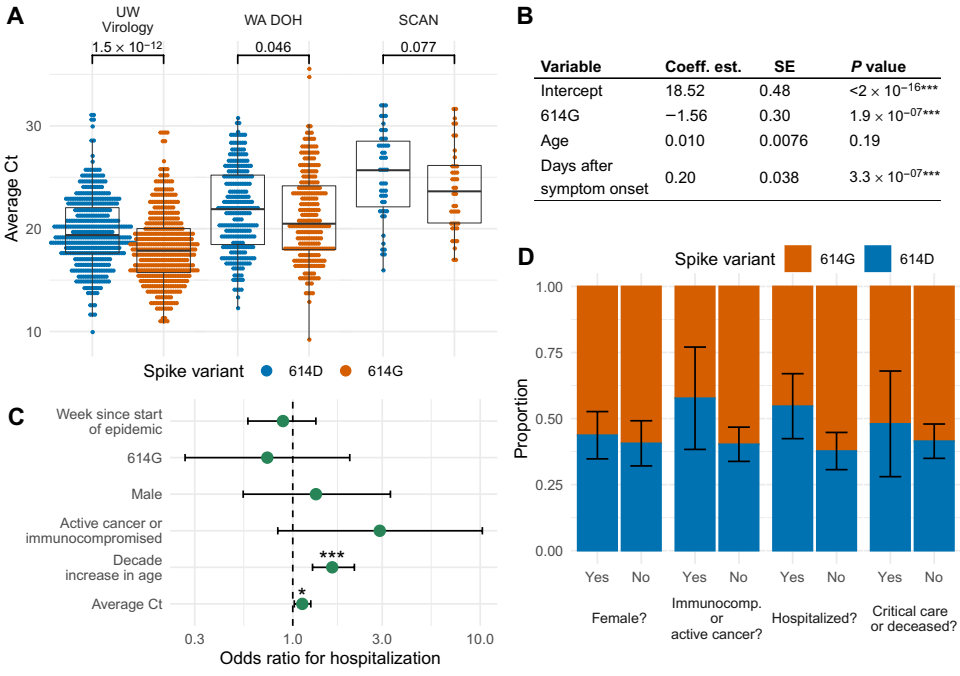

Fig. 4. Factors affecting viral load and disease severity at an individual level. (**A**) Comparison between cycle threshold (Ct) values for viruses with 614G and 614D variants. (**B**) GLM analysis of Ct values using spike variant, age, and days since symptom onset as predictors. (**C**) Odds ratio of being hospitalized given infection with SARS-CoV-2. Error bars show 95% CI, corrected for multiple hypothesis testing using a Bonferroni correction. (**D**) Estimates of the average chance that a patient from a given group was infected with a virus from the 614D clade. The error bars denote the SE of the average chance that a patient from a group was infected with a virus from the 614D clade.

enough genetic diversity in our dataset to identify Ct differences with respect to the other viral clades.

Next, we explored the relationship between the number of amino acid substitutions different from Wuhan/Hu-1/2019 and Ct value. We did not find a significant correlation between amino acid substitutions and Ct with either of variants (614D: Pearson's = −0.052, $P = 0.14$; 614G: Pearson's = 0.061, $P = 0.066$) (fig. S11, A and B). However, a GLM of 614D variant samples predicted a 0.42 decrease in cycle threshold with each additional amino acid substitution ($P = 0.011$) (fig. S11C). In the same GLM applied to 614G variants, amino acid substitutions were not predictive of Ct ($P = 0.66$) (fig. S11D). Within the 614D variant, there was not a specific protein in which additional amino acid substitutions affected Ct values. This might suggest that within our dataset, 614G variants are at a local fitness maxima, whereas 614D variants are not. Thus, there could be more opportunity for amino acid substitutions in 614D variants to affect viral load. We may, however, miss some potentially confounding predictors in this analysis, such as age or mutations in a primer binding region, which could inflate the confidence in the results.

We also found a difference in the age of people infected between the two lineages (fig. S12). In samples from UW Virology, the average age of patients infected with viruses from the 614D and 614G lineages was 56.6 and 52.4, respectively ($P = 5.8 \times 10^{-4}$, Student's t test). In SCAN samples, the average age of patients was 45.8 for 614D and 38.4 for 614G ($P = 0.088$). Age differences may be caused by increased testing, resulting in detection of less severe, younger cases later in the epidemic when 614G was more prevalent. However, we tested this hypothesis in a GLM with week of sample collection and D614G variant as potential predictors of age. Individuals with 614G variant were 3.5 years younger on average ($P = 0.0098$), whereas sample week was not a significant predictor of age ($P = 0.20$) (fig. S12). A skew toward younger individuals is consistent either with a more transmissible virus or with more severe infection as this would result in a larger fraction of younger patients seeking testing. However, the absolute difference in age of infection was still small.

We had access to additional clinical information for 248 of the 1128 sequences from patients referred for SARS-CoV-2 testing by a health care provider. One hundred four of these patients were infected with viruses from the 614D clade, and 144 patients were infected with viruses from the 614G clade. We used data from electronic health records to examine whether differences in Ct values held after correcting for additional potentially confounding factors. We performed the same GLM analysis as above but omitted days since symptom onset as it was missing from most samples. We included additional potential predictors, such as sex, active cancer or immunocompromised status, hospitalization, and whether a patient required intensive care or died. We again found the D614G variant to be significantly associated with Ct values (N1 and N2 primers, $n = 184$, $P = 0.03$). Sex was also a significant predictor of Ct with male individuals having Ct values 1.09 units lower than female individuals (SE = 0.48, $P = 0.02$). None of the other predictors were found to be significant in predicting Ct values, which might be driven by small sample size (table S1). With ORF1ab primers, the D614G variant was not significantly associated with Ct values nor were residuals normally distributed ($n = 63$) (table S2). ORF1ab primers were used later in the epidemic when the 614D variant was less abundant (fig. S7B).

