
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

# Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study

*Luke W Meredith*†, William L Hamilton*†, Ben Warne, Charlotte J Houldcroft, Myra Hosmillo, Aminu S Jahun, Martin D Curran, Surendra Parmar, Laura G Caller, Sarah L Caddy, Fahad A Khokhar, Anna Yakovleva, Grant Hall, Theresa Feltwell, Sally Forrest, Sushmita Sridhar, Michael P Weekes, Stephen Baker, Nicholas Brown, Elinor Moore, Ashley Popay, Iain Roddick, Mark Reacher, Theodore Gouliouris, Sharon J Peacock, Gordon Dougan, M Estée Török*‡, Ian Goodfellow*‡*

## Summary

**Background** The burden and influence of health-care associated severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections is unknown. We aimed to examine the use of rapid SARS-CoV-2 sequencing combined with detailed epidemiological analysis to investigate health-care associated SARS-CoV-2 infections and inform infection control measures.

**Methods** In this prospective surveillance study, we set up rapid SARS-CoV-2 nanopore sequencing from PCR-positive diagnostic samples collected from our hospital (Cambridge, UK) and a random selection from hospitals in the East of England, enabling sample-to-sequence in less than 24 h. We established a weekly review and reporting system with integration of genomic and epidemiological data to investigate suspected health-care associated COVID-19 cases.

**Findings** Between March 13 and April 24, 2020, we collected clinical data and samples from 5613 patients with COVID-19 from across the East of England. We sequenced 1000 samples producing 747 high-quality genomes. We combined epidemiological and genomic analysis of the 299 patients from our hospital and identified 35 clusters of identical viruses involving 159 patients. 92 (58%) of 159 patients had strong epidemiological links and 32 (20%) patients had plausible epidemiological links. These results were fed back to clinical, infection control, and hospital management teams, leading to infection-control interventions and informing patient safety reporting.

**Interpretation** We established real-time genomic surveillance of SARS-CoV-2 in a UK hospital and showed the benefit of combined genomic and epidemiological analysis for the investigation of health-care associated COVID-19. This approach enabled us to detect cryptic transmission events and identify opportunities to target infection-control interventions to further reduce health-care associated infections. Our findings have important implications for national public health policy as they enable rapid tracking and investigation of infections in hospital and community settings.

**Funding** COVID-19 Genomics UK (supported by UK Research and Innovation, the National Institute of Health Research, the Wellcome Sanger Institute), the Wellcome Trust, the Academy of Medical Sciences and the Health Foundation, and the National Institute for Health Research Cambridge Biomedical Research Centre.



## Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in the human population in December, 2019,[1] originating from an intermediate animal host.[2] Owing to the error prone nature of the viral replication process, RNA viruses, such as SARS-CoV-2, accumulate mutations over time resulting in sequence diversity. The current mutation rate of SARS-CoV-2 is estimated to be approximately 2·5 nucleotides per month.[3] Sequencing of SARS-CoV-2 can provide valuable information on virus biology, transmission, and population dynamics.[4–7] When linked with detailed epidemiological data and on a timescale of days, genomic data can support epidemiological investigations of potential hospital-acquired infections. On a larger population scale, genomic surveillance of SARS-CoV-2 can inform which lineages of the virus are circulating in the human population, how these change over time as an indicator of the success of control measures, how often new sources of virus are introduced from other geographical areas, and how the virus evolves in response to interventions.

Health-care associated infections can affect both patients (increasing morbidity and mortality) and health-care workers (impacting on staff sickness and morale) to the detriment of patient care. It is crucial to rapidly detect and manage health-care associated infections effectively to prevent both complications and onward transmission to susceptible patients and staff. The burden of nosocomial COVID-19 infections is unknown with one early study from China reporting a prevalence of 41% among hospitalised patients.[8] Worldwide, more than

*Lancet Infect Dis* 2020; 20: 1263–72

Published **Online** July 14, 2020 https://doi.org/10.1016/S1473-3099(20)30562-4

**This online publication has been corrected. The first corrected version appeared at thelancet.com/infection on January 22, 2021 and the second on February 11, 2021**

