Clinical Chemistry 71:1
192–202 (2025)

SARS-CoV-2 Diversity and Transmission on a University Campus across Two Academic Years during the Pandemic

Amanda M. Casto [a,b,*] Miguel I. Paredes,[b,c] Julia C. Bennett,[a,c] Kyle G. Luiten,[a] Peter D. Han,[d] Luis S. Gamboa,[d,e] Evan McDermot,[d,e] Geoffrey S. Gottlieb,[a,f,g] Zachary Acker,[d,e] Natalie K. Lo,[a] Devon McDonald,[a] Kathryn M. McCaffrey,[d] Marlin D. Figgins,[b] Christina M. Lockwood [d,h] Jay Shendure,[d,e,i] Timothy M. Uyeki,[j] Lea M. Starita,[d,e] Trevor Bedford [b,c,d,e,i] Helen Y. Chu,[a,c,†] and Ana A. Weil[a,g,†]

BACKGROUND: Institutions of higher education (IHE) have been a focus of SARS-CoV-2 transmission studies but there is limited information on how viral diversity and transmission at IHE changed as the pandemic progressed.

METHODS: Here we analyze 3606 viral genomes from unique COVID-19 episodes collected at a public university in Seattle, Washington from September 2020 to September 2022.

RESULTS: Across the study period, we found evidence of frequent viral transmission among university affiliates with 60% ($n = 2153$) of viral genomes from campus specimens genetically identical to at least one other campus specimen. Moreover, viruses from students were observed in transmission clusters at a higher frequency than in the overall dataset while viruses from symptomatic infections were observed in transmission clusters at a lower frequency. Although only a small percentage of community viruses were identified as possible descendants of viruses isolated in university study specimens, phylodynamic modeling suggested a high rate of transmission events from campus into the local community, particularly during the 2021–2022 academic year.

CONCLUSIONS: We conclude that viral transmission was common within the university population throughout the study period but that not all university affiliates were equally likely to be involved. In addition, the transmission rate from campus into the surrounding community may have increased during the second year of the study, possibly due to return to in-person instruction.

Introduction

The SARS-CoV-2 pandemic has been characterized by geographical and chronologic variation in circulating variants revealed through viral genome sequencing surveillance (1). These efforts were integral for identifying functional differences among variants and helped to predict fluctuations in incidence of COVID-19. Institutions of higher education (IHE), have been a major focus of surveillance efforts during the pandemic. Although IHE vary widely in size, demographics, and setting, most IHE populations are predominantly made up of young, healthy adults who are at a relatively low risk of severe disease from SARS-CoV-2 (2, 3). Risk may be further mitigated by high rates of vaccine uptake in the setting of vaccine mandates at some IHE (4). However, features of IHE environments, such as communal housing and frequent social events, and a high level of mobility due to travel during academic breaks, may promote the spread of respiratory viruses, including SARS-CoV-2 (5, 6).

SARS-CoV-2 epidemiology and transmission studies at IHE have estimated varying viral transmission rates between IHE populations and their surrounding communities (5, 7–12). Some of this variation is likely due to fixed differences among IHE, as well as temporal variation in viral transmission dynamics across the pandemic. Most studies of SARS-CoV-2 in IHE populations

[a]Division of Allergy and Infectious Diseases, Department of Medicine, University of Washington, Seattle, WA, United States; [b]Vaccine and Infectious Diseases Division, Fred Hutchinson Cancer Center, Seattle, WA, United States; [c]Department of Epidemiology, University of Washington, Seattle, WA, United States; [d]Brotman Baty Institute for Precision Medicine, Seattle, WA, United States; [e]Department of Genome Sciences, University of Washington, Seattle, WA, United States; [f]Environmental Health and Safety Department, University of Washington, Seattle, WA, United States; [g]Department of Global Health, University of Washington, Seattle, WA, United States; [h]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA, United States; [i]Howard Hughes Medical Institute, Chevy Chase, MD, United States; [j]Influenza Division, National Center for Immunization and Respiratory Diseases, Centers for Disease Control, Atlanta, GA, United States.
*Address correspondence to this author at: Division of Allergy and Infectious Diseases, Department of Medicine, University of Washington, 750 Republican St., Box 358061, Seattle, WA 98109, United States. E-mail amcasto@uw.edu.
[†]These authors contributed equally to this work.
Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.
Received April 4, 2024; accepted September 13, 2024.
https://doi.org/10.1093/clinchem/hvae194

have analyzed data from short time frames (such as single academic quarters or semesters), preventing assessment of how transmission changed over time and how new viral variants and changes in mitigation strategies and human behavior shaped transmission as the pandemic progressed.

Here we examine SARS-CoV-2 genomic sequence data and demographic, epidemiologic, and clinical data collected across from September 2020 to September 2022 as part of a university testing program, the Husky Coronavirus Testing (HCT) study. We used 3 different means to assess viral diversity and transmission. First, we characterized diversity of viruses seen on campus and compared this to viral diversity within the state. Second, we identified clusters of closely related HCT sequences and assessed the impact of demographic, epidemiologic, and clinical factors of study participants on cluster membership. Third, we examined phylogenetic relationships between university and community viruses.

Methods and Materials

DATA AND SAMPLE COLLECTION

Data analyzed were collected as part of the HCT study. HCT provided SARS-CoV-2 testing (both by invitation and on demand) for affiliates (students, faculty, and staff) for the University of Washington (UW) main campus and 2 satellite campuses from September 2020 to July 2023. SARS-CoV-2 testing was performed via participant self-swab collection observed by study personnel at testing sites on campus, unobserved self-swab collection returned to an on-campus drop box, and unobserved self-swab collection picked up by courier. Samples were tested for SARS-CoV-2 using a laboratory developed multiplexed RT-qPCR with targets in the *Orf1b* and *S*-genes. Additional details about the study population and setting, the testing program and facilities, data collection, and sample analysis can be found in the Supplementary Methods and have been described previously (13–15).

GENOME SEQUENCING AND ANALYSIS

Genome sequencing was attempted on all specimens that tested positive for the presence of SARS-CoV-2 with an average cycle threshold of 30 or less. Processing of raw sequence data and generation of consensus genomes was performed using a publicly available bioinformatic pipeline (https://github.com/seattleflu/assembly). All genome sequences used in this study were submitted to Global Initiative on Sharing All Influenza Data (GISAID) (16, 17) and SARS-CoV-2 genomes used in analyses that were generated outside the HCT study were downloaded from GISAID. Sequence alignment, masking of problematic loci, and phylogenetic tree generation were performed using Nextstrain (18). Trees were visualized using Auspice. Groups of identical SARS-CoV-2 sequences were identified using a previously described R package (https://github.com/blab/size-genetic-clusters) (19). Additional details of our approach to genome sequencing and analysis and our transmission modeling analysis are described in the Supplementary Methods.

