RESEARCH ARTICLE
Host-Microbe Biology

# SARS-CoV-2 Genomic Variation in Space and Time in Hospitalized Patients in Philadelphia

**John Everett,[a] Pascha Hokama,[a] Aoife M. Roche,[a] Shantan Reddy,[a] Young Hwang,[a] Lyanna Kessler,[a] Abigail Glascock,[a] Yize Li,[a] Jillian N. Whelan,[a] Susan R. Weiss,[a] Scott Sherrill-Mix,[a] Kevin McCormick,[a] Samantha A. Whiteside,[b] Jevon Graham-Wooten,[b] Layla A. Khatib,[b] Ayannah S. Fitzgerald,[b] Ronald G. Collman,[b] Frederic Bushman[a]**

[a]Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[b]Pulmonary, Allergy and Critical Care Division, Department of Medicine; University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA

John Everett and Pascha Hokama are co-first authors. John Everett led the bioinformatic analysis, and Pascha Hokama led the wet-side workup of patient samples and sequence acquisition.

**ABSTRACT** The severe acute respiratory coronavirus 2 (SARS-CoV-2) is the cause of the global outbreak of COVID-19. The epidemic accelerated in Philadelphia, PA, in the spring of 2020, with the city experiencing a first peak of infections on 15 April, followed by a decline through midsummer. Here, we investigate spread of the epidemic in the first wave in Philadelphia using full-genome sequencing of 52 SARS-CoV-2 samples obtained from 27 hospitalized patients collected between 30 March and 17 July 2020. Sequences most commonly resembled lineages circulating at earlier times in New York, suggesting transmission primarily from this location, though a minority of Philadelphia genomes matched sequences from other sites, suggesting additional introductions. Multiple genomes showed even closer matches to other Philadelphia isolates, suggestive of ongoing transmission within Philadelphia. We found that all of our isolates contained the D614G substitution in the viral spike and belong to lineages variously designated B.1, Nextstrain clade 20A or 20C, and GISAID clade G or GH. There were no viral sequence polymorphisms detectably associated with disease outcome. For some patients, genome sequences were determined longitudinally or concurrently from multiple body sites. In both cases, some comparisons showed reproducible polymorphisms, suggesting initial seeding with multiple variants and/or accumulation of polymorphisms after infection. These results thus provide data on the sources of SARS-CoV-2 infection in Philadelphia and begin to explore the dynamics within hospitalized patients.

**IMPORTANCE** Understanding how SARS-CoV-2 spreads globally and within infected individuals is critical to the development of mitigation strategies. We found that most lineages in Philadelphia had resembled sequences from New York, suggesting infection primarily but not exclusively from this location. Many genomes had even nearer neighbors within Philadelphia, indicating local spread. Multiple genome sequences were available for some subjects and in a subset of cases could be shown to differ between time points and body sites within an individual, indicating heterogeneous viral populations within individuals and raising questions on the mechanisms responsible. There was no evidence that different lineages were associated with different outcomes in patients, emphasizing the importance of individual-specific vulnerability.

**KEYWORDS** SARS-CoV-2, COVID-19, coronavirus, genome sequencing, Philadelphia

The disease COVID-19 is caused by infection with the betacoronavirus SARS-CoV-2 (1). As with other coronaviruses, transmission typically takes place via droplets and aerosols or by contact with contaminated surfaces (2–4). SARS-CoV-2 was identified

**Citation** Everett J, Hokama P, Roche AM, Reddy S, Hwang Y, Kessler L, Glascock A, Li Y, Whelan JN, Weiss SR, Sherrill-Mix S, McCormick K, Whiteside SA, Graham-Wooten J, Khatib LA, Fitzgerald AS, Collman RG, Bushman F. 2021. SARS-CoV-2 genomic variation in space and time in hospitalized patients in Philadelphia. mBio 12:e03456-20. https://doi.org/10.1128/mBio.03456-20.

**Editor** Christina L. Stallings, Washington University School of Medicine in St. Louis

Address correspondence to Ronald G. Collman, collmanr@pennmedicine.upenn.edu, or Frederic Bushman, bushman@pennmedicine.upenn.edu.

This article is a direct contribution from Frederic Bushman, a Fellow of the American Academy of Microbiology, who arranged for and secured reviews by Efrem Lim, Arizona State University, and Eric Delwart, Blood Systems Research Institute.

**Received** 8 December 2020
**Accepted** 11 December 2020
**Published** 19 January 2021

first in China in December 2019 and later in 2020 in most countries. The World Health Organization declared COVID-19 a global pandemic on 11 March 2020.

In the United States, SARS-CoV-2 infection was first recognized in the state of Washington in January 2020. By March 2020, outbreaks had been detected in all 50 states (5). High infection rates were detected in New York City in the spring of 2020, with the first case identified on 29 February 2020 (6).

Here, we investigate samples from the later wave of infection in the city of Philadelphia, PA, taking advantage of viral whole-genome sequencing. The first case of COVID-19 was detected in Philadelphia on 8 March 2020. The epidemic spread rapidly, reaching a first peak of 601 newly diagnosed cases on 17 April 2020 and waned during the remainder of our sampling period, which closed 17 July 2020 (Fig. 1A).

A variety of polymorphisms have been described in the SARS-CoV-2 genome, which provide a means of tracking infections (5–11) and also a window on viral biology. One notable substitution encodes the spike protein (S) variant D614G. The viral spike protein is present on the viral surface and is responsible for binding to the ACE2 host cell receptor and directing membrane fusion and viral entry. The D614G substitution has been proposed to promote infection of human cells, and this variant has spread globally at the expense of other genotypes (12–15). The D614G variant in recently isolated genomes often cooccurs with a mutation encoding P314L in the virus-encoded RNA-dependent RNA polymerase (RdRp) located on ORF1b.

Here, we report the sequence of 52 high-quality genomes from 27 subjects in Philadelphia, allowing investigation of the origins of the local epidemic and assessment of viral variation in patients. All genomes contained the D614G spike substitution and the P314L RdRp substitution. Several nomenclatures have been proposed for this SARS-CoV-2 lineage—these include lineage B.1, Nextstrain clade 20A or 20C, and GISAID clade G or GH; an older designation is the A2a clade (16, 17). Comparison of viral genomes from Philadelphia to sequences from other locations showed that those from New York were commonly the most similar, suggesting a source of the local epidemic. A minority of cases showed nearest neighbors from other sites, suggesting additional introductions from other locations. Comparison of genomes from within Philadelphia showed that many had even better matches to local isolates, suggestive of community spread. We also investigate viral genome variants present at different body sites in patients and longitudinally over time, revealing reproducible polymorphisms in some cases. Lastly, we did not find any strong association of viral genome variation with patient outcomes, as in previous work (18), indicating that subject-specific factors rather than virus-specific factors likely dominate clinical course.

## RESULTS

**The COVID-19 epidemic in Philadelphia.** Whole-genome sequences were obtained from 27 patients hospitalized at the Hospital of the University of Pennsylvania in Philadelphia (Table 1; see also Table S1 in the supplemental material). The sampling period began on 30 March, 22 days after the first detection of SARS-CoV-2 in Philadelphia, extended through the peak daily case number of 601 on 15 April, and was closed after daily cases fell to 195 on 17 July 2020 (Fig. 1A). Of the 27 patients, 12 were female and 15 were male, 20 were Black, five were white, one was Asian, and one identified as other. The median age was 64 years (range 28 to 90), and all but one had at least one underlying major organ system comorbidity (Table 1). Twenty-one subjects received corticosteroids, 12 received hydroxychloroquine, and five received remdesivir. Seventy-eight percent required intubation. Nineteen recovered and left the hospital, and eight died. There were no significant differences between surviving and nonsurviving patients in age, gender, race, underlying comorbidities, or treatment, although the numbers compared were modest.

High-quality complete genome sequences (Table S2) were obtained from 52 samples. Samples yielding high-quality genome sequences included nasopharyngeal (NP) swabs $n = 6$), oropharyngeal (OP) swabs ($n = 8$), pooled NP+OP swabs ($n = 21$), saliva ($n = 1$), and