See **Comment** page 1218

*Contributed equally

†Joint first author

‡Joint last author

**Department of Pathology** (L W Meredith PhD, M Hosmillo PhD, A S Jahun PhD, L G Caller PhD, A Yakovleva BSc, G Hall BS, Prof I Goodfellow PhD) **and Department of Medicine** (W L Hamilton PhD, B Warne MB Bchir, C J Houldcroft PhD, S L Caddy PhD, F A Khokhar BSc, T Feltwell, S Forrest BSc, S Sridhar BS, M P Weekes PhD, Prof S Baker PhD, Prof S J Peacock FRCP, Prof G Dougan PhD, M E Török FRCP)**, University of Cambridge, Cambridge, UK; Cambridge University Hospitals National Health Service Foundation Trust, Cambridge, UK** (W L Hamilton, B Warne, E Moore MBBS, T Gouliouris PhD, M E Török)**; Public Health England Clinical Microbiology and Public Health Laboratory, Cambridge, UK** (M D Curran PhD, S Parmar PhD, N Brown MD, T Gouliouris)**; Francis Crick Institute, London, UK** (L G Caller)**; Cambridge Institute for Therapeutic Immunology and Infectious Disease, Cambridge, UK** (S L Caddy, F A Khokhar, S Forrest, S Sridhar, M P Weekes, Prof S Baker, Prof G Dougan)**; Wellcome Sanger Institute, Hinxton, UK** (S Sridhar);

**Field Epidemiology, Field Service, National Infection Service, Public Health England, Cambridge, UK** (A Popay BSc, I Roddick BSc, M Reacher MD); **and National Infection Service, Public Health England, London, UK** (Prof S J Peacock)

Correspondence to:
Dr M Estée Török, Department of Medicine, University of Cambridge, Cambridge CB2 0QQ, UK
**et317@cam.ac.uk**

or

Prof Ian Goodfellow, Department of Pathology, University of Cambridge, Cambridge CB2 0QQ, UK
**ig299@cam.ac.uk**

## Research in context

### Evidence before this study

Since the emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in December, 2019, the infection has spread worldwide, infecting more than 9·1 million people and causing more than 472 000 deaths as of June 23, 2020. Despite investigation, substantial gaps remain in our understanding of virus biology and transmission. We searched PubMed Central, medRxiv, and bioRxiv with combinations of "SARS-CoV-2", "genome", or "genomic" and "hospital-acquired" or "healthcare-associated" for articles published in English from database inception until May 11, 2020, and returned few relevant results. Previous studies have analysed SARS-CoV-2 biology, diversity and evolution, transmission networks, and health-care worker infections. Very few have applied genomic epidemiology to tackle health-care associated infection and, to our knowledge, none have been at the scale of a hospital COVID-19 epidemic comprising hundreds of patients in real-time. Attempts with other pathogens have often been assessed retrospectively, and in a timeframe that was not actionable.

### Added value of this study

We present the first report, to our knowledge, applying rapid genome sequencing to systematically investigate SARS-CoV-2 health-care associated infections, integrating genomic and epidemiological data to identify transmission networks and inform targeted infection control interventions. In 6 weeks, we sequenced 1000 genomes, including 70% of all COVID-19 cases tested at our hospital. We uncovered ward outbreaks of hospital-acquired infections and substantial transmission in health-care associated community settings. Genomic analysis identified cryptic transmission that had not been suspected by clinical or infection control teams. These complex transmission networks involved patients and health-care workers and spanned hospital and community health-care settings. By feeding results back to the hospital weekly, the data could be actioned by the infection control team during the course of outbreak investigations.

### Implications of all the available evidence

Rapid viral sequencing can contribute to health-care associated infection investigations by uncovering evidence for or against transmission events. As transmission shifts from the community to health-care settings, the use of rapid sequencing integrated with epidemiological investigations can help to reveal complex transmission chains and inform targeted infection control and public health interventions. This strategy could support the new test, track, and trace initiative of the UK Government, enabling a more targeted approach to disease control.