Results

VIRAL LINEAGES AND CLADES COMMON IN WASHINGTON STATE WERE OBSERVED AMONG HCT SPECIMENS

We sequenced 3855 of 6485 SARS-CoV-2 positive specimens collected by HCT from September 2020 to September 2022. These sequences represent 3% of all SARS-CoV-2 genomes in the GISAID EpiCoV database (16, 17) generated from specimens collected in Washington State (WA) during this time. We retained 1 sequence per person per infection and filtered out poor quality sequences, resulting in 3606 sequences analyzed (Fig. 1). Of these sequences, 3195 were collected during academic year 2 (September 1, 2021–September 30, 2022 is referred to as year 2; September 1, 2020–August 31, 2021 is year 1) with 1813 collected between December 1, 2021 and February 28, 2022 (Supplemental Fig. 1). The final sequence set contained sequences from 19 different Nextstrain clades and 115 Pango lineages.

To provide context for diversity seen among SARS-CoV-2 genomes from HCT specimens, we downloaded all SARS-CoV-2 genomes from specimens collected in WA outside the HCT study from September 2020 to September 2022 from the GISAID EpiCoV database (16, 17). After filtering out poor quality and duplicate sequences, 119 215 WA genomes remained, representing 27 clades and 333 lineages. All clades with a frequency of >0.2% and all lineages with a frequency of >0.4% among WA genomes were represented by at least 1 HCT genome. Most lineages in WA were rare (<0.4% of all WA genomes) and more than half ($n = 224$, 67.3%) of all WA lineages were not observed among HCT genomes. There were 6 lineages that were represented in HCT but not the WA sequence set. These were all from samples collected in early January or late March 2022. The percent of WA clades and lineages observed in HCT fluctuated over time; in year 2, these percentages appeared to spike at the beginning of academic quarters (Supplemental Fig. 2; Supplemental Table 1).

AVERAGE DELAY OF 1 MONTH BETWEEN VARIANT OBSERVATION IN WA AND IN HCT

The prevalence of clades among HCT and WA sequences over time is shown in Fig. 2. For clades and

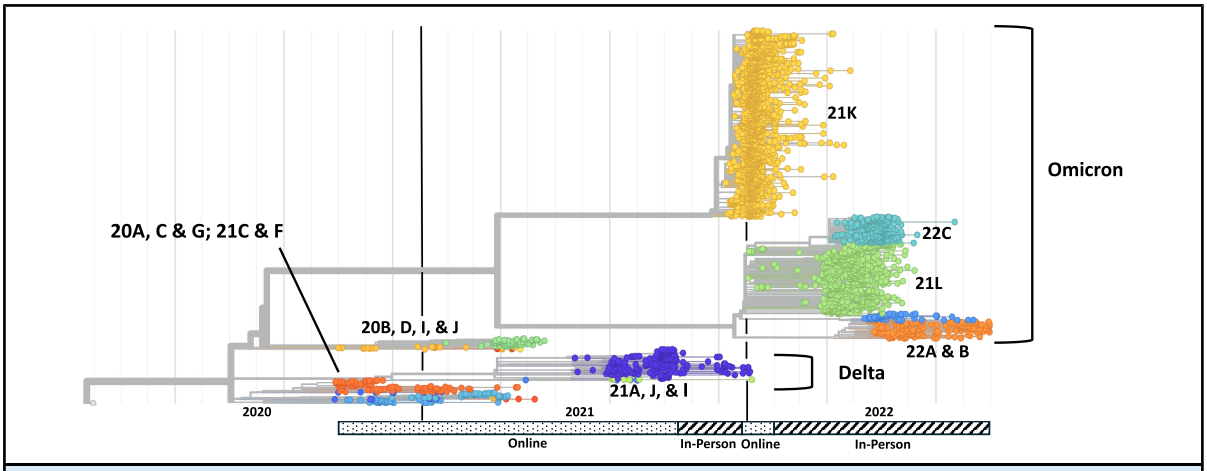

Fig. 1. Phylogenetic tree of all 3606 HCT sequences. Tips are colored by Nextstrain clade with Nextstrain clade designations (i.e., 21K, 22C, etc.) given next to groups of tips of the same color. Date of specimen collection is on the *x*-axis. A bar below the *x*-axis shows timeline of online vs in-person instruction.

lineages seen among both HCT and WA genomes, the average number of days from first observation in WA to observation among HCT specimens was 35.1 days (median 24, range 0–116) for clades and 35.5 days (median 28, range −56–170) for lineages (Supplemental Fig. 3). Ten lineages were observed among HCT specimens before WA specimens. Notably, the BA.2 lineage, which was represented by 428 (11.9%) HCT sequences and 6704 (5.6%) WA sequences and from which all currently circulating SARS-CoV-2 are descended, was among these and was first observed on campus on January 3, 2022. Three of the other lineages first observed in HCT were also collected in January 2022. If we restricted the WA genomes to those from the county where the university campus is located (King County or KC), time from first observation in KC to first observation in HCT was 22.0 days (median 12, range −11 to 84) for clades and 17.0 days (median 10, range −116 to 116) for lineages.

MOST HCT SARS-COV-2 SPECIMENS WERE CLOSELY RELATED TO AT LEAST ONE OTHER HCT SPECIMEN

We used 2 different approaches to identify groups of closely related HCT SARS-CoV-2 genomes. First, we identified groups of identical genomes (which we refer to as "zero distance clusters"). There were 1730 unique haplotypes among HCT sequences, including 2153 sequences that were identical to at least one other HCT sequence and 277 haplotypes represented by >1 HCT sequence (Fig. 3A, Supplemental Table 2). A single Omicron haplotype (clade 21K, lineage BA.1.1) was observed for 655 different sequenced specimens collected from December 17, 2021 until March 8, 2022 (18.2% of all HCT genomes). Of the 277 zero distance clusters, 26 included 10 or more sequences. For each clade, the average size of zero distance clusters decreased with time since clade introduction (Supplemental Fig. 4) (19). The longest period over which a single haplotype was observed was 153 days (clade 21L, lineage BA.2).

We also identified groups of identical sequences among a combined HCT and WA dataset. Of 277 HCT zero distance clusters, 133 (48%) represented a haplotype not observed among WA genomes (we refer to these as "HCT-only zero distance clusters"). The largest HCT-only zero distance cluster (clade 21K, BA.1.20) included 12 sequences and the most persistent cluster (longest period from collection of first to last specimen) was observed over a period of 35 days (clade 21K, BA.1.1). To assess for possible "spill-over" of virus from university affiliates into other populations (Supplemental Note 1), we identified WA viruses that appeared to be descendants of 1 of 133 HCT-only zero distance clusters (see Supplemental Methods). We found a total of 81 such non-HCT viruses, associated with 19 clusters. Over half ($n = 42$, 51.9%) of these 81 viruses were of the BA.2 lineage (clade 21L). The largest number of non-HCT descendants of a single cluster was 37 (clade 21L, BA.2).