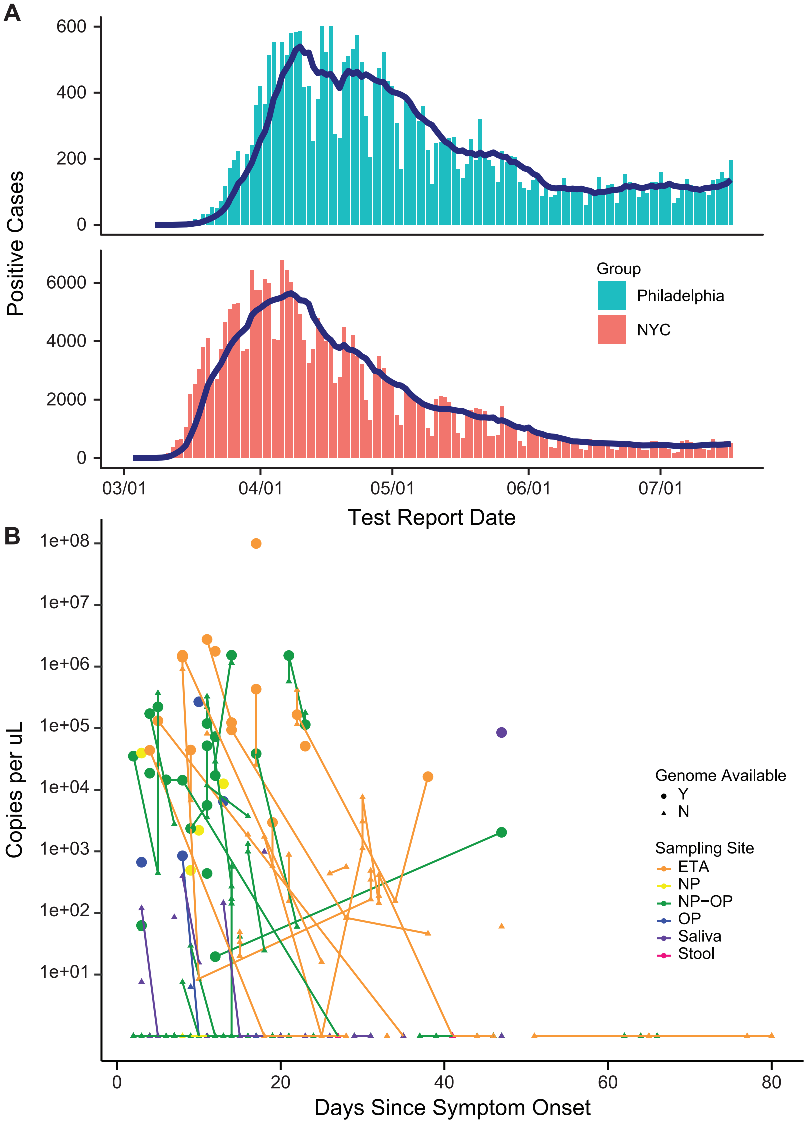


**FIG 1** The first wave of the COVID-19 epidemic in Philadelphia. (A) Number of positive cases recorded per day in Philadelphia and New York City from March to July 2020. Data are provided by the city of New York (https://www1.nyc.gov/site/doh/covid/covid-19-data-testing.page) and the city of Philadelphia (https://www.phila.gov/programs/coronavirus-disease-2019-covid-19/testing-and-data/). (B) Levels of SARS-CoV-2 RNA detected in patient samples using RT-qPCR. The *x* axis shows days after symptom onset, and the *y* axis shows the number of viral RNA copies in samples after RNA purification. Colors show sample type. Colored lines connect samples of the same type obtained longitudinally from within each patient. All samples available are shown for each subject who provided at least one high-quality genome sequence of that sample type. ETA, endotracheal aspirate; NP, nasopharyngeal; OP, oropharyngeal.

endotracheal aspirates (ETA) of lung secretions from intubated patients (*n* = 16). In nine cases, viral genomes were expanded in cell culture prior to sequence acquisition. Sequences were also obtained from two samples of the reference isolate USA-WA1-2020, which was isolated in Seattle, WA, from the first patient identified in the United States on 20 January 2020.

Viral genome copy numbers were assayed in each sample (Fig. 1B). Viral RNA copies per sample typically fell with time after symptom onset, as has been reported previously in many studies. A minority of patients had prolonged RNA detection, allowing analysis of genome sequences over 1 month of viral persistence in the infected subjects.

**TABLE 1** Characteristics of the study participants[a]

| | All ($n = 27$) | Nonsurvivors ($n = 8$) | Survivors ($n = 19$) |
|---|---|---|---|
| Age, median (range) | 64 (28–90) | 66 (46–85) | 62 (28–90) |
| Male/female | 15/12 | 4/4 | 11/8 |
| Race | | | |
| Black | 20 | 5 | 15 |
| White | 5 | 2 | 3 |
| Asian | 1 | 1 | 0 |
| Other | 1 | 0 | 1 |
| Hispanic/Latinx | 0 | 0 | 0 |
| Major comorbidities | | | |
| Diabetes | 15 | 5 | 10 |
| Hypertension | 24 | 8 | 16 |
| CAD/CVD | 12 | 5 | 7 |
| Cancer | 4 | 3 | 1 |
| HIV | 2 | 1 | 1 |
| Organ transplant | 1 | 0 | 1 |
| Chronic lung disease | 12 | 3 | 9 |
| Renal disease (≥stage 4) | 8 | 2 | 6 |
| None | 1 | 0 | 1 |
| BMI, median (range) | 31.1 (17–48) | 27.5 (17–46) | 31.1 (17–48) |
| Mod/severe obesity (BMI >35) | 8 | 2 | 6 |
| Treatment | | | |
| Corticosteroids | 21 | 7 | 14 |
| Hydroxychloroquine | 12 | 3 | 9 |
| Remdesivir | 5 | 2 | 3 |
| Max WHO score, median (range) | | 10 | 8 (4–9) |
| Days from hospitalization to discharge/death, median (range) | 24 (3–61) | 22.5 (10–39) | 32 (4–63) |

[a] Abbreviations: CAD, coronary artery disease; CVD, cardiovascular disease; BMI, body mass index; Mod, moderate.

**Whole-genome sequence analysis of SARS-CoV-2.** Genomes were analyzed by reverse transcription of the viral RNA to make a cDNA copy, PCR amplification of genome segments, Nextera library preparation, and Illumina sequencing. Initially, we devised a protocol based on amplifying the viral genome in six segments. We successfully sequenced several SARS-CoV-2 isolates but noticed that high concentrations of viral RNA were required for efficient amplification. We thus substituted the ARTIC primer set, which amplifies the SARS-CoV-2 genome as 98 shorter amplicons (19). We found that this protocol yielded complete genome sequences more efficiently than the six-amplicon protocol, likely because of greater PCR efficiency with shorter amplicons. We note that our sequencing approach yields the average base at each position in the population—in the interest of high throughput, no effort was made to isolate single genomes prior to sequence acquisition.

An analytical pipeline was devised based on the POLAR protocol for sequence assembly and characterization (19). Sequence reads were aligned to the USA-WA1-2020 reference genome, and variants were identified (Table S3). To be accepted for analysis, genome sequences were required to have least 95% coverage of the USA-WA1-2020 reference (20) and a minimum read depth of 5 reads per position. Results from two subjects were excluded due to failure to meet our sample quality control standards.

**Investigating the origin of the epidemic in Philadelphia.** Polymorphisms were identified in viral genomes isolated in Philadelphia by comparison to the USA-WA1-2020 reference isolate (20). Sites of polymorphisms were cataloged, and genomes were arranged on a phylogenetic tree (Fig. 2 and Tables S2 and S3). Samples from the same subjects clustered together for 26 of the 27 subjects. In the exceptional case, subject 211, the comparisons showing differences involved genomes that were sequenced

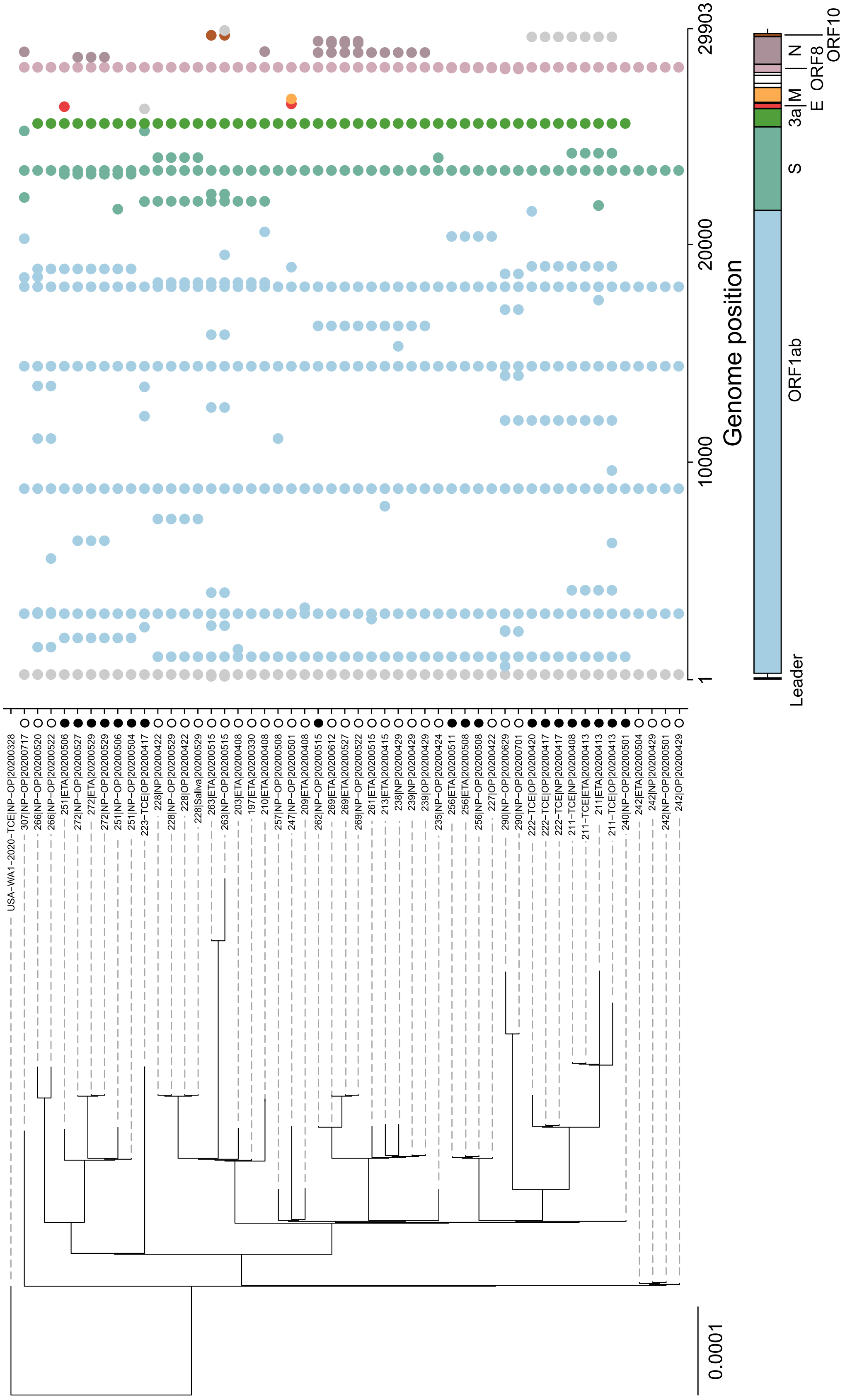


**FIG 2** Sequence polymorphisms in SARS-CoV-2 isolates from Philadelphia. Polymorphisms are shown relative to the USA-WA1-2020 reference isolate. A tree generated using hierarchical clustering (UPGMA method) is shown to the

(Continued on next page)

directly versus isolates that were expanded in tissue culture prior to sequencing and so may be a consequence of the expansion procedure; this is discussed further below.