We next investigated which factors associated with clinical outcome. We grouped cases into inpatient (hospitalized) or outpatient (not hospitalized) and then performed a logistic regression with inpatient or outpatient as potential outcomes. As factors predicting outcome, we considered clade membership, sex, immunocompromised/active cancer, age, week of testing, and measured Ct value. Age ($P = 3.2 \times 10^{-6}$) and measured Ct value ($P = 0.012$) were significant predictors for hospitalization after Bonferroni correction for multiple hypothesis testing. Whether a patient was suffering from active cancer or was immunocompromised had an estimated odds ratio of 2.9 (0.8 to 10.8) but was not significant ($P = 0.14$). We did not find any evidence that D614G variant affected clinical outcome (Fig. 4C). This is consistent with neither variant being significantly enriched among males, immunocompromised/active cancer patients, hospitalized patients, or patients who required intensive care or succumbed due to COVID-19 (Fig. 4D).

DISCUSSION

The COVID-19 pandemic has greatly affected lives around the world. As a virus that just recently made the jump into humans, understanding its transmission dynamics and the drivers of its spread is of utmost importance. The emergence of more transmissible strains of SARS-CoV-2 based on an increase in relative frequencies over time has been suggested previously (*25*).

Consistent with trends from other locations around the world (*4*), we found that cases of the spike 614D variant initially dominated in Washington State but were later taken over by spike 614G. However, the trends for 614G and 614D cases that we observed in Washington State appeared to be explained by differences in when action to curb the spread of SARS-CoV-2 were taken on a county level. The trends in effective reproduction numbers between the two clades 614G and 614D coincided with the different trends in mobility of King County (which includes Seattle) and other areas that experienced substantial spread of SARS-CoV-2. The observed patterns are consistent with initial spread of the 614D clade being largely concentrated in King County, which was then mitigated early on. Spread of 614G on the other hand, although present in King County, dominated in other areas of the state and the reduction in the R_e of this variant coincided with a reduction in mobility in these areas, which happened about 9 days after King County. The spread of SARS-CoV-2 in Yakima County, however, seems to be poorly captured by mobility trends.

We additionally inferred introductions play a larger role in driving cases of the 614D variant than of the 614G variant. This suggests that differences in the relative frequencies of the two variants are at least, in part, driven by differences in when and where lineages were introduced into the state. Overall, we find that we can explain the changes in relative frequency of the 614D and 614G variants over time by nonviral factors in absence of intrinsic transmission rate differences. This does, however, not exclude the possibility that such differences exist and have led to the replacement of 614D by 614G in other parts of the world. The observation that changes in patterns of which lineages are introduced into a location can drive changes in local frequencies of a variant is important when evaluating whether new variants (such as B.1.1.7) are more transmissible. In particular, it means that an increase in relative frequency of a new variant in different places does not necessarily provide independent evidence about whether or not the new variant is more transmissible.

We did find evidence for lower Ct values in patients infected with viruses of the 614G variant, suggesting higher viral loads. This holds even after including several additional factors, such as the age of a patient and days since symptom onset, as potential predictors for Ct values. However, we did not find evidence that D614G has an impact on risk of hospitalization although testing policy would bias toward finding a variant with greater virulence as hospitalized patients are overrepresented in the dataset (*32*, *33*). The differences in Ct values translate to an about 0.47 $\log_{10}$ increase in viral load (95% CI: 0.29 to 0.64 $\log_{10}$). This difference might not be large enough to lead to large differences in severity or transmissibility that can be observed in a dataset of this size.

Our findings are broadly consistent with other analyses on the spike D614G substitution. A previous study found evidence of lowered Ct but limited clinical difference for viruses of the 614G clade in Sheffield, UK (*4*). Recent in vitro studies showed that pseudovirus containing spike protein with a 614G substitution exhibits greater infectivity (*5*, *8*, *9*). Other work suggests the increased transmissibility of 614G over 614D in an analysis of thousands of sequences from the United Kingdom (*6*).

Although our results are broadly consistent with other analyses, they are not without limitations. First, the sample collection is likely biased toward more symptomatic cases. In addition, the collection of SARS-CoV-2 samples was limited initially and improved during the study period and likely differed across different geographic areas. In other words, the sampling regime likely differed across space and time, potentially affecting the results.

The phylogenetic analyses conditioned on specific clustering of sequences in Washington State by incorporating background sequences from other locations. Differences in sampling and sequencing regimes in potential source locations of SARS-CoV-2 relative to Washington State could bias this clustering, which, in turn, could affect the estimated rates of introductions into Washington State and potentially also the effective reproduction numbers over time. Last, the phylodynamic methods used here make a few simplifying assumptions about how SARS-CoV-2 is spread, such as random sampling of infected individuals, homogeneous mixing of individuals, or the absence of superspreading. Although we addressed the latter in our simulation study, it is not fully clear how some of these simplifying assumptions affected the inference results.

Overall, we found evidence for higher viral loads in individuals with viruses from the 614G clade, which theoretically could affect transmissibility and severity. However, we did not see strong evidence that this degree of difference in Ct manifested in substantial differences in transmissibility or severity of infection with SARS-CoV-2 in the spring/summer 2020 Washington State epidemic.

MATERIALS AND METHODS

Study design

The aim of this study was to characterize the drivers of the SARS-CoV-2 outbreak in Washington State (USA) over several months and to investigate how different viral variants affected the spread of SARS-CoV-2. We collected genetic sequence data from SARS-COV-2 viruses isolated in Washington State. Here, we analyzed 3940 SARS-CoV-2 genomes sequenced from samples collected in Washington State between February and July 2020 as our primary dataset. These sequences were produced as part of an effort to survey the spread and evolution of SARS-CoV-2 in the state. These samples were pooled from three different channels: UW Virology, WA DOH, and SCAN, as described below.