We created a phylogenetic tree for each clade that included all HCT and WA genomes. We used these trees to identify clusters (which we refer to as "phylogenetic clusters") of HCT genomes that descend from a single introduction event (see Supplemental Methods). These clusters ranged in size from 2 to 70 sequences with 19 clusters including >10 sequenced specimens

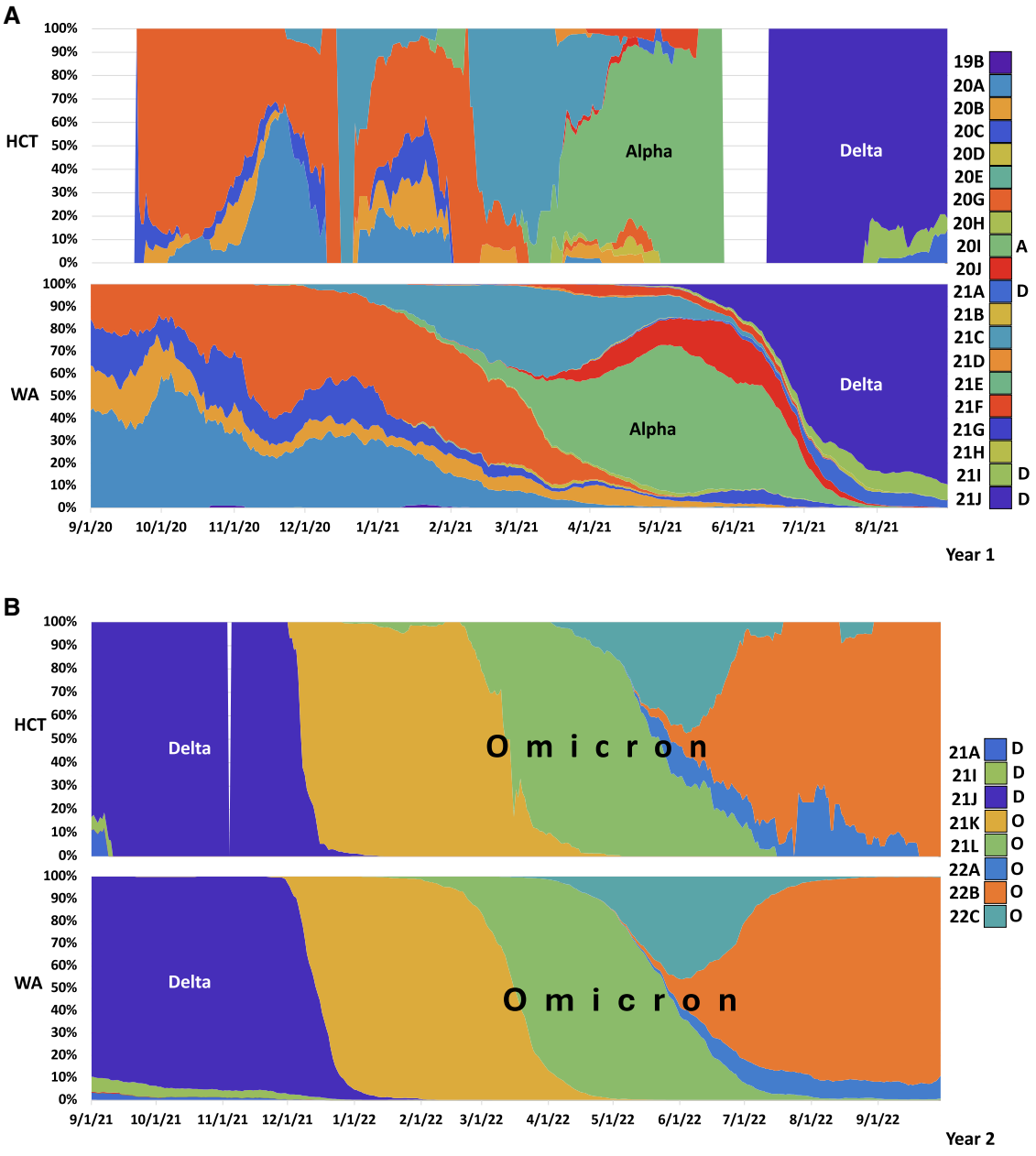

Fig. 2. Frequency of Nextstrain clades among HCT and WA SARS-CoV-2 genomes over time. Each time point on the *x*-axis represents a 2-week sliding window centered on that date. The *y*-axis shows the distribution of genomes collected in that window among different clades. Blank sections for HCT represent 2-week windows during which no sequenced specimens were collected. Chart colors correspond to Nextstrain clades as shown in the legend. Alpha variant clades are labeled A, Delta variant clades are labeled D, and Omicron variant clades are labeled O in the legend. (A), Specimens collected during year 1 (September 1, 2020 to August 31, 2021); (B), Specimens collected during year 2 (September 1, 2021 to September 30, 2022).

(Fig. 3, Supplemental Table 3). Most ($n = 198$, 84.6%) of the 234 HCT phylogenetic clusters included only HCT sequences. However, a total of 218 WA sequences were part of an HCT sequence cluster. The largest number of non-HCT sequences in a single HCT phylogenetic cluster was 37 (clade 20I, lineage B.1.1.7).