All genomes from Philadelphia were found to encode the D614G spike polymorphism suggested to promote efficient spread in humans (13, 21, 22). Philadelphia sequences also all encoded P314L in the virus-encoded RdRp (ORF1b), marking them as lineage B.1, Nextstrain clade 20A or 20C, GISAID clade G or GH, and clade A2a (16, 17). All genomes contained further polymorphisms distinguishing them from the USA-WA1-2020 reference isolate (Fig. 2).

SARS-CoV-2 genomes from Philadelphia were compared to global sequences at several time points in the epidemic. The lineage B.1/Nextstrain clade 20A or 20C/GISAID clade G or GH/A2a (16, 17) variants were circulating in New York (6) prior to expansion of the epidemic in Philadelphia. Figure 3 shows clustering of strains from Philadelphia with New York and global strains. To investigate the geographical origin of the epidemic in Philadelphia more carefully, each genome was aligned to database genomes and neighbors with the lowest edit distances were recorded (Table S4). In this analysis, only global isolates were selected for comparison that were reported prior to the date of symptom onset for each patient queried. Comparisons commonly showed nearest neighbors at multiple locations; in the text below, the majority location is emphasized.

For 22 of the 27 subjects, the most frequently identified closest-matched database genomes were from subjects in New York. Other Philadelphia genomes showed most frequent best alignments to sequences from Massachusetts (subject 223), Sweden (subject 242), California (subjects 263 and 307), and New Jersey (subject 266). In these comparisons, mismatches between Philadelphia and global sequences ranged from 0 to 10 substitutions (mean = 2.3). None of the subjects with sequences linked to Sweden, Massachusetts, or California had known direct contact with these locations prior to illness, suggesting community spread as the proximal source.

To assess the contribution of local circulation of lineages, our genome sequences from Philadelphia were compared to each other and polymorphisms were assessed (Table S4). Genomes differed from their nearest neighbor within this data set by 0 to 9 substitutions (mean = 1.9). In 14 of 27 of these cases, the match to another Philadelphia sequence was closer than the match to genomes from any other location. Analysis is complicated by incomplete sampling at all sites, but these observations are consistent with circulation of closely related lineages within the Philadelphia community.

Thus, the picture that emerges is that the first wave of the epidemic in Philadelphia was mostly introduced from New York, with additional less prominent introductions from a few other sites, and subsequently driven by spread due to circulation within the Philadelphia community. We attempted to specify the geographic origin of infection chains more precisely by analyzing possible clustering of the tree in Fig. 2 by subject zip code but did not find any significant clustering (permutational multivariate analysis of variance [PERMANOVA] *P* value of 0.6).

**Lack of association between viral polymorphisms and patient outcomes.** We next assessed possible associations of viral polymorphisms versus patient disease severity based on WHO scores for maximum severity reached (23) and outcomes. All patients studied (*n* = 27) were hospitalized (WHO score ≥4 [23]) and were grouped as moderate disease (nonintubated; maximum WHO score 4 to 6; *n* = 6), severe disease (intubated; WHO score 7 to 9; *n* = 13), or fatal outcomes (WHO score 10; *n* = 8) (Table 1 and Table S1). Polymorphisms were compared to severity (moderate/severe/fatal) and final outcomes (survivor/nonsurvivor) using PERMANOVA, which queries global associ-

**FIG 2** Legend (Continued)
left; a map of the SARS-CoV-2 genome with sequence polymorphisms indicated is shown to the right. Each row indicates a single genome, and each column indicates a nucleotide position. The dots indicate sequence polymorphisms; each is color coded by the SARS-CoV-2 gene in which it was found (code at bottom). Patient outcomes are encoded by the column of symbols between the tree and the map of polymorphisms, coded as follows: open circle, survived infection; filled circle, fatal outcome. The sequence of the USA-WA1-2020 isolate was verified twice independently by sequencing.

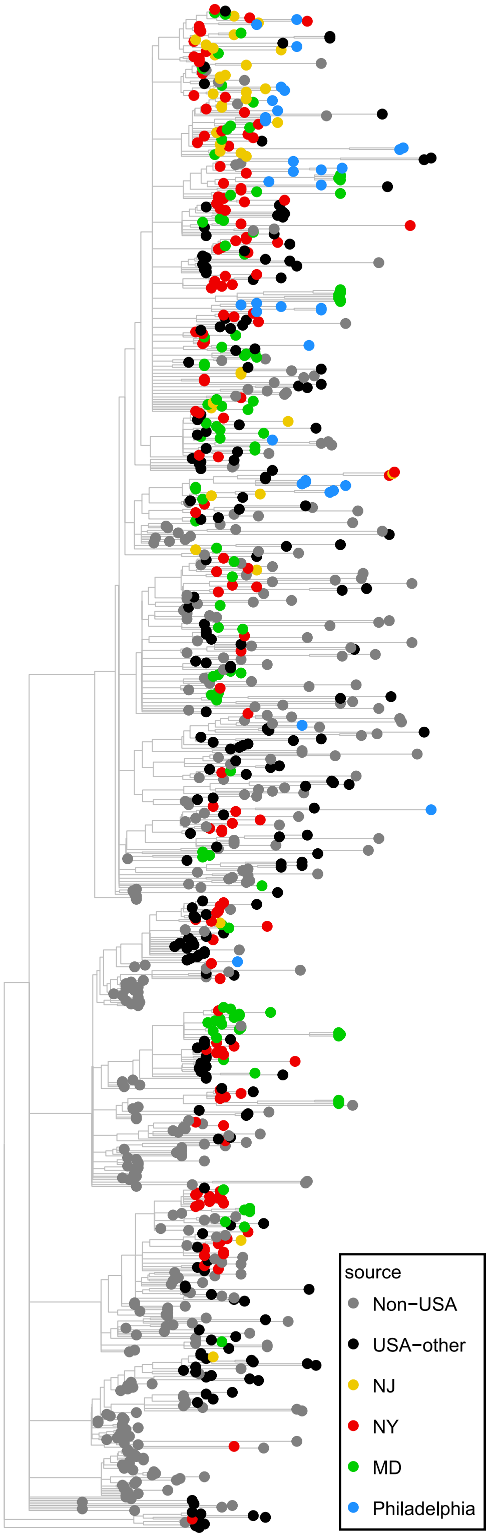


**FIG 3** SARS-CoV-2 isolates in the global context. Isolates are indicated by the global site of origin. The phylogenetic tree shows isolates from 30 March 2020 and later, including the lineages sampled here. Phylogenetic trees were generated via the IQ-TREE algorithm (48). The geographic origin of selected sequence isolates is shown by the color code.

ations of the sequence-based tree with outcome. No significant associations were found ($P = 0.38$ for WHO score and $P = 0.06$ for survival; prior to correction for multiple comparisons). Specific polymorphisms were next queried using Fisher's exact test comparing polymorphisms versus outcomes. The closest to significance were a set of 4 single nucleotide polymorphisms (SNPs) (C18998T, C23230T, G29540A, and T1918C) that are shared between two patients (211 and 222) who both died (Fisher's *P* value = 0.074; prior to correction for multiple comparison). Three of the polymorphisms are synonymous, while one caused A1844V in ORF1b. Thus, we did not detect viral polymorphisms that significantly increased or decreased pathogenic potential in our cohort, paralleling previous work (18).

**Longitudinal variation of viral sequences within subjects.** We next investigated possible longitudinal variation (Fig. 4 and Table S5). We obtained high-quality complete genome sequences from the same body site at more than one time point for 8 patients. Intervals between samples ranged from 2 to 39 days. In the longest sampling period, subject 228 yielded identical viral genome sequences over 39 days. Identical genomes were recovered from NP, OP, NP-OP, and saliva samples over this period. Subject 228 is a 65-year-old male who reached WHO level 9 and was hospitalized for 55 days but ultimately survived to be discharged. This emphasizes that viral populations can show notable longitudinal stability in some subjects.

In four of eight cases, sequence data showed polymorphisms between time points. To verify the authenticity of polymorphisms, we repeated the sequencing procedure on an independent aliquot of the samples (Fig. 4, "replicate" designation), validating reproducible polymorphisms in all four cases. One polymorphism resulted in a silent substitution in the spike coding region, and the others resulted in amino acid coding changes in the large orf1ab gene. In all four cases, longitudinal variation was seen in upper respiratory tract samples, and not in the two ETA samples queried. In most cases, the variant present in the earlier samples was evident as a minor variant in the later samples, suggesting that standing viral populations commonly encode multiple variants at single loci and that proportions can change over time.

Either of two processes could account for the observed variation. The data are consistent with the idea that the virus is mutating within subjects, and rapid cell turnover and particle washout result in appearance of new variants. The alternative is that the subjects were infected initially with viral populations containing multiple variants, and different variants became predominant at different times during infection.

**Differences in viral genome sequences between body sites.** Another question turns on whether there are separately evolving viral populations at different body sites. We were able to investigate 10 pairs of genomes from different body sites at the same time point (Table S6). Comparisons include different upper airway sites and upper versus lower airway. In 3 out of 10 cases, we identified polymorphisms differing at different body sites at the same time point (Fig. 5). In several cases, the polymorphic variant present in one of the sites was evident as a minor variant in the other site. In all three cases, the polymorphisms involved comparison of upper respiratory tract samples to an endotracheal aspirate. For the comparison of subject 211, there is a possible alternative explanation—the viruses compared were grown out first in tissue culture, so variants could have alternatively accumulated at this step. Nevertheless, these data together suggest possible independently replicating viral populations along the respiratory tract with incomplete intermixing in some subjects. As with the longitudinal data, variation could be the result of either *de novo* mutation within subjects or initial infection with multiple variants.