22 000 cases of COVID-19 infection in health-care workers were reported in the WHO situation report from May, 2020, which is likely to be an underestimate.[9] As the number of community-acquired COVID-19 cases reduces, health-care settings are likely to act as reservoirs of infection. Identifying transmission events in these settings will therefore become increasingly important to manage outbreaks and effectively monitor infection control.

See Online for appendix

We aimed to examine the use of rapid sequencing of SARS-CoV-2, combined with detailed epidemiological analysis, to investigate health-care associated COVID-19 infections and to inform infection control measures in our hospital.

## Methods

### Study design and participants

We did a prospective surveillance study of SARS-CoV-2 infections at Cambridge University Hospitals National Health Service Foundation Trust (CUH; Cambridge, UK), a secondary care provider and tertiary referral centre in the East of England. Sufficient supplies of personal protective equipment (PPE) to meet national recommendations were available in CUH at all times during the study period. We adhered to national guidance on PPE use. Patients were isolated either pre-emptively (suspected COVID-19) or immediately after a confirmed PCR positive result. Clinical specimens collected from patients presenting to 18 hospitals in the East of England were submitted to the Public Health England Clinical Microbiology and Public Health Laboratory at CUH for diagnostic testing. Samples underwent nucleic acid extraction and were tested for the presence of SARS-CoV-2 with a validated in-house RT-qPCR assay developed by the Public Health England Clinical Microbiology and Public Health Laboratory (appendix p 1). The test was reported as SARS-CoV-2 PCR positive if the cycle threshold (Ct) value was 36 or less.

The study was done as part of surveillance for COVID-19 under the auspices of Section 251 of the National Health Service Act 2006. It therefore did not require individual patient consent or ethical approval. The COVID-19 Genomics UK (COG-UK) study protocol was approved by the Public Health England Research Ethics Governance Group.

### Procedures

Demographic, clinical, and laboratory data were extracted from the hospital information system (Epic Systems, Verona, WI, USA; appendix p 5). Sample collection, metadata curation, and linkage to sequencing identification numbers involved coordination between multiple teams and depended on daily email communications and face-to-face contact between the sequencing and diagnostic laboratory staff (appendix p 5). Each day samples with positive SARS-CoV-2 PCR results from the last 24–72 h were identified from the hospital information system and clinical metadata was extracted, formatted, and integrated

| | Label | Definition | Total (n) |
|---|---|---|---|
| 1 | Community onset, community associated | Diagnostic sample positive within 48 h of admission and no health-care contact during the previous 14 days | 263 |
| 2 | Community onset, suspected health-care associated | Diagnostic sample positive within 48 h of admission with health-care contact during the previous 14 days | 32 |
| 3 | Hospital onset, indeterminate health-care associated | Diagnostic sample positive after 48 h but less than 7 days post-admission | 13 |
| 4 | Hospital onset, suspected health-care associated | Diagnostic sample positive 7–14 days post-admission | 14 |
| 5 | Hospital onset, health-care associated | Diagnostic sample positive more than 14 days post-admission | 43 |
| 6 | Health-care worker | Any positive test identified as coming from a health-care worker* | 9 |

Different categories of hospital onset infection reflect the increasing likelihood of hospital acquired infection based on the incubation period of the virus. N=374 is the number of CUH patients included in the study period, of whom 262 (70%) had sequencing available for analysis (appendix p 9). CUH=Cambridge University Hospitals National Health Service Foundation Trust. *These data do not include 37 health-care workers identified through the CUH screening programme.