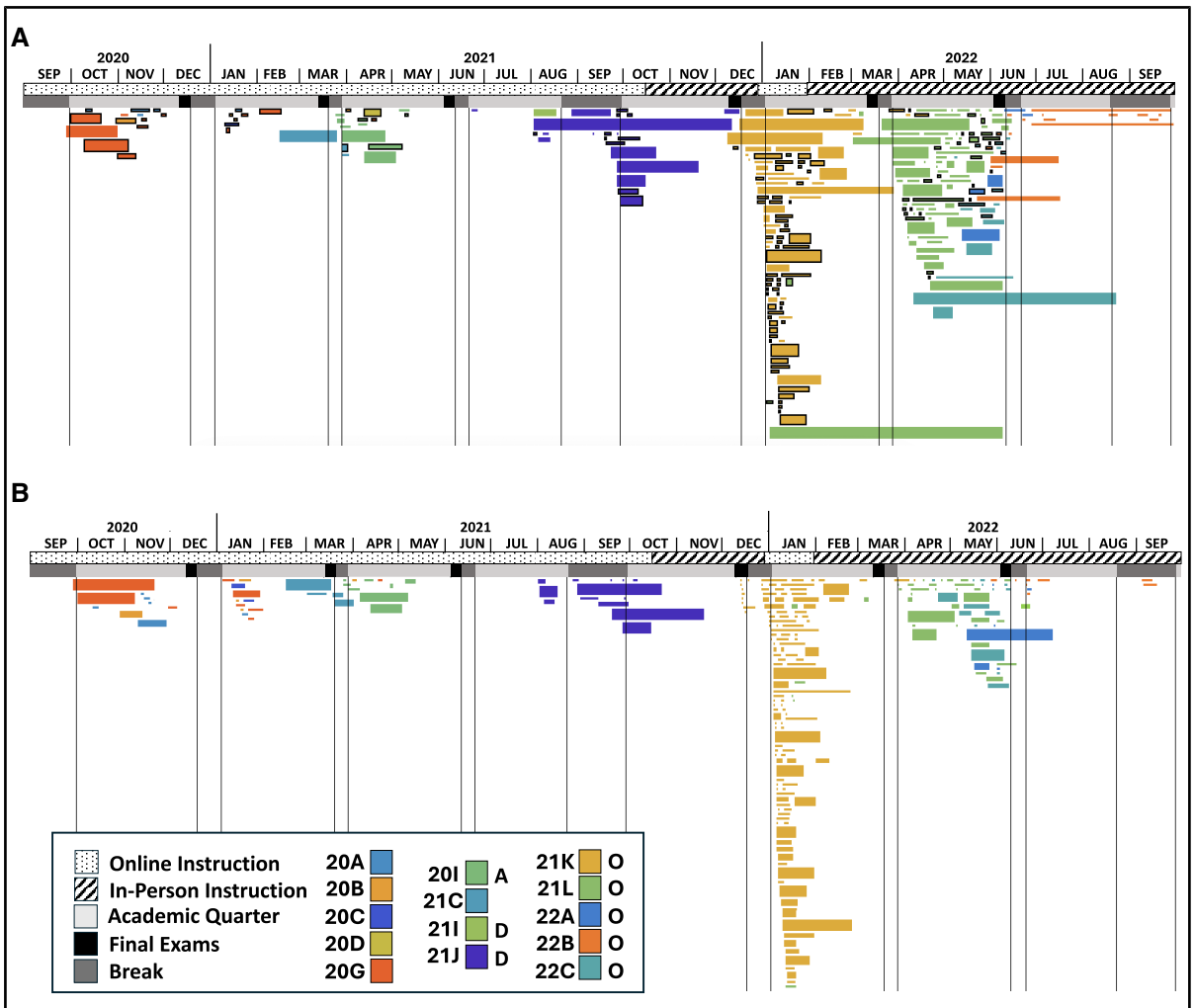

Fig. 3. Chronology of clusters of HCT sequences. Each bar represents a sequence cluster that extends from the collection date of the first specimen in the cluster to the collection date of the last specimen. Month and year of specimen collection is denoted at the top with additional bars to denote dates of academic quarters and timeline of online/in-person instruction. The colors of the cluster bars correspond to the viral clade as indicated in the legend. Alpha variant clades are labeled A, Delta variant clades are labeled D, and Omicron variant clades are labeled O in the legend. The height of the cluster bar is proportional to the number of sequences in the cluster. (A), Zero distance clusters (groups of identical sequences). Bars with black outlines represent HCT-only zero distance clusters (groups of identical sequences with haplotype unique to HCT); (B), Phylogenetic clusters (groups of sequences that cluster phylogenetically).

MODEL SUGGESTS HIGH TRANSMISSION RATE FROM THE UNIVERSITY INTO THE SURROUNDING COMMUNITY

To further explore the relationship between SARS-CoV-2 in university affiliates and the surrounding community, we modeled transmission dynamics to and from the HCT population. We limited the WA sequences in this analysis to those from KC to reflect the community immediately surrounding the university. In addition to including an HCT and KC region in the model, we also included an "other" region representing sequences from outside KC in WA and the rest of the world. After subsampling (see Supplemental Methods), a total of 1137 genomes were used for the model. Results suggested a higher forward migration rate from HCT into KC than vice versa [Fig. 4; 10.8 migration events/lineage/year (95% highest posterior density 4.3–19.9) vs 0.13 migration events/lineage/year (95% highest posterior density 0.068–0.179)].

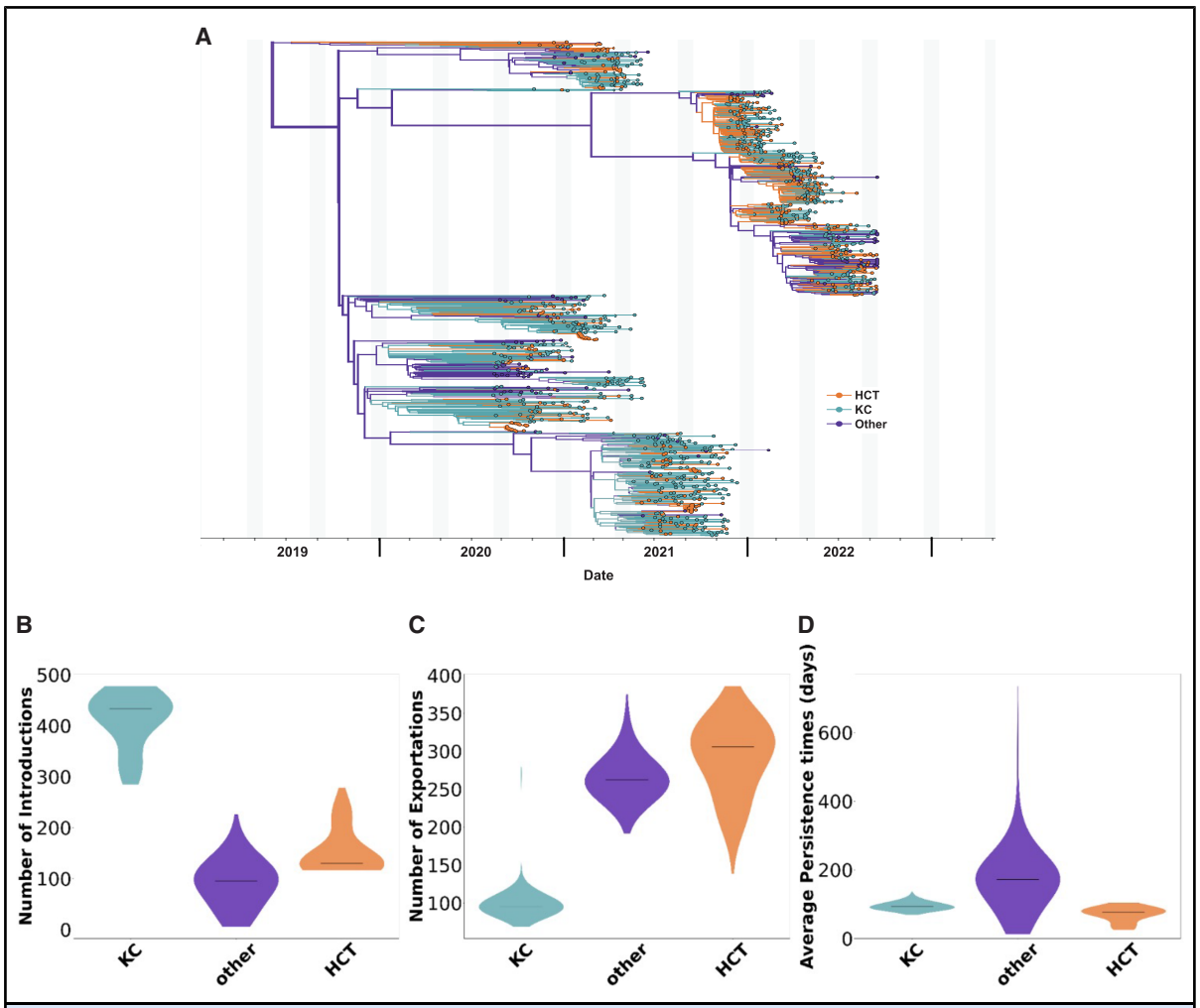

Fig. 4. Phylodynamic analysis of SARS-CoV-2 transmission between KC and HCT populations. (A), Maximum clade credibility tree summary of the Bayesian inference conducted using MASCOT-Skyline on 1137 sequences. Colors correspond to the locations in the legend. KC = Sequences from King County (excluding those from HCT), HCT = Sequences from specimens collected by HCT, other = global contextual sequences from outside King County sampled to increase spatiotemporal diversity. Estimated number of introductions (B), exports (C), and average time of local persistence in days (D) for each region. Horizontal black line denotes median estimates.