**Variation associated with expansion in cell culture.** Nine of the viral isolates were expanded by growth in cell culture prior to sequencing, and for samples from two subjects (211 and 266), genomes were sequenced from patient samples before expansion as well. Samples were expanded by growth for 5 to 6 days on human A549 cells engineered to express the ACE2 viral receptor. Successfully expanded viral stocks were recovered from NP, OP, and ETA samples (Table S2).

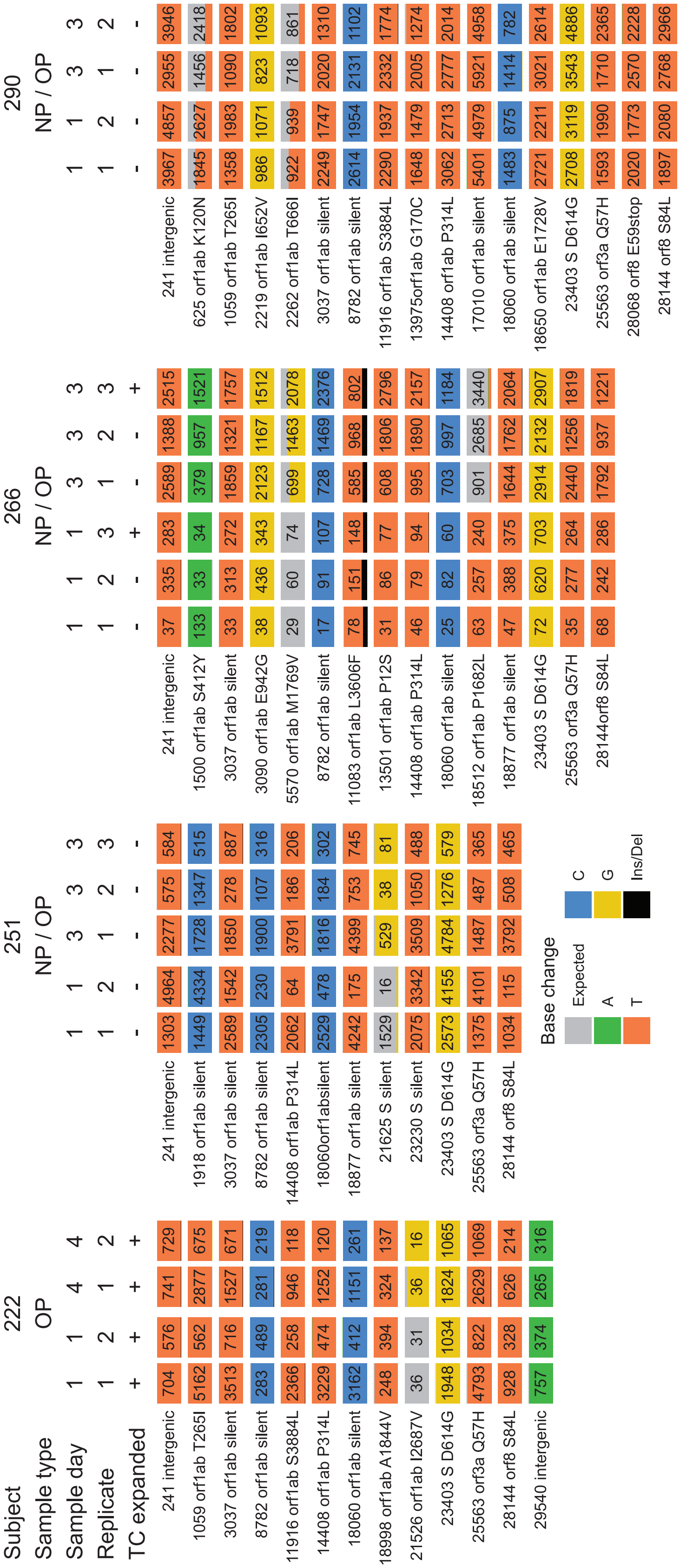
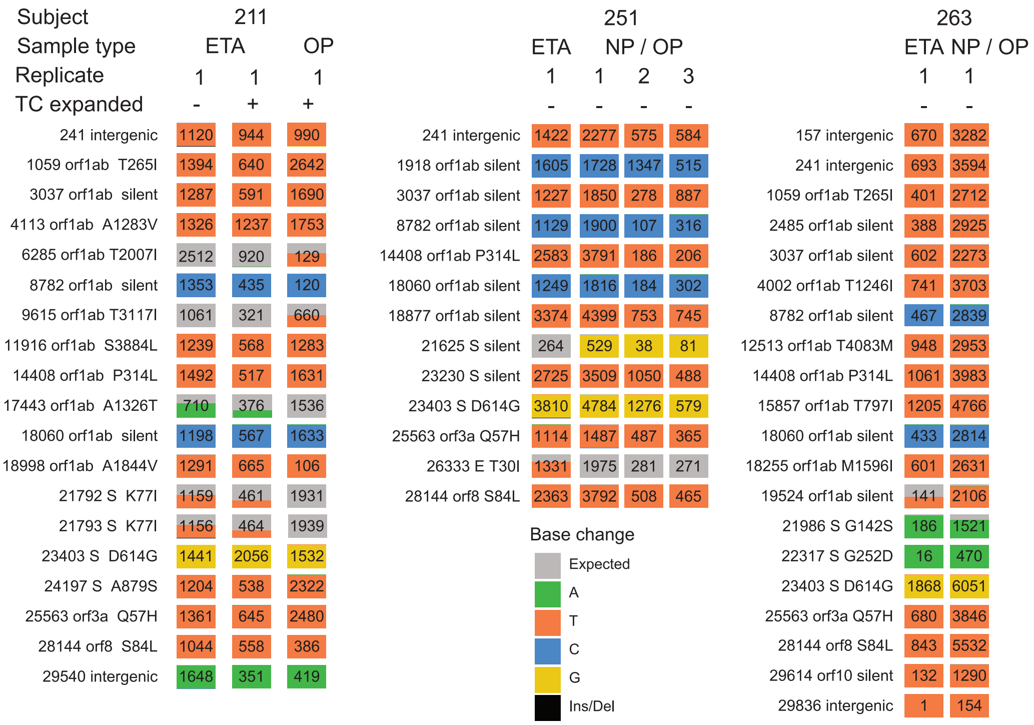


**FIG 4** Examples of longitudinal variation detected in SARS-CoV-2 genome sequences. The plots show base substitutions differing from the USA-WA1-2020 reference isolate as colored tiles. Each column

(Continued on next page)



| Subject | 211 | | |
|---|---|---|---|
| Sample type | ETA | | OP |
| Replicate | 1 | 1 | 1 |
| TC expanded | - | + | + |
| 241 intergenic | 🟧 1120 | 🟧 944 | 🟧 990 |
| 1059 orf1ab T265I | 🟧 1394 | 🟧 640 | 🟧 2642 |
| 3037 orf1ab silent | 🟧 1287 | 🟧 591 | 🟧 1690 |
| 4113 orf1ab A1283V | 🟧 1326 | 🟧 1237 | 🟧 1753 |
| 6285 orf1ab T2007I | ⬜ 2512 | ⬜ 920 | 🟧 129 |
| 8782 orf1ab silent | 🟦 1353 | 🟦 435 | 🟦 120 |
| 9615 orf1ab T3117I | ⬜ 1061 | ⬜ 321 | 🟧 660 |
| 11916 orf1ab S3884L | 🟧 1239 | 🟧 568 | 🟧 1283 |
| 14408 orf1ab P314L | 🟧 1492 | 🟧 517 | 🟧 1631 |
| 17443 orf1ab A1326T | ⬜🟩 710 | ⬜🟩 376 | ⬜ 1536 |
| 18060 orf1ab silent | 🟦 1198 | 🟦 567 | 🟦 1633 |
| 18998 orf1ab A1844V | 🟧 1291 | 🟧 665 | 🟧 106 |
| 21792 S K77I | ⬜🟧 1159 | ⬜ 461 | ⬜ 1931 |
| 21793 S K77I | 🟧⬜ 1156 | ⬜ 464 | ⬜ 1939 |
| 23403 S D614G | 🟨 1441 | 🟨 2056 | 🟨 1532 |
| 24197 S A879S | 🟧 1204 | 🟧 538 | 🟧 2322 |
| 25563 orf3a Q57H | 🟧 1361 | 🟧 645 | 🟧 2480 |
| 28144 orf8 S84L | 🟧 1044 | 🟧 558 | 🟧 386 |
| 29540 intergenic | 🟩 1648 | 🟩 351 | 🟩 419 |

| | 251 | | | |
|---|---|---|---|---|
| | ETA | NP / OP | | |
| | 1 | 1 | 2 | 3 |
| | - | - | - | - |
| 241 intergenic | 🟧 1422 | 🟧 2277 | 🟧 575 | 🟧 584 |
| 1918 orf1ab silent | 🟦 1605 | 🟦 1728 | 🟦 1347 | 🟦 515 |
| 3037 orf1ab silent | 🟧 1227 | 🟧 1850 | 🟧 278 | 🟧 887 |
| 8782 orf1ab silent | 🟦 1129 | 🟦 1900 | 🟦 107 | 🟦 316 |
| 14408 orf1ab P314L | 🟧 2583 | 🟧 3791 | 🟧 186 | 🟧 206 |
| 18060 orf1ab silent | 🟦 1249 | 🟦 1816 | 🟦 184 | 🟦 302 |
| 18877 orf1ab silent | 🟧 3374 | 🟧 4399 | 🟧 753 | 🟧 745 |
| 21625 S silent | ⬜ 264 | 🟨 529 | 🟨 38 | 🟨 81 |
| 23230 S silent | 🟧 2725 | 🟧 3509 | 🟧 1050 | 🟧 488 |
| 23403 S D614G | 🟨 3810 | 🟨 4784 | 🟨 1276 | 🟨 579 |
| 25563 orf3a Q57H | 🟧 1114 | 🟧 1487 | 🟧 487 | 🟧 365 |
| 26333 E T30I | ⬜🟧 1331 | ⬜ 1975 | ⬜ 281 | ⬜ 271 |
| 28144 orf8 S84L | 🟧 2363 | 🟧 3792 | 🟧 508 | 🟧 465 |