Sequencing and analysis of samples from the Seattle Flu Study was approved by the Institutional Review Board (IRB) at the University of Washington (protocol STUDY00006181). Informed consent was obtained for all community participant samples and survey data. Informed consent for residual sample and clinical data collection was waived. Sequencing and analysis of samples from SCAN was approved by the IRB at the University of Washington (protocol STUDY00010432). Informed consent was obtained for all community participant samples and survey data. For UW Virology Lab, use of residual clinical specimens was approved by the IRB at the University of Washington (protocol STUDY00000408) with a waiver of informed consent.

Sample collection and testing for SARS-CoV-2

For the 1236 UW Virology samples, nasopharyngeal/oropharyngeal swabs were obtained as part of clinical testing for SARS-CoV-2 ordered by local health care providers or collected at drive-up testing

sites. RNA was extracted and the presence of SARS-CoV-2 was detected by reverse transcription–PCR (RT-PCR) as previously described using either the emergency use-authorized UW Centers for Disease Control (CDC)–based laboratory-developed test, Hologic Panther Fusion test, or Roche cobas SARS-CoV-2 test (*34*).

For the 2601 WA DOH samples, nasopharyngeal/oropharyngeal/bronchoalveolar/sputum samples were obtained for SARS-CoV-2 clinical testing, as requested by submitting health care entities. RNA was extracted and the presence of SARS-CoV-2 was detected via either the CDC 2019-nCoV RT-PCR Diagnostic Panel or the Applied Biosystems TaqPath COVID-19 Combo Kit.

For the 103 SCAN samples, specimens were shipped to the Brotman Baty Institute for Precision Medicine via commercial couriers or the U.S. Postal Service at ambient temperatures and opened in a class II biological safety cabinet in a biosafety level-2 laboratory. Two or three 650-μl aliquots of Universal Transport Media were collected from each specimen and stored at 4°C until the time of nucleic acid extraction, performed with a MagNA Pure 96 small volume total nucleic acid kit (Roche). SARS-CoV-2 detection was performed using real-time RT-PCR with a probe set targeting Orf1b and S with FAM fluor (Life Technologies 4332079, assay nos. APGZJKF and APXGVC4APX) multiplexed with an RNaseP probe set with VIC or HEX fluor (Life Technologies A30064 or IDT custom) each in duplicate on a QuantStudio 6 instrument (Applied Biosystems).

Viral sequencing and genome assembly

For UW Virology samples, sequencing was attempted on all specimens with Ct < 32 either using a metagenomic approach described previously (*2*, *35*), via oligonucleotide probe-capture (*36*), or using an amplicon sequencing–based approach (*37*). Libraries were sequenced on Illumina MiSeq or NextSeq instruments using 1 × 185 or 1 × 75 runs, respectively. Consensus sequences were assembled using a custom bioinformatics pipeline (https://doi.org/10.5281/zenodo.4701603) that combines de novo assembly and read mapping to generate a per-sample consensus sequence. Consensus sequences were deposited to GenBank and GISAID and raw reads to SRA under Bioproject PRJNA610428.

For samples from WA DOH and SCAN, sequencing was attempted on all specimens with Ct < 30 using a hybrid-capture approach. RNA was fragmented and converted to cDNA using random hexamers and reverse transcriptase (SuperScript IV, Thermo Fisher Scientific) and a sequencing library was constructed using an Illumina TruSeq RNA Library Prep for Enrichment kit. Using Ct value as a proxy for viral load, samples were balanced and pooled in 24-plex for the hybrid capture reaction. Capture pools were incubated overnight with probes targeting the Wuhan-Hu-1 isolate, synthesized by Twist Bioscience. The manufacturer's protocol was followed for the hybrid capture reaction and target enrichment washes. Final pools were sequenced on the Illumina NextSeq or NovaSeq instrument using 2 × 150–base pair reads. The resulting reads were assembled against the SARS-CoV-2 reference genome Wuhan/Hu-1/2019 (GenBank accession MN908947) using the bioinformatics pipeline at (https://doi.org/10.5281/zenodo.4701970). Consensus sequences were deposited to GenBank and GISAID. Samples sequenced by UW Virology had a higher proportion of 614G variants (54.7%) than SCAN and WA DOH samples (48.6%), which were sequenced using a different pipeline (chi-square test: $P = 0.017$). Investigating differences in Ct independently for each primer type should control for differences in the spike variant proportion, as primer types did not overlap between sequencing pipelines.

Clustering

To distinguish between sequences that were connected by local transmission, we clustered all sequences from Washington State together on the basis of their pairwise genetic distance. We first built a timed tree using sequences from Washington State and from around the world using the Nextstrain pipeline (*3*). Overall, we used 4023 sequences from Washington State and 6028 from the rest of the world. Of all sequences, 2601 were from the Washington Department of Health, 1236 were from the UW Virology Lab, and 103 were from SCAN. All other sequences were downloaded from the GISAID EpiCoV database (*38*, *39*).

We then use a parsimony-based approach to reconstruct the locations of internal nodes. We considered all sequences from Washington State as one location and all sequences from anywhere else on the globe to be from another location. We then reconstructed the internal node locations using the Fitch parsimony algorithm. We considered each group of sequences to be on the same local transmission cluster if all their common ancestor nodes are inferred to be in Washington State. We additionally tested the sensitivity of this approach to having less background samples by randomly removing sequences from outside of Washington State and computing the number of clusters again. Although we do expect that including more background sequences would increase the number of clusters detected, we did not find a large impact on the number of background sequences on the number of clusters identified or the average size of clusters identified (fig. S13).