We estimated that KC had at least 433 (IQR 415–444) viral introduction events during the study period with at least 130 events (IQR 126–137) coming from the HCT population (Supplemental Note 1). Our model indicated that viral lineages are more likely to circulate longer in the larger KC region (92.6 days, IQR 86.4–101.1 days) than in the HCT population (77.2 days, IQR 71.5–82.6 days). When analyzing transmission patterns across time, we find that viral flow between HCT and KC was dominated by spread from KC to HCT during year 1, and from HCT to KC during year 2 (Fig. 5).

PARTICIPANTS WITH A SEQUENCED VIRAL GENOME WERE REPRESENTATIVE OF THOSE TESTING POSITIVE FOR SARS-COV-2

The 3606 HCT genomes were from 3560 unique individuals (Supplemental Table 4). Most (85.4%) were students, 57.5% identified as female, 8.6% were Latinx, and most were White (50.4%) or Asian (32.0%). Average age at the time of infection was 25.1 years (median 21.3 years, range 17.4–78.7). HCT participants with a sequenced specimen were overall demographically representative of all HCT participants with a positive test (Supplemental Table 5). In total, 3514 individuals

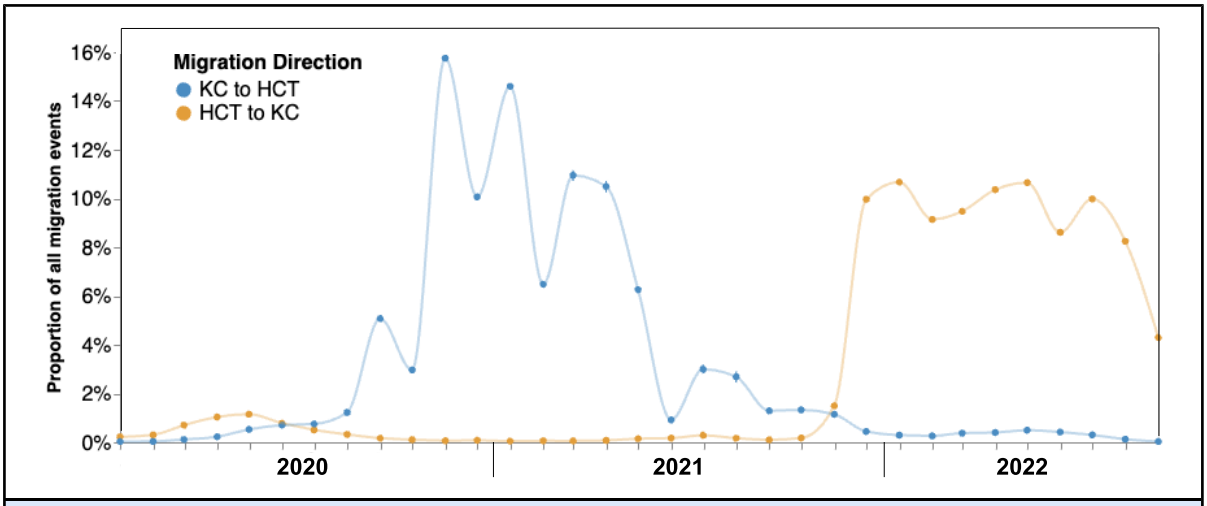

Fig. 5. Proportion of all migration events between HCT and KC. Proportions do not add up to 100% as migration events including the "other" region were excluded.

had only 1 sequence in the dataset while 46 individuals, who experienced infection with >1 clade/lineage of SARS-CoV-2 during the study period, had 2 sequences included in the dataset (Supplemental Table 6, Supplemental Note 2).

SEQUENCES FROM STUDENTS, YOUNGER PEOPLE WERE MORE LIKELY TO CLUSTER WITH OTHER HCT SEQUENCES

Students were overrepresented among reinfected individuals relative to their frequency in the complete dataset (1 proportion *z*-test, 95.7% vs 85.4%, $P = 0.048$). Among reinfected individuals, the proportion of infections associated with symptoms (47.8% or 44/92) was also significantly lower than the proportion of infections associated with symptoms among all participants (2722 out of 3606, $P < 0.0001$, Fig. 6, Supplemental Table 7). Additionally, average age at the time of infection was lower for those who experienced reinfection than for all participants with a sequenced virus (21.1 vs 25.1, $P < 0.0001$). This was also observed when sequences from students and those from faculty/staff were considered separately (20.5 vs 21.7, $P < 0.0001$ and 33.0 vs 44.8, $P = 0.0096$).

Sequences from students were overrepresented among those in zero distance, HCT-only zero distance, and phylogenetic clusters (90.9%, 97.0%, 97.6% vs 85.5%, $P < 0.0001$ for all, Fig. 6, Supplemental Table 7) while sequences from nonstudents (faculty/staff/other) were underrepresented in all 3 cluster types (9.1%, 3.0%, 2.4% vs 14.5%, $P < 0.0001$ for all). Sorority/fraternity members were also overrepresented in all 3 cluster types (21.5%, 23.5%, 31.3% vs 18.6%, $P = 0.00022$, 0.0276, <0.0001), while sequences from symptomatic infections were underrepresented among those in HCT-only zero distance and phylogenetic clusters (64.4%, 65.9% vs 75.5%, $P < 0.0001$ for both). Finally, average age at the time of infection was lower for sequences in all 3 cluster types compared to average age for all sequences (23.4, 21.3, 21.1 vs 25.1, $P < 0.0001$ for all). This difference was also observed when sequences from students were considered separately (21.3, 20.6, 20.6 vs 21.7, $P < 0.0001$ for all). Average ages at infection for sequences from nonstudents in all 3 cluster types did not differ from the overall average age for nonstudents (44.0, 39.2, 43.2 vs 44.8, $P = 0.3922$, 0.6292, and 0.2305). Results of similar analyses for other demographic and epidemiologic variables are shown in Supplemental Fig. 5.