Base change

⬜ Expected
🟩 A
🟧 T
🟦 C
🟨 G
⬛ Ins/Del

| | 263 | |
|---|---|---|
| | ETA | NP / OP |
| | 1 | 1 |
| | - | - |
| 157 intergenic | 🟧 670 | 🟧 3282 |
| 241 intergenic | 🟧 693 | 🟧 3594 |
| 1059 orf1ab T265I | 🟧 401 | 🟧 2712 |
| 2485 orf1ab silent | 🟧 388 | 🟧 2925 |
| 3037 orf1ab silent | 🟧 602 | 🟧 2273 |
| 4002 orf1ab T1246I | 🟧 741 | 🟧 3703 |
| 8782 orf1ab silent | 🟦 467 | 🟦 2839 |
| 12513 orf1ab T4083M | 🟧 948 | 🟧 2953 |
| 14408 orf1ab P314L | 🟧 1061 | 🟧 3983 |
| 15857 orf1ab T797I | 🟧 1205 | 🟧 4766 |
| 18060 orf1ab silent | 🟦 433 | 🟦 2814 |
| 18255 orf1ab M1596I | 🟧 601 | 🟧 2631 |
| 19524 orf1ab silent | ⬜🟧 141 | 🟧 2106 |
| 21986 S G142S | 🟩 186 | 🟩 1521 |
| 22317 S G252D | 🟩 16 | 🟩 470 |
| 23403 S D614G | 🟨 1868 | 🟨 6051 |
| 25563 orf3a Q57H | 🟧 680 | 🟧 3846 |
| 28144 orf8 S84L | 🟧 843 | 🟧 5532 |
| 29614 orf10 silent | 🟧 132 | 🟧 1290 |
| 29836 intergenic | 🟧 1 | 🟧 154 |

**FIG 5** Examples of SARS-CoV-2 genome variation at different body sites at the same time of sampling. Markings are as in Fig. 4. The sample types are marked at the top of each column. Technical replicates of the sequencing procedure are shown separately.

For subject 211 (Fig. 5), five SNPs were identified that distinguish the tissue-culture expanded virus from the original ETA sample. Examination of the sequence reads showed that all five variants were detectable as minor populations that changed in proportion upon tissue culture, so that the majority form changed at the five loci. It is unknown whether this reflects superior replication of these variants under conditions of tissue culture or instead stochastics of sampling during the virus isolation procedure.

For subject 266 (Fig. 4), NP/OP samples from two dates were each expanded and compared to preexpansion sequences. In each case, the expanded sequence was identical to the preexpansion sequence.

**Possible consequences of the observed substitutions.** The data could be interrogated to investigate several additional aspects of viral evolution. The drug remdesivir is coming into wide use (24), and the binding site for remdesivir has been defined by structural analysis (25–27). Thus, we asked whether polymorphisms are accumulating in the binding site, suggestive of evolution to drug resistance. While only 3 of our subjects received remdesivir, such polymorphisms could indicate accumulation of resistant lineages circulating within the community sampled. However, no polymorphisms were detected in or near the region encoding the remdesivir binding site.

**FIG 4** Legend (Continued)

shows results of an independent determination of the genome sequence; the top of the column indicates (1) the subject of origin; (2) the sample type; (3) the relative days of sampling; (4) replicate number, i.e., indicating independent sequence determinations for the same patient sample; and (5) whether the virus was isolated and expanded in tissue culture ("TC expanded"). Results of technical replicates of the sequencing procedure are marked with numbers to indicate independent sequence determinations for the same initially isolated patient sample. The numbers of sequence reads contributing to the sequence call are marked on each tile. Tiles with more than one color indicate the presence of minor sequence variants at that position. The key to tile colors is at the bottom; gray indicates a match to the consensus USA-WA1-2020; Ins/Del indicates insertion/deletion.

Koonin and coworkers identified regions of coronavirus genomes that they proposed to be associated with increased pathogenesis in humans, including regions encoding the spike protein and subcellular sorting motifs in the nucleoprotein (28). No polymorphisms were found in any of these sites, providing no support to the idea that SARS-CoV-2 is evolving to be more pathogenic in these subjects by these pathways.

The furin cleavage site in the spike protein has been proposed to be a locus of evolution in coronaviruses (29–32). However, no changes were observed in the region encoding the furin cleavage site in the genomes studied here.

ORF8 encodes a protein suggested to promote immune evasion by downregulating major histocompatibility complex (MHC) (33), and previously mutations in ORF8 were found and suggested to be associated with reduced severity of disease or altered immune evasion (34, 35). We found an example of a genome that in fact contained a stop codon in ORF8 (Fig. 4; subject 290). We resequenced this genome and verified that the substitution was indeed present. The possible consequences for replication are unknown; however, we note that subject 290 reached a WHO score of only 4 (hospitalized without supplemental oxygen) and survived to be discharged from the hospital, consistent with possible attenuation by the ORF8 substitution.

**Polymorphisms potentially disrupting SARS-CoV-2 detection.** Lastly, we checked whether commonly used reverse transcription-PCR (RT-PCR) and reverse transcription–loop-mediated isothermal amplification (RT-LAMP) primer sets for detecting SARS-CoV-2 (Table S7) were fully matched with Philadelphia isolates, or whether polymorphisms might disrupt viral detection (36, 37). For this, we compared the CDC RT-PCR primers and several widely used RT-LAMP primer sets (38, 39). No polymorphisms were found in binding sites for the CDC RT-PCR primers. Of the three RT-LAMP primer sets studied, each did have at least one primer for which binding would be disrupted by a patient polymorphism; in each case, the polymorphism was found in one subject only. Thus, we conclude that all of these primer sets are suitable for detecting the great majority of SARS-CoV-2 lineages in Philadelphia, but in rare instances sensitivity for some may be decreased by target site polymorphisms.

## DISCUSSION

Here, we present an analysis of the SARS-CoV-2 epidemic in Philadelphia using viral whole-genome sequencing of 52 isolates from 27 hospitalized patients. We find that all of the viral genomes recovered contained the spike D614G substitution suggested to promote spread among humans, and all contained the linked RdRp P314L substitution, marking them as the lineage B.1, Nextstrain clade 20A or 20C, GISAID clade G or GH, and clade A2a (16, 17). The majority of these genomes had nearest neighbors outside Philadelphia that were most commonly from New York, providing a probable origin for much of the outbreak in Philadelphia. In a few cases, genomes had best matches to genomes from other locations, consistent with additional independent introductions. When comparing sequences within Philadelphia, many were found to have even closer matches to local sequences, suggesting community transmission chains. We thus propose that the epidemic in Philadelphia was seeded primarily from New York, followed by local spread. Although power to detect differences was small, in no case did any polymorphism correlate with patient outcome, consistent with results of others (18).

Use of whole-genome sequencing to postulate transmission chains involves several approximations, so that conclusions must be taken as likely scenarios and not strictly established. When aligning the Philadelphia isolates to database genomes, it was common to find multiple equally good closest matches, and usually this collection contained isolates from several different locations. In the interpretation, we focused on the most frequently seen location, typically New York, but it is not ruled out that transmission could have been from one of the other less frequently captured locations with high-homology genomes. A bias in this analysis is that sampling effort is not distributed equally at different global sites, resulting in potentially more frequent identification of isolates from heavily sampled regions. In addition, the sharing of two sequences between different locations at different times is taken to suggest transmission from one location to the other, but both might have been infected independently from some third location. Our interpretation nevertheless seems warranted—the finding of nearest external neighbors in New York was the case for many of our genomes, and similarly the frequency of still more homologous within-Philadelphia best matches was also high. Thus, the data support a model of infection primarily from New York followed by local spread.

Our study recovered multiple complete genome sequences from individual patients at different times after infection, allowing assessment of within-host sequence heterogeneity. To exclude possible error in sequence determination, where possible we resequenced virus with possible polymorphisms from an independent sample aliquot to validate the variants detected. In four out of eight of the subjects analyzed, reproducible polymorphisms were detected. This finding implies either that populations of SARS-CoV-2 are accumulating substitutions and turning over at a high rate or else that patients were initially infected with multiple variants and different variants predominated at different times. High mutation rates are well known in many RNA viruses such as HIV and hepatitis C virus (HCV) (40).

We were also able to study 10 cases where high-quality genomes were available from different body sites at the same time in the same individual. Of these, three showed polymorphisms. Although the numbers are small, this suggests that there can be heterogeneity of viral populations at different body sites and limited exchange between sites. All three cases involved comparison of lung (endotracheal aspirate) samples to upper respiratory tract samples, potentially consistent with distinct viral populations replicating in the upper and lower respiratory sites in these patients. As with the analysis of longitudinal variation, it is unknown whether this reflects initial infection with a mixed virus population or accumulation of sequence variation during growth within subjects.

This study has several limitations. All subjects studied were hospitalized, leaving open the possibility that different genotypes predominate in less sick subjects. We analyzed small numbers of subjects with genomes in samples collected longitudinally or contemporaneously at multiple body sites—it will be valuable to examine more subjects to determine whether the extent of polymorphism seen here is reproducible in other cohorts. Happenstance of sampling may have also affected recovery of sequence polymorphism—for example, if different nostrils were sampled at different times, a temporal difference might be inferred that actually represented partitioning by body site. We note that these observations of potential differences in viral populations in time and space should be amenable to further investigation using experimental infections in model organisms.