***Table:*** **Definitions of health-care associated COVID-19**

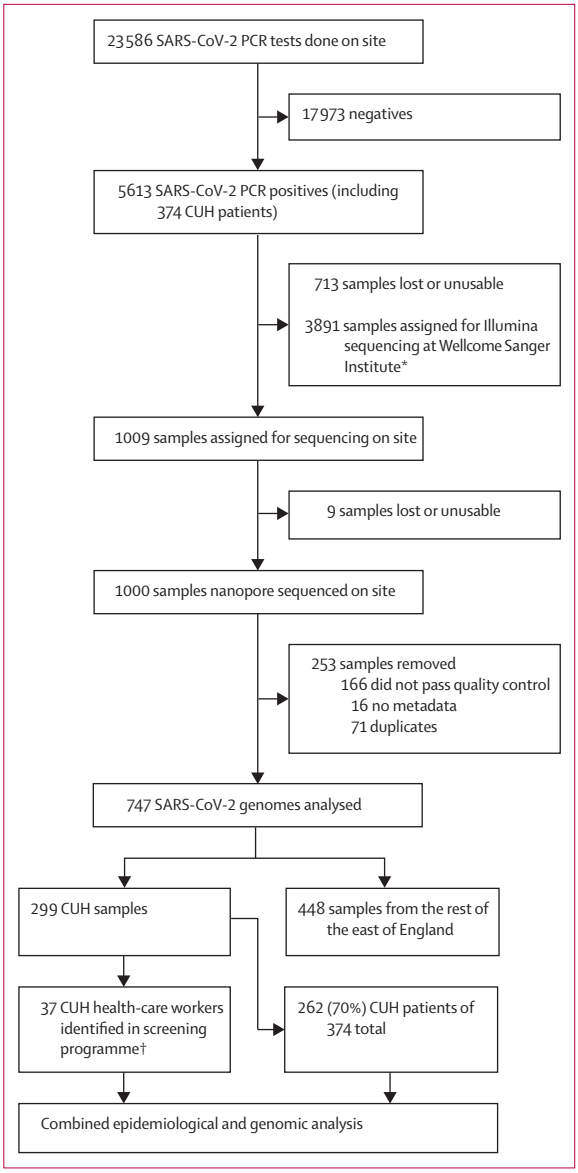


***Figure 1:*** **Study profile**
We prioritised CUH samples for nanopore sequencing on site for quick turnaround to investigate health-care associated infections. Of the 166 samples that did not pass quality control, two were removed as their genomes were less than 29 kb and 164 were removed as they had more than 2990 undefined nucleotides. SARS-CoV-2=severe acute respiratory syndrome coronavirus 2. CUH=Cambridge University Hospitals National Health Service Foundation Trust. *Of 3891 samples assigned for sequencing, 2940 were uploaded to the CLIMB server, as have all of the 1000 genomes sequenced on site, for national COVID-19 Genomics UK analyses. †The 37 health-care workers were identified through a CUH screening programme. In addition, nine self-presented to CUH admission units and are counted as patients (n=46).

into a master metadata file. A 15 µL aliquot of the RNA extract of all CUH samples and a random selection of samples from the East of England were collected from the diagnostic microbiology laboratory for local sequencing. We aimed for each sampling site in the East of England to be represented across the study period. Typically, two positive samples per site were selected per day for sequencing, including both high and low Ct samples when possible. Samples were also included from a local health-care worker screening programme (n=37 in this dataset).[10] Most samples were from patients with symptomatic COVID-19; two samples were from asymptomatic individuals.

Samples were assigned COG-UK sequencing codes that were integrated back into the master metadata file. For samples that were not sequenced locally, the remaining RNA extract was collected from the diagnostic microbiology laboratory and sent to the Wellcome Sanger Institute (Hinxton, UK) for sequencing as part of the COG-UK consortium. 14 samples sequenced on site were also sent to the institute to compare consistency across sequencing platforms. Each week, clinical metadata and sequencing data were combined and formatted for upload to the Medical Research Council CLIMB system. Data manipulations were done in R (version 3.6.2) with the tidyverse packages (version 1.3.0) installed onto computers within the Trust network.