Estimating population dynamics jointly from multiple local outbreak clusters

To estimate the population dynamics of the Washington State outbreak, we used a coalescent approach to infer these dynamics jointly from all known local outbreak clusters. We modeled the coalescence and migration of lineages within Washington State as a structured coalescent process with known migration history. Under this model, lineages can coalesce within the sampled subpopulation and have originated from outside the sampled subpopulation. We a priori assumed that we know where on the tree lineages were introduced into the sampled subpopulation (fig. S14). This known migration history is given by the clustering of sequences into local outbreak clusters. The migration events from anywhere outside WA into WA were always assumed to have happened before the common ancestor of all sequences in each local outbreak cluster. How long before this common ancestor time was inferred during the Markov chain Monte Carlo (MCMC) run. The rate at which we expect coalescent events to occur is exponentially distributed with mean = $n \times (n - 1)/2N_e$, and the rate at which we expect to observe introductions events is exponentially distributed with mean $n \times m$, with n being the number of lineages in any given local transmission cluster that coexist at a point in time and m being the rate of introduction. Everything that happened outside the sampled subpopulation was ignored, or in other words, we ignored how exactly the individual local outbreak clusters related to each other.

We then inferred the effective population size and rates of introduction through time using a skyline approach. Effective population sizes and rates of introduction were allowed to change at predefined time points. The rates were interpolated between these predefined time points where the rates are estimated. This is equivalent to assuming exponential growth or decline between the effective population sizes at these time points.

We then used two different ways to account for correlations between adjacent scaled effective population sizes ($N_e\tau$). First, we used the classic skyride (*14*) approach, where we assumed that the logarithm of adjacent $N_e\tau$ is normally distributed with mean 0 and an estimated σ. In addition, we used an approach where we assumed that differences in growth rates are normally distributed with mean 0 and an estimated σ (*40*). This is equivalent to using an exponential coalescent model with time varying growth rates. We implemented this multitree coalescent approach as an extension to the Bayesian phylogenetics software BEAST2 (*41*). The code for the multitree coalescent is available here (https://doi.org/10.5281/zenodo.4697903) and is validated in fig. S3. We allowed the effective population sizes to change every 3.5 days and the rates of introduction to change every 7 days. The inference of the effective population sizes and rates of introductions was performed using an adaptive multivariate Gaussian operator (*42*) implemented at (https://doi.org/10.5281/zenodo.4705996), and the analyses were run using adaptive Metropolis-coupled MCMC (*43*).

In contrast to backward-in-time coalescent approaches, we can consider different local outbreak clusters as independent observations of the same underlying population process using birth-death models. We inferred the effective reproduction number using the birth-death skyline model (*12*) by assuming that the different local outbreak clusters are independent observations of the same process with the same parameters (*13*). We allowed the effective reproduction number to change every 3.5 days. As for the coalescent approach, we assumed adjacent effective reproduction numbers to be normally distributed in log space with mean 0 and an estimated σ. We further assumed the becoming uninfectious rate to be 52.3 per year, which corresponds to an average duration of infectivity of 7 days (*44*). We allowed the probability of an individual to be sampled and sequenced upon recovery to change every 7 days.

Simulation study

To test our implementation of the multitree coalescent, we performed two different sets of simulation studies. In the first simulation study, we simulated 10 phylogenetic trees under the structured coalescent using 1000 samples from the same location in MASTER (*45*). For each of the 10 simulations, we randomly sampled the N_e at time 0 from a normal distribution with mean = 0 and σ = 0.5 and then randomly drew the N_e at subsequent time points $t + 1$ randomly from a normal distribution with mean = $N_e(t)$ and σ = 0.5. This is equivalent to randomly sampling N_e trajectories under a skygrid distribution (*14*). We performed the same for the rate of introductions at different points in time. We then simulated a single phylogenetic tree under the structured coalescent using these randomly sampled parameters. Next, we splitted this tree into several local transmission clusters and then inferred the N_es and rate of introductions over time from only the local transmission clusters (fig. S15).

In the second simulation study, we simulated 10 phylogenetic trees under a structured infected (I) only model with superspreading. We assumed that there was a constant number of introductions per unit of time from outside into Washington State. After an introduction into the state, each infected individual was transmitting to n other individuals. We assumed the number of newly infected individuals to be negatively binomially distributed such that the mean number of introductions at any point in time t was equal to $R_e(t)$ and the dispersion parameter $k = 1$. We next simulated a structured phylogenetic tree from this approach. We then simulated genetic sequences on top of this phylogenetic tree using Seq-Gen (*46*).

Subsampling of sequences

We analyzed the population dynamics in total for four different datasets. In the first datasets, we randomly subsampled 1500 of the sequences from Washington State, excluding sequences from Yakima County. One thousand five hundred sequences were chosen because of computational limitations of the Bayesian phylodynamic inference. For the second and third datasets, we distinguished between two different clades that we call D and G. The D clade consists of all sequences with an aspartic acid at site 614 of the spike protein, and the G clade consists of all sequences with a glycine at this position (visible at https://nextstrain.org/ncov/global?c=gt-S_614). For the 614D datasets, we used the same subsampling procedure as for the above dataset but with 500 sequences, and 750 sequences for the 614G clade. For the Yakima County dataset, we used 750 randomly subsampled sequences.

Estimating the percentage of overall new cases from independent introductions

We estimated the relative contribution of introductions compared to local transmission using the coalescent approach introduced here. In addition to the regular assumptions of the coalescent approach that all samples are taken at random from a well-mixed population, we assumed that differences in effective population size between adjacent time intervals can be used to compute the transmission rate. We then computed the transmission rate as the sum of the growth rate of the effective population size and the becoming uninfectious rate (that is, we used the relationship $\frac{dNe}{dt}$ = transmission rate – becoming uninfectious rate, to compute the transmission rate). We assumed an average time of infectiousness of 7 days. In addition, we assumed that dN_e/dt is independent from the rate of introduction. We then computed the percentage of introductions in overall cases using the rate of introduction and the transmission rate. The rate of introduction can be expressed as the total number of introductions divided by the number of infected in WA, that is, rate of introduction = No. introductions/No. infected. The total number of new infections locally can be expressed as transmission rate × infected, which, in turn, means that ratio of introductions over local infections can be expressed as (rate of introduction × infected)/(transmission rate × infected). From this ratio, we can then compute the percentage of introductions of the overall cases.