Discussion

We studied SARS-CoV-2 cases with associated viral genomic data in a large, public university population over the first 2 academic years of the pandemic with a focus on characterizing viral diversity and transmission dynamics. To our knowledge, this represents the largest survey to date of IHE SARS-CoV-2 cases with viral sequence data and one of the few based on data collected for >1 year. We found that some measures of viral diversity and transmission dynamics differed between year 1, during which online-only instruction occurred, and year 2, during which classes were conducted mostly in-person, or appeared to be impacted by the academic calendar. We also observed that not all university affiliates were equally likely to be involved in campus-related viral transmission. These findings provide context for and aid

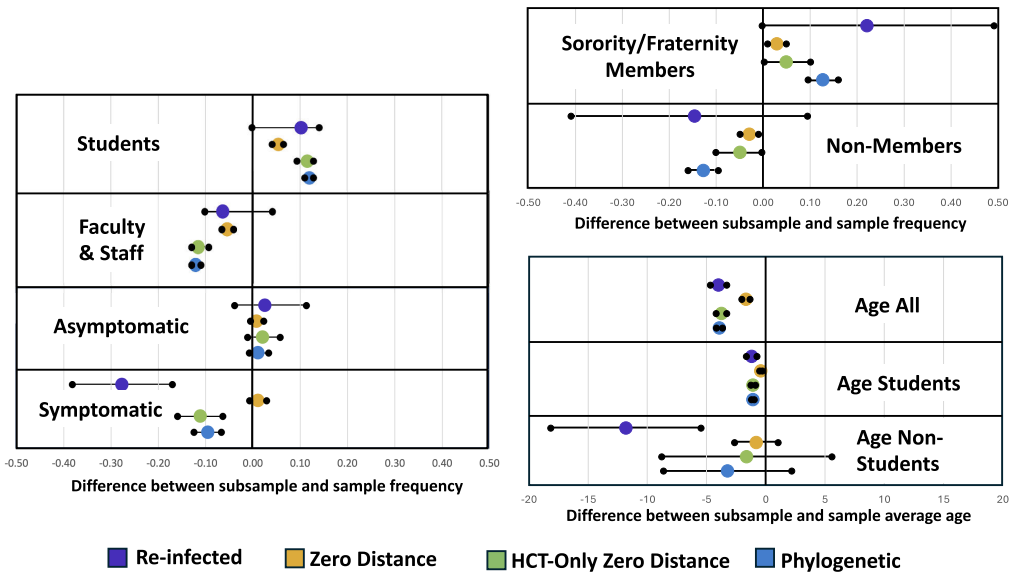

Fig. 6. Demographic and epidemiologic characteristics of participants who experienced reinfection and of those with sequenced specimens in clusters. For students, faculty and staff, asymptomatic, symptomatic, sorority/fraternity members, and nonmembers categories, the mid-line ($x = 0$) represents the percent representation of that category in the whole dataset. The purple, yellow, green, and blue dots give the percentage difference between the frequency of that category among those who experienced reinfection and those with sequences in zero distance (groups of identical sequences), HCT-only zero distance (groups of identical sequences with haplotype unique to HCT), and phylogenetic clusters (groups of sequences that cluster phylogenetically), respectively, and its frequency in the whole dataset. Intervals marked by black dots connected by a bar indicate the 95% confidence interval. For the age all, age students, and age nonstudents categories, the mid-line ($x = 0$) represents the average age at time of infection for all sequences, sequences from students, and sequences from nonstudents, respectively. The purple, yellow, green, and blue dots give the difference between the above averages and the average age of those who experienced reinfection and for sequences in zero distance, HCT-only zero distance, and phylogenetic clusters, respectively. Intervals marked by black dots connected by a bar indicate the 95% confidence interval.

in interpretation of other studies of SARS-CoV-2 at IHE, particularly those with shorter study periods. They also can help administrators of IHE target mitigation strategies toward affiliates at highest risk of campus-related SARS-CoV-2 transmission.

Our study has several unique features, including evaluation of viral diversity on campus compared to the state over time. We found that clades/lineages common in the state were reliably observed within HCT even during year 1, when the number of samples collected was relatively small. In year 2, HCT clade frequencies were almost identical to statewide frequencies. The average delay between first observation of a clade or lineage on campus relative to first observation in the state was shorter during year 2 relative to year 1. These differences may be explained by differences in sample size, although we hypothesize that changes in infection control measures and in the virus, such as a return to in-person instruction, the introduction of more transmissible variants, and overall higher infection rates in year 2, also played a role. The academic calendar appeared to influence the relationship between viral diversity on campus and in the state. We saw spikes in the percent of WA clades/lineages represented on campus at the beginning of academic quarters in year 2. Lineages unique to HCT or seen first in HCT were mostly collected at the beginning of academic quarters in year 2. These observations are consistent with increased importation of viral diversity into the HCT population when students, 10% of whom are international and 15% of whom are out-of-state residents, were returning to campus for the start of in-person instruction. These patterns were not observed in year 1 of the study when classes were held exclusively online.

Our cluster analysis suggested that campus-related transmission was common throughout the study period

as most specimens were closely related to at least one other HCT specimen. Most clusters were confined to a single academic quarter, but there were exceptions to this, suggesting some ongoing transmission among university affiliates even during academic breaks. The cluster analysis also indicated that campus transmission chains could persist for weeks to months despite frequent lineage replacement within the SARS-CoV-2 population. The number of clusters per academic quarter was stable during year 1 and increased during year 2. This seems to be due to the increase in the number of specimens collected in year 2 relative to year 1, rather than a change in transmission dynamics, as the percentage of specimens falling into clusters remained stable across the study period. Finally, we observed that average cluster size for a particular variant decreased with increasing time since emergence of that variant. This is consistent with prior observations in WA (19) and is thought to be indicative of a decrease in the effective reproduction number for viral variants the longer they have been circulating in a population.

Transmission modeling suggested that the average persistence time of viral lineages on campus was 77.2 days, providing further support for the hypothesis that viral transmission among university affiliates was common during the study period. Model results also suggested that viral flow between KC and HCT was mostly into the campus population during year 1 and then from the campus population during year 2. This could be the result of spread of new viral variants (Delta, Omicron), the return to in-person instruction in year 2 with an increase in campus population relative to year 1, relative vaccination rates in the 2 populations (Supplemental Note 3), or some combination of these. Given that few WA sequences appeared to descend from HCT clusters, we were surprised that the model estimated a substantial number of transmission events from HCT to KC (estimated minimum of 130) and a significantly higher forward migration rate than from KC to HCT. Most of this difference is likely attributable to the vastly different population sizes of the 2 regions (Supplemental Note 1). Notably, the results of the modeling analysis do not suggest that SARS-CoV-2 cases in the HCT population had a disproportionate impact on KC, but do indicate that nearly all HCT transmission chains resulted in an infection in the KC population.