In summary, our complete genome sequence analysis indicates that the SARS-CoV-2 epidemic in Philadelphia was primarily seeded from New York, which experienced an earlier expansion of COVID-19, followed by local spread. We were able to survey several subjects longitudinally, in some cases observing acquisition of viral polymorphisms that imply either initial infection with heterogeneous viral populations or accumulation of variants during growth within subjects. We also saw examples of polymorphisms in different body sites in some subjects, suggesting that populations may distribute in part independently at different anatomical locations. These observations present new hypotheses for viral dynamics that should be readily amenable to further study.

## MATERIALS AND METHODS

**Human subjects.** Following informed consent obtained under protocol no. 823392 approved by the University of Pennsylvania IRB, samples were collected beginning within 2 days of hospitalization. Oropharyngeal (OP) and nasopharyngeal (NP) swabs were obtained using flocked swabs (Copan Diagnostics) eluted in 1.5 to 3 ml of viral transport medium. In some instances, OP and NP samples were eluted together (NP-OP). Saliva and endotracheal aspirate samples were obtained from nonintubated or intubated patients, respectively. Patients were classified clinically based on survival to discharge or in-hospital mortality and on the maximum score reached during hospitalization based on the 11-point WHO clinical COVID-19 progression scale (23).

**RT-qPCR to detect SARS-CoV-2.** RNA was extracted from 140 $\mu$l of clinical sample (swab eluate or neat endotracheal aspirate or saliva) using the Qiagen QIAamp viral RNA minikit. The RT-quantitative PCR (qPCR) assay targeted the SARS-CoV-2 nucleocapsid region using the CDC 2019-nCoV_N1 primer-probe set (2019-nCoV_N1-F, GACCCCAAAATCAGCGAAAT; 2019-nCoV_N1-R, TCTGGTTACTGCCAGTTGAATCTG; 2019_nCoV_N1-P, 6-carboxyfluorescein (FAM)-ACCCCGCATTACGTTTGGTGGACC-Iowa black fluorescent quencher (IBFQ). The RT-qPCR master mix was prepared according to the following protocol: 8.5 $\mu$l distilled water ($dH_2O$), 0.5 $\mu$l N1-F (20 $\mu$M), 0.5 $\mu$l N1-R (20 $\mu$M), 0.5 $\mu$l N1-P (5 $\mu$M), 5.0 $\mu$l TaqMan Fast Virus 1-Step master mix per reaction. Five microliters of extracted RNA was added to 15 $\mu$l of prepared master mix for a final volume of 20 $\mu$l per reaction. Final concentrations of both 2019-nCoV_N1-F and 2019-nCoV_N1-R primers were 500 nM, and the final concentration of the 2019-nCoV_N1-P probe was 125 nM as suggested by the CDC protocol. The assay was performed using the Applied Biosystems QuantStudio 5 real-time PCR system. The thermocycler conditions were as follows: 5 min at 50°C, 20 s at 95°C, and 40 cycles of 3 s at 95°C and 30 s at 60°C.

**Cells.** Human A549 cells expressing ACE2, constructed by lentivirus transduction of *hACE2* (41), were cultured in RPMI 1640 (Gibco catalog no. 11875) supplemented with 10% fetal bovine serum (FBS), 100 U/ml of penicillin, and 100 $\mu$g/ml streptomycin.

**Viral isolation.** Nine viral isolates from patients were expanded in cell culture prior to sequencing. Successfully expanded viral stocks were recovered from NP, OP, and ETA samples. Briefly, NP or OP swabs were incubated with 1 ml of viral isolation medium (DMEM [Gibco catalog no. 11965] with 200 U penicillin and 200 $\mu$g/ml streptomycin) for 1 h at room temperature. For ETA samples, 500 $\mu$l or 100 $\mu$l of eluate was mixed with 500 or 900 $\mu$l viral isolation medium and then inoculated onto A549$^{ACE2}$ cells in 48-well plates. After 1 h of inoculation, the media were removed from the wells, and 1 ml of culture medium (RPMI 1640 [Gibco catalog no. 11875] with 2% FBS and 200 U penicillin and 200 $\mu$g/ml streptomycin) was added to each well. Three to 4 days postinfection, supernatants were harvested and 300 $\mu$l was used to inoculate A549$^{ACE2}$ cells in 6-well plates. Forty-eight hours postinfection, the supernatants were collected, the cells were lysed using RLT Plus lysis buffer, and the RNA was extracted using the RNeasy Plus minikit (Qiagen). For viral isolates, please contact Susan Weiss.

**Viral genome sequence acquisition.** Most viral genome sequencing was carried out using the POLAR protocol (19). For each sample, 1 $\mu$l to 5 $\mu$l of viral RNA was used along with 0.5 $\mu$l of 10 mM deoxynucleoside triphosphate (dNTP) mix (Thermo Fisher, 18427013), 0.5 $\mu$l of 50 $\mu$M random hexamers (Thermo Fisher, N8080127), and additional nuclease-free water to reach a total volume of 6.5 $\mu$l. The mixture was incubated at 65°C for 5 min followed by a 1-min incubation at 4°C. Reverse transcription was performed by the addition of 0.5 $\mu$l SuperScript III reverse transcriptase (Thermo Fisher, 18080085), 2 $\mu$l of 5× First-Strand buffer (Thermo Fisher, 18080085), 0.5 $\mu$l of RNaseOUT (Thermo Fisher, 18080051), and 0.5 $\mu$l of 0.1 M dithiothreitol (DTT) (Thermo Fisher, 18080085). The reverse transcription mixture was incubated at 42°C for 50 min and 70°C for 10 min and then held at 4°C. To amplify the cDNA, we used the artic-ncov2019 version 3 primers designed by the ARTIC Network. The two pooled ARTIC primer sets were provided by IDT. For the PCR, 5 $\mu$l of 5× Q5 reaction buffer (NEB, M0493S), 0.5 $\mu$l of 10 mM dNTP mix (NEB, N0447S), 0.25 $\mu$l Q5 Hot Start DNA polymerase (NEB, M0493S), and either 4.0 $\mu$l of pooled primer set 1 or 3.98 $\mu$l of pooled primer set 2 were prepared for each reaction. Additionally, 12.7 $\mu$l or 12.8 $\mu$l of nuclease-free water and 2.5 $\mu$l of cDNA were added to reach a total volume of 25 $\mu$l. The reaction mixture was incubated at 98°C for 30 s for 1 cycle, followed by 25 cycles at 98°C for 15 s and 65°C for 5 min, and then held at 4°C. Prior to using the ARTIC nCoV-2019 amplicon sequencing protocol, we designed 12 primers that amplified six regions along the SARS-CoV-2 genome. In these early experiments, the random hexamers and ARTIC primers were replaced with our genome-specific primers.

PCR products of the same genome that were generated from the two primer sets were pooled. A 1:1 volume AMPure XP (Beckman Coulter, A63881) bead cleanup using 80% ethanol washes and 15 $\mu$l elution in nuclease-free water was carried out before quantifying the DNA content with the Qubit dsDNA HS assay kit (Thermo Fisher, Q32851). PCR products were diluted to 0.25 ng/$\mu$l, and the Nextera library was prepared using the Nextera XT library prep kit (kit (Illumina, FC-131-1096) and the Nextera XT Index kit (Illumina, FC-131-2001, FC-131-2002). The DNA tagmentation reaction, adaptor ligation, and amplification for the library followed the protocols in the Nextera XT library prep reference guide provided by Illumina. Following the Nextera amplification, a second 1:1 volume AMPure XP bead cleanup was done, and samples were eluted in 35 $\mu$l of nuclease-free water. RT-PCR was used to quantify the DNA of each sample using the KAPA SYBR Fast Universal qPCR kit (Roche, KK4903, KK4622). The samples were pooled in equal quantities, and an additional RT-PCR was performed on the pooled library. The library was sequenced on an Illumina MiSeq.

**Genome analysis.** A custom informatics pipeline was created to align quality trimmed reads to the USA-WA1-2020 reference genome with the Burrows-Wheeler aligner and create read pile-ups and call variants with the SAMtools/bcftools software packages (42–45). Variants were accepted if they possessed a PHRED score of $\geq$20, the position had a read depth of at least 5 reads, and $>$50% of the reads included the variant. The pipeline created data objects for each sequenced sample and subsequently compiled these data objects into detailed subject reports including variant heatmaps and read pileups for both individual experiments and compilations of multiple experiments. One genome sequence was excluded because repeated attempts at sequence acquisition failed to yield a consensus sequence. Genomes from another subject were excluded when parallel analysis of HLA-A and HLA-B RNA failed to confirm that samples were from the same subject. The phylogenetic tree of local genomes shown in Fig. 2 was created via hierarchical cluster analysis using the UPGMA (unweighted pair group method with arithmetic mean) algorithm (46) where global genome alignments were first created with the MAFFT alignment software package (47). Maximum-likelihood phylogenies of global genome representatives shown in Fig. 3 (sampling of PANGOLIN lineages) were built with the IQ-TREE software package

(48), implemented through the Augur informatics package (49), which employs a combination of hill-climbing and stochastic perturbation methods to build trees from large data sets.

**Data availability.** All sequence data acquired in this study are available at NCBI under MW001232 to MW001286; Table S2 in the supplemental material relates the sample metadata to sequence accession numbers. Computer code used is available at https://doi.org/10.5281/zenodo.4046252. A detailed list of materials used is in Table S8.

## SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

**TABLE S1**, XLSX file, 0.01 MB.

**TABLE S2**, XLSX file, 0.01 MB.

**TABLE S3**, XLSX file, 0.01 MB.

**TABLE S4**, XLSX file, 0.01 MB.

**TABLE S5**, XLSX file, 0.01 MB.

**TABLE S6**, XLSX file, 0.01 MB.

**TABLE S7**, XLSX file, 0.01 MB.

**TABLE S8**, XLSX file, 0.01 MB.

## ACKNOWLEDGMENTS

We are grateful to patients and their families who volunteered to provide specimens; Rebecca Trotta, Robin Hermann, and the University of Pennsylvania nursing staff who assisted with sample collection; members of the Bushman and Collman laboratories for help and suggestions; and Laurie Zimmerman for artwork and help with manuscript preparation. We acknowledge help from all the staff of the Philadelphia Department of Public Health who conducted surveillance on COVID and provided the data in Fig. 1A.

Funding was provided by philanthropic donations to the Penn Center for Research on Coronaviruses and Other Emerging Pathogens, and in part by NIH grant R61/33-HL137063 and AI140442—supplement for SARS-CoV-2. Additional assistance was provided by the Penn Center for AIDS Research (P30-AI045008).

J.E., P.H., R.G.C., and F.B. designed the study; R.G.C., J.G.-W., L.K., A.M.R., S.A.W., and A.S.F. managed acquisition of clinical specimens; Y.L. and S.R.W. isolated replication competent viral stocks; P.H., A.M.R., S.R., Y.H., L.K., K.M., S.S.-M., and A.G. carried out sequencing, biochemical and bioinformatic analysis; J.E., P.H., R.G.C., and F.B. wrote the paper.

We declare that we have no competing interests.

## REFERENCES

1. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395:565–574. https://doi.org/10.1016/S0140-6736(20)30251-8.
2. Morawska L, Milton DK. 2020. It is time to address airborne transmission of COVID-19. Clin Infect Dis 71:2311–2313. https://doi.org/10.1093/cid/ciaa939.
3. Morawska L, Cao J. 2020. Airborne transmission of SARS-CoV-2: the world should face the reality. Environ Int 139:105730. https://doi.org/10.1016/j.envint.2020.105730.
4. van Doremalen N, Bushmaker T, Morris DH, Holbrook MG, Gamble A, Williamson BN, Tamin A, Harcourt JL, Thornburg NJ, Gerber SI, Lloyd-Smith JO, de Wit E, Munster VJ. 2020. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. N Engl J Med 382:1564–1567. https://doi.org/10.1056/NEJMc2004973.
5. Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, Watts AG, Vogels CBF, Brito AF, Alpert T, Muyombwe A, Razeq J, Downing R, Cheemarla NR, Wyllie AL, Kalinich CC, Ott IM, Quick J, Loman NJ, Neugebauer KM, Greninger AL, Jerome KR, Roychoudhury P, Xie H, Shrestha L, Huang M-L, Pitzer VE, Iwasaki A, Omer SB, Khan K, Bogoch II, Martinello RA, Foxman EF, Landry ML, Neher RA, Ko AI, Grubaugh ND. 2020. Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. Cell 181:990–996.e5. https://doi.org/10.1016/j.cell.2020.04.021.
6. Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, Fabre S, Kleiner G, Polanco J, Khan Z, Alburquerque B, van de Guchte A, Dutta J, Francoeur N, Melo BS, Oussenko I, Deikus G, Soto J, Sridhar SH, Wang Y-C, Twyman K, Kasarskis A, Altman DR, Smith M, Sebra R, Aberg J, Krammer F, García-Sastre A, Luksza M, Patel G, Paniz-Mondolfi A, Gitman M, Sordillo EM, Simon V, van Bakel H. 2020. Introductions and early spread of SARS-CoV-2 in the New York City area. Science 369:297–301. https://doi.org/10.1126/science.abc1917.
7. Oude Munnink BB, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, Kamga SK, Schapendonk C, Pronk M, Lexmond P, van der Linden A, Bestebroer T, Chestakova I, Overmars RJ, van Nieuwkoop S, Molenkamp R, van der Eijk AA, GeurtsvanKessel C, Vennema H, Meijer A, Rambaut A, van Dissel J, Sikkema RS, Timen A, Koopmans M, The Dutch-Covid-19 response team. 2020. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nat Med 26:1405–1410. https://doi.org/10.1038/s41591-020-0997-y.
8. Moreno GK, Braun KM, Riemersma KK, Martin MA, Halfmann PJ, Crooks CM, Prall T, Baker D, Baczenas JJ, Heffron AS, Ramuta M, Khubbar M, Weiler AM, Accola MA, Rehrauer WM, O'Connor SL, Safdar N, Pepperell CS, Dasu T, Bhattacharyya S, Kawaoka Y, Koelle K, O'Connor DH, Friedrich TC. 2020. Distinct patterns of SARS-CoV-2 transmission in two nearby communities in Wisconsin, USA. medRxiv https://doi.org/10.1101/2020.07.09.20149104.
9. Rockett RJ, Arnott A, Lam C, Sadsad R, Timms V, Gray K-A, Eden J-S, Chang S, Gall M, Draper J, Sim EM, Bachmann NL, Carter I, Basile K, Byun R, O'Sullivan MV, Chen SC-A, Maddocks S, Sorrell TC, Dwyer DE, Holmes EC,

Kok J, Prokopenko M, Sintchenko V. 2020. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. Nat Med 26:1398–1404. https://doi.org/10.1038/s41591-020-1000-7.