We tested that we can retrieve the percentage of introductions from simulations, where we simulated phylogenetic trees using an infected recovered (IR) compartmental model with superspreading using MASTER (*45*). We then simulated genetic sequence data using those trees and then inferred the percentage of new cases due to introductions from those sequences (figs. S15 and S16).

Chart review

Clinical record review of UW affiliated patients was performed under University of Washington IRB: STUDY00000408. This included patients who visited UW affiliated clinics and patients who were hospitalized at UW Medical Center, both the Montlake and Northwest locations, and Harborview Medical Center. Sex, age, the presence of active cancer or immunosuppresive medication, hospital admission, critical care admission, and deceased status were extracted from all charts.

Statistical analysis

Factors affecting Ct and clinical outcomes of individuals

R/3.6.2 was used for Ct and clinical record analysis. The code and data cleaned of all patient identifiers is available at (https://doi.org/10.5281/zenodo.4701583).

UW Virology used three different primer sets and platforms over the timeframe of the dataset (fig. S7). Because it is difficult to compare Ct across primer sets, we ran both tests comparing Ct by viral clade and the generalized linear model predicting Ct separately for N1 and N2, and ORF1ab primers. There were insufficient samples amplified with Egene/RdRp primers for statistical analysis ($n = 20$).

We chose to use Wilcoxon rank sum test to compare differences in Ct between viral lineages and Student's t test to compare differences in age between viral lineages. Age was reported as a decade bin converted into a numerical equivalent, and Wilcoxon rank sum test underestimates differences with duplicate numbers. Tukey's range test was used to identify differences in Ct between viral clades, and we used Pearson's correlation coefficient to examine the relationship between Ct and number of amino acid and synonymous substitutions. P values less than 0.05 were considered significant. Data were plotted as a univariate histogram to check for normal distribution before testing with Tukey's range test and Student's t test.

For GLMs predicting Ct and age, we used a multivariate linear regression of form

$$y_i = \beta_0 + \Sigma\, \beta_j x_{i,j} + \epsilon_i$$

where y is the dependent variable (either Ct or age), β is the coefficient of the predictor variable, x is the predictor variable, and ϵ is the residual error. Models were run with the glm package in R.

UW Virology and SCAN samples were used to estimate predictors of Ct as age was not available for WA DOH samples. The predictor variables were the amino acid at Spike 614 (binary variable), days since symptom onset (continuous variable), and age of patient (continuous variable). In the GLM of Ct with only samples from UW Medicine affiliates, we excluded days since symptom onset as it was not available for most samples. We additionally included sex (binary variable), active cancer or immunocompromised (binary variable), hospitalized (binary variable), and required critical care or deceased (binary variable) as predictors of Ct. When considering viral clade as a predictor of Ct, we applied the same GLM as above with addition of binary variables for clade 19A, 20A, and 20C. Clades 20B and 19B were excluded due to collinearity.

To test the relationship between number of substitutions (synonymous and amino acid) and Ct, we applied a GLM predicting Ct from amino acid substitutions (continuous variable), synonymous substitutions (continuous variable), days since symptom onset (continuous variable, week since start of the Washington State epidemic (continuous variable), and binary variables for ORF1ab, WA DOH, and SCAN primers. We ran the GLM separately spike 614D and 614G variants as the correlation between the number of amino acid substitutions and Ct differed between variants. In the GLM, we excluded samples with greater than 20 nucleotide substitutions as outliers, because all other samples had between 3 and 17 nucleotide substitutions.

To estimate predictors of patient age, we used all SCAN & UW Virology samples with age available ($n = 1172$). The predictor variables were amino acid at spike 614 (binary variable) and week since community spread of COVID-19 was reported in Washington (continuous variable).

To estimate predictors of hospitalization if infected with SARS-CoV-2, we used a multivariate logistic regression

$$\text{logit}(P_i) = \beta_0 + \Sigma\, \beta_j x_{i,j} + \epsilon_i$$

where P is the probability of hospitalization, β is the coefficient of the predictor variable, x is the predictor variable, and ϵ is the residual error. Predictor variables were week since first sample in dataset (continuous variable), sex (binary variable), active cancer or immunocompromised (binary variable), age in decade (continuous variable), amino acid at Spike 614 (binary variable), and average Ct (continuous variable). To fit the logistic regression, we again used the glm package in R, specifying family as "binomial". P values and CIs for risk of hospitalization were adjusted for multiple hypothesis testing using a Bonferroni correction.

Chi-square tests were used to compare proportions of viral lineages by sex, immunocompromised status, clinical outcome (inpatient or outpatient), and severe outcome (critical care or death). P values were adjusted for multiple hypothesis testing using the Bonferroni correction.

SUPPLEMENTARY MATERIALS

stm.sciencemag.org/cgi/content/full/13/595/eabf0202/DC1

Fig. S1. Number of lineages through time for different local transmission clusters.
Fig. S2. Workplace mobility trends of different counties in Washington State compared to King County.
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.
Fig. S5. Substitutions and success of a SARS-CoV-2 introduction.
Fig. S6. Probability that a newly sampled case reveals a new introduction.
Fig. S7. Histogram of primers used by UW Virology across time.
Fig. S8. Comparison of Ct across SARS-CoV-2 Spike variant.
Fig. S9. Symptom and Ct values across time.
Fig. S10. Comparing Ct by viral clade.
Fig. S11. Cycle threshold by number of substitutions.
Fig. S12. Age of infected individuals by 614D or 614G variant over time.
Fig. S13. Dependence of the local outbreak clusters and the number of background sequences used.
Fig. S14. Principle of the multitree coalescent.
Fig. S15. Estimation of effective population sizes and rates of introductions from simulations.
Fig. S16. Estimation of the percentage of new cases due to introductions from simulations.
Table S1. GLM of Ct with N1, N2 primers in patients at UW affiliates.
Table S2. GLM of Ct with ORF1ab primers in patients at UW affiliates.
Data file S1. GISAID acknowledgment table (tsv).
Data file S2. Cycle threshold values for isolates (tsv).