While findings that viruses from students were disproportionately represented in clusters and that average age of those with viruses in clusters was younger than the population average are not surprising, these results provide evidence to support the direction of limited infection control resources at IHE to those most likely to be involved in transmission chains. Studies with more detailed information about participant housing, activities, and behaviors could help to further delineate drivers of this association between student status, age, and cluster membership to more strategically target infection control resources. Our results indicated that sorority/fraternity members were disproportionally represented in transmission clusters. Data on membership in other social groups, such as sport teams or clubs, were not collected by HCT and we are unable to comment on the impact of participation in these activities on involvement in campus-related viral transmission. We note, however, that unlike most social groups, sorority and fraternity members frequently live together in communal housing, which could be a major driver of their risk of involvement in viral transmission chains. Finally, the appearance of differences in clustering observed for symptomatic vs asymptomatic cases was an unanticipated result. One possible explanation is differences in behavior of the 2 groups, such as increased social distancing and isolation by symptomatic individuals.

Limitations of our study included fluctuations in the collection rate of sequenced samples. We believe that the proportion of HCT SARS-CoV-2 cases for which a genome was generated was roughly consistent throughout the study period and that these fluctuations were primarily driven by changes in SARS-CoV-2 transmission rates on campus. Limiting consideration to a set number of sequences per unit time through subsampling would have permitted proportionate representation of different timeframes within the overall study period. However, it would have also restricted the number of genomes used in our dataset, affecting our ability to describe viral diversity and detect putative transmission clusters. This variation in sampling density through time and the resulting over- and underrepresentation of some time periods should be taken into consideration when interpreting the results. Particularly notable is the large spike in sequenced samples collected from December 2021 to March 2022, corresponding to early Omicron transmission. Observations made about sample diversity and transmission dynamics during this time should not be assumed to be applicable to times of lower SARS-CoV-2 incidence. Limitations also included incomplete case identification and sampling on the university campus and in WA state during the study period. Additionally, sequence data could not be generated for all HCT cases. University affiliates could test outside the HCT and sequenced specimens from affiliates collected outside HCT were classified as non-HCT WA sequences. This reflects the broader challenge of the lack of associated demographic and clinical data for most SARS-CoV-2 genomes in GISAID. This limits our understanding of relationships between viral transmission and factors such as gender, age, race/ethnicity, symptoms, and place of residence below the level of state. Changing availability of genomic surveillance data over time and unequal sampling across WA and the world affected the probability

that a case was represented by a sequence in our dataset. We attempted to mitigate bias in our modeling analysis by using spatiotemporal subsampling, which has been shown to improve inferential power of similar models (20). However, conclusions of our modeling analysis, and all similar modeling analyses, are limited by the fact that results are based on assumptions about population sizes and migration rates, which may be inaccurate, and on the input sequence set, which represents a small fraction of all SARS-CoV-2 cases occurring in the 3 regions during the study period.

Populations of IHE have been and will continue to be a focus of SARS-CoV-2 research out of concern that these populations are prone to frequent transmission, which may have significant impacts on IHE and surrounding communities. Varying results have made it challenging to derive generalizable lessons from studies conducted in IHE. Here we have characterized viral diversity and transmission at a single IHE over 2 years to gain an understanding of how viral diversity and transmission dynamics at a single institution can vary over time. These results aid in the synthesis of results from previous studies into a cohesive knowledge base, which is vital to the optimization of interventions to limit spread of SARS-CoV-2 and other respiratory viruses in IHE populations.

Data Availability Statement

GISAID dataset identifier is EPI_SET_240110xb. All genome sequences and associated metadata in this dataset are published in GISAID's EpiCoV database. To view the contributors of each individual sequence with details such as accession number, virus name, collection date, originating laboratory, submitting laboratory, and list of Authors, visit 10.55876/gis8.240110xb.

Additional data and software files are available on GitHub (https://github.com/amcasto/huskytesting_SARSCoV2genomics_First2Years) and Zenodo (DOI: 10.5281/zenodo.13997411).

Author Declaration

A version of this paper was previously posted as a preprint on medRxiv as https://doi.org/10.1101/2024.02.29.24303285.

Supplemental Material

Supplemental material is available at *Clinical Chemistry* online.

Nonstandard Abbreviations: IHE, Institutes of higher education; HCT, husky coronavirus testing; UW, University of Washington; GISAID, global initiative on sharing all influenza data; WA, Washington State; KC, King County.

Genes: *Orf1b*, Sars-CoV-2 open reading frame 1b; *S*, Sars-CoV-2 spike gene.

Author Contributions: *The corresponding author takes full responsibility that all authors on this publication have met the following required criteria of eligibility for authorship: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved. Nobody who qualifies for authorship has been omitted from the list.*

Amanda Casto (Conceptualization-Lead, Formal analysis-Lead, Methodology-Lead, Software-Equal, Validation-Lead, Visualization-Lead, Writing—original draft-Lead, Writing—review & editing-Lead), Miguel Paredes (Conceptualization-Supporting, Formal analysis-Supporting, Methodology-Supporting, Software-Equal, Validation-Supporting, Visualization-Supporting, Writing—original draft-Supporting, Writing—review & editing-Supporting), Julia Bennett (Data curation-Equal, Investigation-Equal, Project administration-Supporting, Writing—review & editing-Supporting), Kyle Luiten (Data curation-Equal, Investigation-Equal, Software-Supporting, Writing—review & editing-Supporting), Peter Han (Data curation-Supporting, Investigation-Supporting, Writing—review & editing-Supporting), Luis Gamboa (Data curation-Supporting, Investigation-Supporting, Writing—review & editing-Supporting), Evan McDermot (Data curation-Supporting, Formal analysis-Supporting, Writing—review & editing-Supporting), Geoffrey Gottlieb (Project administration-Supporting, Resources-Supporting, Supervision-Supporting, Writing—review & editing-Supporting), Zack Acker (Data curation-Supporting, Investigation-Supporting, Project administration-Supporting, Writing—review & editing-Supporting), Natalie Lo (Data curation-Supporting, Investigation-Supporting, Project administration-Supporting, Writing—review & editing-Supporting), Devon McDonald (Data curation-Supporting, Investigation-Supporting, Writing—review & editing-Supporting), Kathryn McCaffrey (Data curation-Supporting, Investigation-Supporting, Writing—review & editing-Supporting), Marlin Figgins (Writing—review & editing-Supporting), Christina Lockwood (Funding acquisition-Supporting, Project administration-Supporting, Resources-Supporting, Supervision-Supporting, Writing—review & editing-Supporting), Jay Shendure (Funding acquisition-Supporting, Project administration-Supporting, Resources-Supporting, Supervision-Supporting, Writing—review & editing-Supporting), Timothy Uyeki (Supervision-Supporting, Writing—review & editing-Supporting), Lea Starita (Funding acquisition-Supporting, Project administration-Supporting, Resources-Supporting, Supervision-Supporting, Writing—review & editing-Supporting), Trevor Bedford (Conceptualization-Supporting, Funding acquisition-Supporting, Methodology-Supporting, Project administration-Supporting, Resources-Supporting, Software-Supporting, Supervision-Supporting, Writing—original draft-Supporting, Writing—review & editing-Supporting), Helen Chu (Conceptualization-Supporting, Funding acquisition-Lead, Project administration-Equal, Resources-Lead, Supervision-Equal, Writing—original draft-Supporting, Writing—review & editing-Supporting), and Ana Weil (Conceptualization-Supporting, Funding acquisition-Supporting, Project administration-Equal, Resources-Supporting, Supervision-Equal, Writing—original draft-Supporting, Writing—review & editing-Supporting)

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form.*

Research Funding: This work was supported by a Howard Hughes Medical Institute Covid Supplement Award to T. Bedford (who is a Howard Hughes Medical Institute Investigator) and by the United States Senate and House of Representatives, Bill 748, Coronavirus Aid, Relief, and Economic Security Act. M.I. Paredes is an ARCS Foundation scholar.