When Ct values were available before sample selection, positive samples with a Ct value of 33 or less were sequenced with a multiplex PCR based approach according to the modified ARTIC version 2 protocol with version 3 primer set (appendix p 2).[11,12] Amplicon libraries were sequenced using MinION flow cells version 9.4.1 (Oxford Nanopore Technologies, Oxford, UK). Genomes were assembled with reference-based assembly and a bioinformatic pipeline[13] with 20× minimum coverage

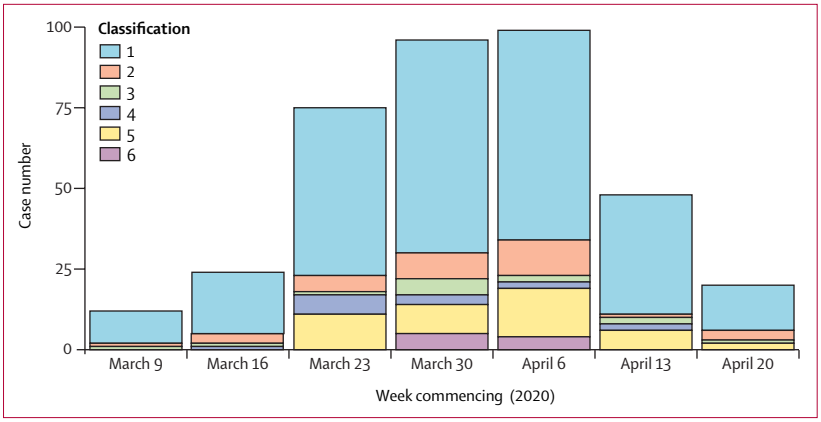


***Figure 2:*** **Epidemic curve of COVID-19 at CUH**
Data for 374 patients tested at CUH. Classification of infection: (1) community onset, community associated; (2) community onset, suspected health-care associated; (3) hospital onset, indeterminate health-care associated; (4) hospital onset, suspected health-care associated; (5) hospital onset, health-care associated; (6) health-care worker. These data do not include 37 health-care workers identified through the CUH screening programme, but do include nine health-care workers that self-presented for testing. Note that data for week commencing April 20, 2020, stops at 8 AM on April 24, 2020, so does not include the weekend. CUH=Cambridge University Hospitals National Health Service Foundation Trust.

cutoff for any region of the genome and 50·1% cutoff for defining single nucleotide polymorphisms (SNPs).

### Analysis

All sequences underwent quality control filtering and de-duplication to remove repeat samples from the same patient. Multiple sequence alignment was done with MAFFT.[14] Phylogenetic trees were produced with IQ-TREE[15] and visualised in Microreact[16] for weekly hospital reports and the R package ggtree (appendix pp 2–3).[17] A pairwise SNP distance matrix was produced from the alignment with use of the snp-dists package. Viral lineages[18] were assigned with the PANGOLIN package, version 1.07.

For epidemiological analysis, patient movement data for all SARS-CoV-2 PCR positive samples were extracted from the hospital information system and transferred to the Public Health England Field Service (Epidemiology). Epidemiological analysis was done with a cloud-based plotter application (Cluster Track; appendix pp 3–4).

Clusters of COVID-19 cases including health-care workers were identified by the clinical and infection control teams and reviewed at a weekly meeting, coordinated by the patient safety team (appendix p 6). The genomic and epidemiological analyses were presented to help establish whether infections were health-care associated and to identify possible causes and interventions. A weekly report was fed back to the clinical, infection control, and hospital management teams to inform changes in infection-control practice and comply with patient safety procedures. Definitions of health-care associated COVID-19 are shown in the table.

### Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

## Results

Between March 13 and April 24, 2020, 1000 samples were selected for sequencing and, after quality control and de-duplication, 747 were used for downstream analysis. 299 (40%) of 747 samples were from CUH including 46 health-care worker samples, of which 37 were identified through a health-care worker screening programme[10] (including two asymptomatic individuals). 374 patients with PCR-confirmed COVID-19 were tested at CUH between March 10 and April 24, 2020 (figure 1, appendix pp 7–8). The median age was 64 years (range 0–98) and 233 (62%) of 374 were male. 74 (20%) of 374 patients were admitted to critical care units and 75 (20%) died. Excluding the health-care workers screening samples, 262 (70%) of 374 CUH COVID-19 samples had sequencing data available (figure 1; appendix p 9). 57 (15%) infections were suspected or highly likely to be hospital-acquired, of which 49 (86%) had genome sequences available. A further 32 (9%) admissions were community-acquired but likely to be health-care associated, and nine (2%) were health-care workers (not counting health-care workers identified through screening). The CUH epidemic curve showed that weekly admissions peaked in week 4 (commencing March 30, 2020) and then declined (figure 2). The UK went into full lockdown on March 23, 2020. In the early stages of the epidemic, community-onset and community-acquired infections predominated but the frequency of health-care associated infections increased from March 23 until April 6, 2020, and then declined.