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. A. Rambaut, Phylogenetic analysis of nCoV-2019 genomes. *Virological* (available at http://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356).
2. T. Bedford, A. L. Greninger, P. Roychoudhury, L. M. Starita, M. Famulare, M.-L. Huang, A. Nalla, G. Pepper, A. Reinhardt, H. Xie, L. Shrestha, T. N. Nguyen, A. Adler, E. Brandstetter, S. Cho, D. Giroux, P. D. Han, K. Fay, C. D. Frazar, M. Ilcisin, K. Lacombe, J. Lee, A. Kiavand, M. Richardson, T. R. Sibley, M. Truong, C. R. Wolf, D. A. Nickerson, M. J. Rieder, J. A. Englund, J. Hadfield, E. B. Hodcroft, J. Huddleston, L. H. Moncla, N. F. Müller, R. A. Neher, X. Deng, W. Gu, S. Federman, C. Chiu, J. Duchin, R. Gautom, G. Melly, B. Hiatt, P. Dykema, S. Lindquist, K. Queen, Y. Tao, A. Uehara, S. Tong, D. MacCannell, G. L. Armstrong, G. S. Baird, H. Y. Chu, J. Shendure, K. R. Jerome, Cryptic transmission of SARS-CoV-2 in Washington State. *Science* **370**, 571–575 (2020).
3. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R. A. Neher, Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
4. B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge,

C. M. Evans, T. M. Freeman, T. I. de Silva, C. McDanal, L. G. Perez, H. Tang, A. Moon-Walker, S. P. Whelan, C. C. LaBranche, E. O. Saphire, D. C. Montefiori, A. Angyal, R. L. Brown, L. Carrilero, L. R. Green, D. C. Groves, K. J. Johnson, A. J. Keeley, B. B. Lindsey, P. J. Parsons, M. Raza, S. Rowland-Jones, N. Smith, R. M. Tucker, D. Wang, M. D. Wyles, Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812–827.e19 (2020).