10. Taboada B, Vazquez-Perez JA, Muñoz-Medina JE, Ramos-Cervantes P, Escalera-Zamudio M, Boukadida C, Sanchez-Flores A, Isa P, Mendieta-Condado E, Martínez-Orozco JA, Becerril-Vargas E, Salas-Hernández J, Grande R, González-Torres C, Gaytán-Cervantes FJ, Vazquez G, Pulido F, Araiza-Rodríguez A, Garcés-Ayala F, González-Bonilla CR, Grajales-Muñiz C, Borja-Aburto VH, Barrera-Badillo G, López S, Hernández-Rivas L, Perez-Padilla R, López-Martínez I, Ávila-Ríos S, Ruiz-Palacios G, Ramírez-González JE, Arias CF. 2020. Genomic analysis of early SARS-CoV-2 variants introduced in Mexico. J Virol 94:e01056-20. https://doi.org/10.1128/JVI.01056-20.
11. Candido DS, Claro IM, de Jesus JG, Souza WM, Moreira FRR, Dellicour S, Mellan TA, Du Plessis L, Pereira RHM, Sales FCS, Manuli ER, Thézé J, Almeida L, Menezes MT, Voloch CM, Fumagalli MJ, Coletti TM, da Silva CAM, Ramundo MS, Amorim MR, Hoeltgebaum HH, Mishra S, Gill MS, Carvalho LM, Buss LF, Prete CA, Ashworth J, Nakaya HI, Peixoto PS, Brady OJ, Nicholls SM, Tanuri A, Rossi ÁD, Braga CKV, Gerber AL, de C Guimarães AP, Gaburo N, Alencar CS, Ferreira ACS, Lima CX, Levi JE, Granato C, Ferreira GM, Francisco RS, Granja F, Garcia MT, Moretti ML, Perroud MW, Castiñeiras TMPP, Lazari CS, Hill SC, Santos AADS, Simeoni CL, Forato J, Sposito AC, Schreiber AZ, Santos MMN, de Sá CZ, Souza RP, Resende-Moreira LC, Teixeira MM, Hubner J, Leme PAF, Moreira RG, Nogueira ML, Brazil-UK Centre for Arbovirus Discovery, Diagnosis, Genomics and Epidemiology (CADDE) Genomic Network, Ferguson NM, Costa SF, Proenca-Modena JL, Vasconcelos ATR, Bhatt S, Lemey P, Wu C-H, Rambaut A, Loman NJ, Aguiar RS, Pybus OG, Sabino EC, Faria NR. 2020. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science 369:1255–1260. https://doi.org/10.1126/science.abd2161.
12. Li X, Giorgi EE, Marichann MH, Foley B, Xiao C, Kong XP, Chen Y, Korber B, Gao F. 2020. Emergence of SARS-CoV-2 through recombination and strong purifying selection. bioRxiv https://doi.org/10.1101/2020.03.20.000885.
13. Zhang L, Jackson CB, Mou H, Ojha A, Rangarajan ES, Izard T, Farzan M, Choe H. 2020. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. bioRxiv https://doi.org/10.1101/2020.06.12.148726.
14. Yurkovetskiy L, Pascal KE, Tompkins-Tinch C, Nyalile T, Wang Y, Baum A, Diehl WE, Dauphin A, Carbone C, Veinotte K, Egri SB, Schaffner SF, Lemieux JE, Munro J, Rafique A, Barve A, Sabeti PC, Kyratsous CA, Dudkina N, Shen K, Luban J. 2020. SARS-CoV-2 spike protein variant D614G increases infectivity and retains sensitivity to antibodies that target the receptor binding domain. bioRxiv https://doi.org/10.1101/2020.07.04.187757.
15. Daniloski Z, Guo X, Sanjana NE. 2020. The D614G mutation in SARS-CoV-2 spike increases transduction of multiple human cell types. bioRxiv https://doi.org/10.1101/2020.06.14.151357.
16. Banu S, Jolly B, Mukherjee P, Singh P, Khan S, Zaveri L, Shambhavi S, Gaur N, Mishra RK, Scaria V, Sowpati DT. 2020. A distinct phylogenetic cluster of Indian SARS-CoV-2 isolates. bioRxiv https://doi.org/10.1101/2020.05.31.126136.
17. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, Du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 5:1403–1407. https://doi.org/10.1038/s41564-020-0770-5.
18. Zhang X, Tan Y, Ling Y, Lu G, Liu F, Yi Z, Jia X, Wu M, Shi B, Xu S, Chen J, Wang W, Chen B, Jiang L, Yu S, Lu J, Wang J, Xu M, Yuan Z, Zhang Q, Zhang X, Zhao G, Wang S, Chen S, Lu H. 2020. Viral and host factors related to the clinical outcome of COVID-19. Nature 583:437–440. https://doi.org/10.1038/s41586-020-2355-0.
19. St Hilaire BG, Durand NC, Mitra N, Pulido SG, Mahajan R, Blackburn A, Colaric ZL, Theisen JWM, Weisz D, Dudchenko O, Gnirke A, Rao S, Kaur P, Aiden EL, Aiden AP. 2020. A rapid, low cost, and highly sensitive SARS-CoV-2 diagnostic based on whole genome sequencing. bioRxiv https://doi.org/10.1101/2020.04.25.061499.
20. Harcourt J, Tamin A, Lu X, Kamili S, Sakthivel SK, Murray J, Queen K, Tao Y, Paden CR, Zhang J, Li Y, Uehara A, Wang H, Goldsmith C, Bullock HA, Wang L, Whitaker B, Lynch B, Gautam R, Schindewolf C, Lokugamage KG, Scharton D, Plante JA, Mirchandani D, Widen SG, Narayanan K, Makino S, Ksiazek TG, Plante KS, Weaver SC, Lindstrom S, Tong S, Menachery VD, Thornburg NT. 2020. Isolation and characterization of SARS-CoV-2 from the first US COVID-19 patient. bioRxiv https://doi.org/10.1101/2020.03.02.972935.
21. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, de Silva TI, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC, Sheffield COVID-19 Genomics Group. 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182:812–827.e19. https://doi.org/10.1016/j.cell.2020.06.043.
22. Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, Zhao C, Zhang Q, Liu H, Nie L, Qin H, Wang M, Lu Q, Li X, Sun Q, Liu J, Zhang L, Li X, Huang W, Wang Y. 2020. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell 182:1284–1294.e9. https://doi.org/10.1016/j.cell.2020.07.012.
23. WHO Working Group on the Clinical Characterisation and Management of COVID-19 infection. 2020. A minimal common outcome measure set for COVID-19 clinical research. Lancet Infect Dis 20:e192–e197. https://doi.org/10.1016/S1473-3099(20)30483-7.
24. Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC. 2020. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review. JAMA 324:782–793. https://doi.org/10.1001/jama.2020.12839.
25. Wang Q, Wu J, Wang H, Gao Y, Liu Q, Mu A, Ji W, Yan L, Zhu Y, Zhu C, Fang X, Yang X, Huang Y, Gao H, Liu F, Ge J, Sun Q, Yang X, Xu W, Liu Z, Yang H, Lou Z, Jiang B, Guddat LW, Gong P, Rao Z. 2020. Structural basis for RNA replication by the SARS-CoV-2 polymerase. Cell 182:417–428.e13. https://doi.org/10.1016/j.cell.2020.05.034.
26. Hillen HS, Kokic G, Farnung L, Dienemann C, Tegunov D, Cramer P. 2020. Structure of replicating SARS-CoV-2 polymerase. Nature 584:154–156. https://doi.org/10.1038/s41586-020-2368-8.
27. Yin W, Mao C, Luan X, Shen D-D, Shen Q, Su H, Wang X, Zhou F, Zhao W, Gao M, Chang S, Xie Y-C, Tian G, Jiang H-W, Tao S-C, Shen J, Jiang Y, Jiang H, Xu Y, Zhang S, Zhang Y, Xu HE. 2020. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. Science 368:1499–1504. https://doi.org/10.1126/science.abc1560.
28. Gussow AB, Auslander N, Faure G, Wolf YI, Zhang F, Koonin EV. 2020. Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. Proc Natl Acad Sci U S A 117:15193–15199. https://doi.org/10.1073/pnas.2008176117.
29. Klimstra WB, Tilston-Lunel NL, Nambulli S, Boslett J, McMillen CM, Gilliland T, Dunn MD, Sun C, Wheeler SE, Wells A, Hartman AL, McElroy AK, Reed DS, Rennick LJ, Duprex WP. 2020. SARS-CoV-2 growth, furin-cleavage-site adaptation and neutralization using serum from acutely infected, hospitalized COVID-19 patients. bioRxiv https://doi.org/10.1101/2020.06.19.154930.
30. Wrobel AG, Benton DJ, Xu P, Roustan C, Martin SR, Rosenthal PB, Skehel JJ, Gamblin SJ. 2020. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nat Struct Mol Biol 27:763–767. https://doi.org/10.1038/s41594-020-0468-7.
31. Johnson BA, Xie X, Kalveram B, Lokugamage KG, Muruato A, Zou J, Zhang X, Juelich T, Smith JK, Zhang L, Bopp N, Schindewolf C, Vu M, Vanderheiden A, Swetnam D, Plante JA, Aguilar P, Plante KS, Lee B, Weaver SC, Suthar MS, Routh AL, Ren P, Ku Z, An Z, Debbink K, Shi PY, Freiberg AN, Menachery VD. 2020. Furin cleavage site is key to SARS-CoV-2 pathogenesis. bioRxiv https://doi.org/10.1101/2020.08.26.268854.
32. Xing Y, Li X, Gao X, Dong Q. 2020. Natural polymorphisms are present in the furin cleavage site of the SARS-CoV-2 spike glycoprotein. Front Genet 11:783. https://doi.org/10.3389/fgene.2020.00783.
33. Zhang Y, Zhang J, Chen Y, Luo B, Yuan Y, Huang F, Yang T, Yu F, Liu J, Liu B, Song Z, Chen J, Pan T, Zhang X, Li Y, Li R, Huang W, Xiao F, Zhang H. 2020. The ORF8 protein of SARS-CoV-2 mediates immune evasion through potently downregulating MHC-I. bioRxiv https://doi.org/10.1101/2020.05.24.111823.
34. Young BE, Fong SW, Chan YH, Mak TM, Ang LW, Anderson DE, Lee CY-P, Amrun SN, Lee B, Goh YS, Su YCF, Wei WE, Kalimuddin S, Chai LYA, Pada S, Tan SY, Sun L, Parthasarathy P, Chen YYC, Barkham T, Lin RTP, Maurer-Stroh S, Leo Y-S, Wang L-F, Renia L, Lee VJ, Smith GJD, Lye DC, Ng LFP. 2020. Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. Lancet 396:603–611. https://doi.org/10.1016/S0140-6736(20)31757-8.
35. Su YCF, Anderson DE, Young BE, Linster M, Zhu F, Jayakumar J, Zhuang Y, Kalimuddin S, Low JGH, Tan CW, Chia WN, Mak TM, Octavia S, Chavatte J-M, Lee RTC, Pada S, Tan SY, Sun L, Yan GZ, Maurer-Stroh S, Mendenhall IH, Leo Y-S, Lye DC, Wang L-F, Smith GJD. 2020. Discovery and genomic characterization of a 382-nucleotide deletion in ORF7b and ORF8 during the early evolution of SARS-CoV-2. mBio 11:e01610-20. https://doi.org/10.1128/mBio.01610-20.

36. Vanaerschot M, Mann SA, Webber JT, Kamm J, Bell SM, Bell J, Hong SN, Nguyen MP, Chan LY, Bhatt KD, Tan M, Detweiler AM, Espinosa A, Wu W, Batson J, Dynerman D, CLIAHUB Consortium, Wadford DA, Puschnik AS, Neff N, Ahyong V, Miller S, Ayscue P, Tato CM, Paul S, Kistler A, DeRisi JL, Crawford ED. 2020. Identification of a polymorphism in the N gene of SARS-CoV-2 that adversely impacts detection by RT-PCR. J Clin Microbiol https://doi.org/10.1128/JCM.02369-20.
37. Artesi M, Bontems S, Göbbels P, Franckh M, Maes P, Boreux R, Meex C, Melin P, Hayette M-P, Bours V, Durkin K. 2020. A recurrent mutation at position 26340 of SARS-CoV-2 is associated with failure of the E gene quantitative reverse transcription-PCR utilized in a commercial dual-target diagnostic assay. J Clin Microbiol 58:e01598-20. https://doi.org/10.1128/JCM.01598-20.
38. Rabe BA, Cepko C. 2020. SARS-CoV-2 detection using an isothermal amplification reaction and a rapid, inexpensive protocol for sample inactivation and purification. medRxiv https://doi.org/10.1073/pnas.2011221117.
39. El-Tholoth M, Bau HH, Song J. 2020. A single and two-stage, closed-tube, molecular test for the 2019 novel coronavirus (COVID-19) at home, clinic, and points of entry. ChemRxiv https://doi.org/10.26434/chemrxiv.11860137.
40. Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9:267–276. https://doi.org/10.1038/nrg2323.
41. Li Y, Renner DM, Coumar CE, Whelan JN, Reyes HM, Cardenas-Diaz FL, Truitt R, Tan LH, Dong B, Alysandratos KD, Huang J, Palmer JN, Adappa ND, Kohanski MA, Kotton DN, Silverman RH, Yang W, Morrisey E, Cohen NA, Weiss SR. 2020. SARS-CoV-2 induces double-stranded RNA-mediated innate immune responses in respiratory epithelial derived cells and cardiomyocytes. bioRxiv https://doi.org/10.1101/2020.09.24.312553.
42. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
43. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123. https://doi.org/10.1093/bioinformatics/bty407.
44. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
45. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330.
46. Michener CD, Sokal RR. 1957. A quantitative approach to a problem of classification. Evolution 11:130–139. https://doi.org/10.1111/j.1558-5646.1957.tb02884.x.
47. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
48. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300.
49. Billion A, Ghai R, Chakraborty T, Hain T. 2006. Augur—a computational pipeline for whole genome microbial surface protein prediction and classification. Bioinformatics 22:2819–2820. https://doi.org/10.1093/bioinformatics/btl466.