Disclosures: A.M. Casto has received a K08 award from the National Institute of Allergy and Infectious Diseases. G.S. Gottlieb has received research grants and/or research support from the US National Institutes of Health, the University of Washington, the Bill & Melinda Gates Foundation, Gilead Sciences, Alere Technologies, Merck & Co., Janssen Pharmaceutica, Cerus Corporation, ViiV Healthcare, Bristol-Myers Squibb, Roche Molecular Systems, Abbott Molecular Diagnostics, and THERA Technologies/TaiMed Biologics, Inc, all outside of the submitted work. C.M. Lockwood is an associate editor for *Clinical Chemistry*, ADLM, and reports that her spouse is an employee of Bayer. H.Y. Chu reports consulting for Ellume, Pfizer, and the Bill and Melinda Gates Foundation; has served on advisory boards for Vir, Merck and Abbvie; has conducted CME teaching with Medscape, Vindico, and Clinical Care Options; and has received research funding from Gates Ventures, and support and reagents from Ellume and Cepheid outside of the submitted work.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

Acknowledgments: We would first like to thank all the study participants and the HCT study team. We would also like to thank the University of Washington, including UW Environment Health and Safety team (Katia Harb, Sheryl Schwartz, Natalie Thiel, Kim Baker, and Julie Skene) and the UW Covid Incident Command team (Margaret Shepherd, Josh Gana, Pamela Schreiber, and Jack Martin). In addition, we would like to thank Jessica O'Hanlon, Caitlin Wolf, Ariana Magedson, Melissa Truong, Tessa Wright, and Deborah Nickerson from the University of Washington; Michael Boeckh from the Fred Hutchinson Cancer Center; and Janet Englund from the Seattle Children's Research Institute for their contributions to the HCT study. Finally, we would like to thank all contributors of data to GISAID.

References

1. Tosta S, Moreno K, Schuab G, Fonseca V, Segovia FMC, Kashima S, et al. Global SARS-CoV-2 genomic surveillance: what we have learned (so far). Infect Genet Evol 2023;108:105405.
2. Dudley J. COVID-19 transmission under the public health radar: high prevalence in young adults for COVID-19 pandemic wave 1. Int J Infect Dis 2022;116:S29.
3. Romero Starke K, Reissig D, Petereit-Haack G, Schmauder S, Nienhaus A, Seidler A. The isolated effect of age on the risk of COVID-19 severe outcomes: a systematic review with meta-analysis. BMJ Glob Health 2021;6:e006434.
4. Petros BA, Turcinovic J, Welch NL, White LF, Kolaczyk ED, Bauer MR, et al. Early introduction and rise of the omicron severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variant in highly vaccinated university populations. Clin Infect Dis 2023;76:e400–8.
5. Valesano AL, Fitzsimmons WJ, Blair CN, Woods RJ, Gilbert J, Rudnik D, et al. SARS-CoV-2 genomic surveillance reveals little spread from a large university campus to the surrounding community. Open Forum Infect Dis 2021;8:ofab518.
6. Nickbakhsh S, Hughes J, Christofidis N, Griffiths E, Shaaban S, Enright J, et al. Genomic epidemiology of SARS-CoV-2 in a university outbreak setting and implications for public health planning. Sci Rep 2022;12:11735.
7. Richmond CS, Sabin AP, Jobe DA, Lovrich SD, Kenny PA. SARS-CoV-2 sequencing reveals rapid transmission from college student clusters resulting in morbidity and deaths in vulnerable populations. Preprint at http://medrxiv.org/lookup/doi/10.1101/2020.10.12.20210294 (2020).
8. Leidner AJ, Barry V, Bowen VB, Silver R, Musial T, Kang GJ, et al. Opening of large institutions of higher education and county-level COVID-19 incidence—United States, July 6–September 17, 2020. MMWR Morb Mortal Wkly Rep 2021;70:14–9.
9. Andersen MS, Bento AI, Basu A, Marsicano CR, Simon KI. College openings in the United States increase mobility and COVID-19 incidence. PLoS One 2022;17:e0272820.
10. Srinivasa VR, Griffith MP, Waggle KD, Johnson M, Zhu L, Williams JV, et al. Genomic epidemiology of severe acute respiratory syndrome coronavirus 2 transmission among university students in Western Pennsylvania. J Infect Dis 2023;228:37–45.
11. Turcinovic J, Kuhfeldt K, Sullivan M, Landaverde L, Platt JT, Alekseyev YO, et al. Transmission dynamics and rare clustered transmission within an urban university population before widespread vaccination. J Infect Dis 2024;229:485–92.
12. Turcinovic J, Kuhfeldt K, Sullivan M, Landaverde L, Platt JT, Doucette-Stamm L, et al. Linking contact tracing with genomic surveillance to deconvolute SARS-CoV-2 transmission on a university campus. iScience 2022;25:105337.
13. Weil AA, Luiten KG, Casto AM, Bennett JC, O'Hanlon J, Han PD, et al. Genomic surveillance of SARS-CoV-2 omicron variants on a university campus. Nat Commun 2022;13:5240.
14. Weil AA, Sohlberg SL, O'Hanlon JA, Casto AM, Emanuels AW, Lo NK, et al. SARS-CoV-2 epidemiology on a public university campus in Washington state. Open Forum Infect Dis 2021;8:ofab464.
15. Bennett JC, Luiten KG, O'Hanlon J, Han PD, McDonald D, Wright T, et al. Utilizing a university testing program to estimate relative effectiveness of monovalent COVID-19 mRNA booster vaccine versus two-dose primary series against symptomatic SARS-CoV-2 infection. Vaccine 2024;42:1332–41.
16. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data—from vision to reality. Euro Surveill 2017;22:30494.
17. Khare S, Gurry C, Freitas L, Schultz M B, Bach G, Diallo A, et al. GISAID's role in pandemic response. China CDC Wkly 2021;3:1049–51.
18. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34:4121–3.
19. Tran-Kiem C, Bedford T. Estimating the reproduction number and transmission heterogeneity from the size distribution of clusters of identical pathogen sequences. Proc Natl Acad Sci U S A. 2024;121:e2305299121.
20. Layan M, Müller NF, Dellicour S, De Maio N, Bourhy H, Cauchemez S, Baele G. Impact and mitigation of sampling bias to determine viral spread: evaluating discrete phylogeography through CTMC modeling and structured coalescent model approximations. Virus Evol 2023;9:vead010.