5. L. Yurkovetskiy, K. E. Pascal, C. Tompkins-Tinch, T. Nyalile, Y. Wang, A. Baum, W. E. Diehl, A. Dauphin, C. Carbone, K. Veinotte, S. B. Egri, S. F. Schaffner, J. E. Lemieux, J. Munro, P. C. Sabeti, C. Kyratsous, K. Shen, J. Luban, SARS-CoV-2 spike protein variant D614G increases infectivity and retains sensitivity to antibodies that target the receptor binding domain. *bioRxiv*, 2020.07.04.187757 (2020).
6. E. M. Volz, V. Hill, J. T. McCrone, A. Price, D. Jorgensen, A. O'Toole, J. A. Southgate, R. Johnson, B. Jackson, F. F. Nascimento, S. M. Rey, S. M. Nicholls, R. M. Colquhoun, Ana Da Silva Filipe, N. Pacchiarini, M. Bull, L. Geidelberg, I. Siveroni, I. G. Goodfellow, N. J. Loman, O. Pybus, D. L. Robertson, E. C. Thomson, A. Rambaut, T. R. Connor, *Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity*, *medRxiv* (2020); https://www.medrxiv.org/content/10.1101/2020.07.31.20166082v1.full.pdf+html.
7. R. P. McNamara, C. Caro-Vegas, J. T. Landis, R. Moorad, L. J. Pluta, A. B. Eason, C. Thompson, A. Bailey, F. C. S. Villamor, P. T. Lange, J. P. Wong, T. Seltzer, J. Seltzer, Y. Zhou, W. Vahrson, A. Juarez, J. O. Meyo, T. Calabre, G. Broussard, R. Rivera-Soto, D. L. Chappell, R. S. Baric, B. Damania, M. B. Miller, D. P. Dittmer, High-density amplicon sequencing identifies community spread and ongoing evolution of SARS-CoV-2 in the Southern United States. *Cell Rep.* **33**, 108352 (2020).
8. L. Zhang, C. B. Jackson, H. Mou, A. Ojha, E. S. Rangarajan, T. Izard, M. Farzan, H. Choe, The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *Nat. Commun.* **11**, 6013 (2020).
9. Z. Daniloski, T. X. Jordan, J. K. Ilmain, X. Guo, G. Bhabha, B. R. tenOever, N. E. Sanjana, The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. *eLife* **10**, e65365 (2021).
10. R. Burstein, H. Hu, N. Thakkar, A. Schroeder, M. Famulare, D. Klein, *Understanding the Impact of COVID-19 Policy Change in the Greater Seattle Area using Mobility Data* (Institute for Disease Modeling, 2020; https://covid.idmod.org/data/Understanding_impact_of_COVID_policy_change_Seattle.pdf).
11. M. Worobey, J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, P. Lemey, The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564–570 (2020).
12. T. Stadler, D. Kühnert, S. Bonhoeffer, A. J. Drummond, Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci. U.S.A.* **110**, 228–233 (2013).
13. N. F. Müller, D. Wüthrich, N. Goldman, N. Sailer, C. Saalfrank, M. Brunner, N. Augustin, H. M. B. Seth-Smith, Y. Hollenstein, M. Syedbasha, D. Lang, R. A. Neher, O. Dubuis, M. Naegele, A. Buser, C. H. Nickel, N. Ritz, A. Zeller, B. M. Lang, J. Hadfield, T. Bedford, M. Battegay, R. Schneider-Sliwa, A. Egli, T. Stadler, Characterising the epidemic spread of Influenza A/H3N2 within a city through phylogenetics. *PLOS Pathog.* **16**, e1008984 (2020).
14. M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, M. A. Suchard, Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013).
15. N. Takahata, The coalescent in two partially isolated diffusion populations. *Genet. Res.* **52**, 213–222 (1988).
16. J. Hein, M. Schierup, C. Wiuf, *Gene Genealogies, Variation and Evolution: A primer in coalescent theory* (Oxford Univ. Press, 2004).
17. A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo; Centre for Mathematical Modelling of Infectious Diseases COVID-19 working group, Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *Lancet Infect. Dis.* **20**, 553–558 (2020).
18. R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, J. Shaman, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493 (2020).
19. N. Thakkar, M. Famulare, *COVID-19 transmission was likely rising through April 22 across Washington State* (Institute for Disease Modeling, 2020; https://covid.idmod.org/data/COVID-19-transmission-likely-rising-through-April22-across-Washington-State.pdf).
20. L. L. C. Google, *Google COVID-19 Community Mobility Reports*; https://www.google.com/covid19/mobility/data_documentation.html?hl=en.
21. N. Thakkar, M. Zimmermann, R. Burstein, E. Wenger, M. Famulare, *Comparing COVID-19 dynamics in King and Yakima counties* (Institute for Disease Modeling, 2020; https://covid.idmod.org/data/Comparing_COVID-19_dynamics_in_King_and_Yakima_counties.pdf).
22. D. Chao, M. Zimmermann, *Mobility and phased re-opening in Washington* (Institute for Disease Modeling, 2020; https://covid.idmod.org/data/mobility_and_phased_re-opening_in_washington.pdf).
23. F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
24. L. van Dorp, D. Richard, C. C. S. Tan, L. P. Shaw, M. Acman, F. Balloux, No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat. Commun.* **11**, 5986 (2020).
25. B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, B. Foley, E. E. Giorgi, T. Bhattacharya, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, C. C. LaBranche, D. C. Montefiori; Sheffield COVID-19 Genomics Group, Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv* 2020.04.29.069054 , (2020).
26. D. Rhoads, D. R. Peaper, R. C. She, F. S. Nolte, C. M. Wojewoda, N. W. Anderson, B. S. Pritt, College of American Pathologists (CAP) Microbiology Committee Perspective: Caution must be used in interpreting the cycle threshold (Ct) value. *Clin. Infect. Dis.* , ciaa1199 (2020).
27. F. Yu, L. Yan, N. Wang, S. Yang, L. Wang, Y. Tang, G. Gao, S. Wang, C. Ma, R. Xie, F. Wang, C. Tan, L. Zhu, Y. Guo, F. Zhang, Quantitative detection and viral load analysis of SARS-CoV-2 in infected patients. *Clin. Infect. Dis.* **71**, 793–798 (2020).
28. X. He, E. H. Y. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. C. Lau, J. Y. Wong, Y. Guan, X. Tan, X. Mo, Y. Chen, B. Liao, W. Chen, F. Hu, Q. Zhang, M. Zhong, Y. Wu, L. Zhao, F. Zhang, B. J. Cowling, F. Li, G. M. Leung, Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675 (2020).
29. L. Zou, F. Ruan, M. Huang, L. Liang, H. Huang, Z. Hong, J. Yu, M. Kang, Y. Song, J. Xia, Q. Guo, T. Song, J. He, H.-L. Yen, M. Peiris, J. Wu, SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *N. Engl. J. Med.* **382**, 1177–1179 (2020).
30. Y. Huang, S. Chen, Z. Yang, W. Guan, D. Liu, Z. Lin, Y. Zhang, Z. Xu, X. Liu, Y. Li, SARS-CoV-2 viral load in clinical samples from critically ill patients. *Am. J. Respir. Crit. Care Med.* **201**, 1435–1438 (2020).
31. J. A. Hay, L. Kennedy-Shaffer, S. Kanjilal, M. Lipsitch, M. J. Mina, Estimating epidemiologic dynamics from single cross-sectional viral load distributions. *medRxiv* 2020.10.08.20204222 , (2020).
32. UW Medicine, SARS-CoV-2 Testing Criteria, March 7, 2020 (2020); https://education.uwmedicine.org/wp-content/uploads/2020/03/1-Testing-Criteria.pdf.
33. UW Medicine, SARS-CoV-2 Testing Criteria, August 20, 2020 (2020); https://one.uwmedicine.org/coronavirus/Screening%20and%20Testing%20Algorithms/01a%20-%20Testing%20Criteria.pdf.
34. J. A. Mays, A. L. Greninger, K. R. Jerome, J. B. Lynch, P. C. Mathias, Preprocedural surveillance testing for SARS-CoV-2 in an asymptomatic population in the Seattle region shows low rates of positivity. *J. Clin. Microbiol.* **58**, e01193-20 (2020).
35. A. L. Greninger, D. M. Zerr, X. Qin, A. L. Adler, R. Sampoleo, J. M. Kuypers, J. A. Englund, K. R. Jerome, Rapid metagenomic next-generation sequencing during an investigation of hospital-acquired human parainfluenza virus 3 infections. *J. Clin. Microbiol.* **55**, 177–182 (2017).
36. A. L. Greninger, P. Roychoudhury, H. Xie, A. Casto, A. Cent, G. Pepper, D. M. Koelle, M.-L. Huang, A. Wald, C. Johnston, K. R. Jerome, Ultrasensitive capture of human herpes simplex virus genomes directly from clinical samples reveals extraordinarily limited evolution in cell culture. *mSphere* **3**, e00283-18 (2018).
37. A. Addetia, H. Xie, P. Roychoudhury, L. Shrestha, M. Loprieno, M.-L. Huang, K. R. Jerome, A. L. Greninger, Identification of multiple large deletions in ORF7a resulting in in-frame gene fusions in clinical SARS-CoV-2 isolates. *J. Clin. Virol.* **129**, 104523 (2020).
38. S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33–46 (2017).
39. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.* **22**, 30494 (2017).
40. E. M. Volz, X. Didelot, Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance. *Syst. Biol.* **67**, 719–728 (2018).
41. R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, M. Matschiner, F. K. Mendes, N. F. Müller, H. A. Ogilvie, L. du Plessis, A. Popinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C.-H. Wu, D. Xie, C. Zhang, T. Stadler, A. J. Drummond, BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019).
42. G. Baele, P. Lemey, A. Rambaut, M. A. Suchard, Adaptive MCMC in Bayesian phylogenetics: An application to analyzing partitioned data in BEAST. *Bioinformatics* **33**, 1798–1805 (2017).
43. N. F. Müller, R. R. Bouckaert, Adaptive Metropolis-coupled MCMC for BEAST 2. *PeerJ* **8**, e9473 (2020).
44. L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, C. Fraser, Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368**, eabb6936 (2020).
45. T. G. Vaughan, A. J. Drummond, A stochastic simulator of birth-death master equations with application to phylodynamics. *Mol. Biol. Evol.* **30**, 1480–1493 (2013).

46. A. Rambaut, N. C. Grassly, Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* **13**, 235–238 (1997).
47. G. Yu, Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinformatics* **69**, e96 (2020).
48. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer Science & Business Media, 2009).

Acknowledgments: We would like to thank the anonymous reviewers for helpful feedback. In addition, we would like to thank T. Stadler and T. Vaughan for comments on the phylodynamic analyses. We also thank K. Allen for providing symptom onset dates. We acknowledge the authors for originating and submitting laboratories of the sequences from GISAID's EpiFlu database, on which this research is based. A full Acknowledgments table is available in the Supplementary Materials. We have tried our best to avoid any direct analysis of genomic data not submitted as part of this paper and use this genomic data as background. **Funding:** N.F.M. is funded by the Swiss National Science Foundation (P2EZP3_191891). C.W. is funded by Achievement Rewards for College Scientists. J.S. is an investigator of the Howard Hughes Medical Institute. T.B. is a Pew Biomedical Scholar and is supported by NIH R35 GM119774-01. The Seattle Flu Study is run through the Brotman Baty Institute for Precision Medicine and funded by Gates Ventures, the private office of Bill Gates. The Scientific Computing Infrastructure at Fred Hutch is supported by NIH ORIP S10OD028685. P.R. is a CFAR New Investigator award recipient supported by NIH AI027757. **Author contributions:** N.F.M. designed and performed phylodynamic analyses. C.W. led individual-level analysis of viral variants. C.D.F. led viral sequencing of SCAN and WA DOH samples. P.R. led viral sequencing of UW Virology samples. J.L., L.H.M., B.P., M.R., and E.R. provided key support for sequence generation and analysis. H.X., L.S., Amin Addetia, V.M.R., N.A.P.L., and M.L.H. processed and sequenced UW Virology samples. R.G., G.M., B.H., and P.D. processed WA DOH samples. Amanda Adler, E.B., P.D.H., K.F., M.I., K.L.. T.R.S., M.T., C.R.W., M.B., J.A.E., M.F., B.R.L., M.J.R., and M.T. processed SCAN samples. S.L., J.S.D., L.M.S., H.Y.C., J.S., and K.R.J. oversaw sample collection and sequencing. A.L.G., D.A.N., and T.B. oversaw the study. N.F.M, C.W., and T.B. wrote the manuscript. All other authors edited the manuscript. **Competing interests:** J.A.E. is a consultant for Sanofi Pasteur and Meissa Vaccines Inc. and receives research support from GlaxoSmithKline, AstraZeneca, and Pfizer. H.Y.C. is a consultant for Merck and GlaxoSmithKline. J.S. is a consultant with Guardant Health, Maze Therapeutics, Camp4 Therapeutics, Nanostring, Phase Genomics, Adaptive Biotechnologies, and Stratos Genomics and has a research collaboration with Illumina. M.B. is a consultant for Merck, VirBio, and Moderna. B.R.L. is a cofounder and CTO of Anavasi Diagnostics Inc., which develops point-of-care tests for COVID-19 and other diseases. N.F.M., C.W., C.D.F., P.R., J.L., L.H.M., B.P., M.R., E.R., H.X., L.S., A. Addetia, V.M.R., N.A.P.L., M.-L.H., R.G., G.M., B.H., P.D., A. Adler, E.B., P.D.H., K.F., M.I., K.L., T.R.S., M.T., C.R.W., M.F., B.R.L., M.J.R., M.T., J.S.D., L.M.S., K.R.J., S.L., A.L.G., D.A.N., and T.B. declare that they have no competing interests. **Data and materials availability:** All data associated with this study are available in the paper or the Supplementary Materials. Data and code associated with this work are available at https://doi.org/10.5281/zenodo.4697912 and https://doi.org/10.5281/zenodo.4701583. These include the R code used to produce the figures [made with ggtree (*47*) and ggplot (*48*)]. SARS-CoV-2 consensus genome sequences associated with this work have been uploaded to GenBank and the GISAID EpiFlu database, and accession numbers are available in the Supplementary Materials (data file S2).

Submitted 29 September 2020
Resubmitted 23 January 2021
Accepted 25 April 2021
Published First Release 3 May 2021
Published 26 May 2021
10.1126/scitranslmed.abf0202

Citation: N. F. Müller, C. Wagner, C. D. Frazar, P. Roychoudhury, J. Lee, L. H. Moncla, B. Pelle, M. Richardson, E. Ryke, H. Xie, L. Shrestha, A. Addetia, V. M. Rachleff, N. A. P. Lieberman, M.-L. Huang, R. Gautom, G. Melly, B. Hiatt, P. Dykema, A. Adler, E. Brandstetter, P. D. Han, K. Fay, M. Ilcisin, K. Lacombe, T. R. Sibley, M. Truong, C. R. Wolf, M. Boeckh, J. A. Englund, M. Famulare, B. R. Lutz, M. J. Rieder, M. Thompson, J. S. Duchin, L. M. Starita, H. Y. Chu, J. Shendure, K. R. Jerome, S. Lindquist, A. L. Greninger, D. A. Nickerson, T. Bedford, Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington State. *Sci. Transl. Med.* **13**, eabf0202 (2021).