Each week a sample set was locked and underwent bioinformatic analysis. Of 1000 sequenced genomes presented here, 253 were excluded from downstream analysis because they did not reach quality control thresholds (n=166), metadata were missing (n=16), or they were repeat samples from the same patient (n=71; figure 1). The median genome depth of coverage was 6612×. We compared Ct value versus depth of coverage and found that the latter declined at Ct values of more than 30 (appendix p 10). We also examined the location and frequency of SNPs across the genomes (appendix p 11). Genomes were assigned to a lineage based on the combination of mutations that have accumulated since the virus emerged. As of March 23, 2020, 12 lineages had been described in the UK.[19] Most samples in both the East of England and CUH belonged to lineage B.1. There were no lineage A samples, which have mainly been identified in China, the USA, South Korea, and Australia (figure 3, appendix pp 12–13).[18]

Phylogenetic trees were used to explore potential genetic clustering and correlation with sampling ward location and cases of suspected hospital-acquired infections (figure 4). Samples collected from the

emergency department were phylogenetically dispersed, probably reflecting unconnected transmission events within the past few days or weeks in community-acquired infections (appendix p 14). By contrast, samples collected from several wards in CUH and an outpatient dialysis unit were genetically clustered (figure 4). The putative ward clusters included a high proportion of suspected hospital-acquired infections, suggestive of linked transmission chains in the hospital.

SARS-CoV-2 has low genetic diversity due to its introduction into the human population within 6 months of sample collection; here, a median of eight SNPs separated any two samples at CUH (appendix pp 15–16). This low sequence diversity makes interpretation of putative clusters challenging, as samples could be identical by chance rather than because of a connected transmission chain. To investigate genomic clustering further, we adopted a combined genetic, clinical, and epidemiological approach. Samples with zero SNP differences were identified and clusters named numerically by decreasing sample size.

Overall 159 (53%) of 299 genomes from CUH in 35 clusters shared at least one identical sequence. The largest cluster had 18 identical genomes. Patients' medical records (including address, social setting, clinical details, and ward movements) were reviewed for all putative genomic clusters to assess whether cases had plausible or probable linked transmission within a few days or weeks (appendix pp 17–20). Of 159 cases from 35 putative clusters, 92 (58%) cases had strong epidemiological evidence to support transmission within a few days or weeks, 32 cases (20%) had intermediate evidence, and 35 (22%) no evidence of connected transmission. Clusters with strong evidence of linked transmission within a few days or weeks included cases in which a connection was already suspected, such as groups of probable hospital-acquired infections seen on multiple CUH hospital wards (figure 4), and clusters that were previously not recognised as being linked, such as a care home outbreak involving health-care workers based in hospital and community settings.

Directed by clinical and infection control teams, we established a process for focused genomic and epidemiological analyses of suspected hospital-acquired infections (appendix p 6). These were discussed in weekly meetings with an accompanying written report submitted to the hospital. A full description of clusters is detailed in the appendix (pp 17–20). In brief, epidemiologically linked cases with identical viral genomes indicating transmission events were identified in 12 hospital wards and an outpatient dialysis unit. Nine of the hospital wards were classed as green (ie, no known patients with COVID-19) at the onset of cluster cases, and three were classed as red (ie, housing patients with confirmed or highly suspected COVID-19). Additionally, community transmission events were identified in three care homes (in both residents and carers), hostel accommodation, and several households.

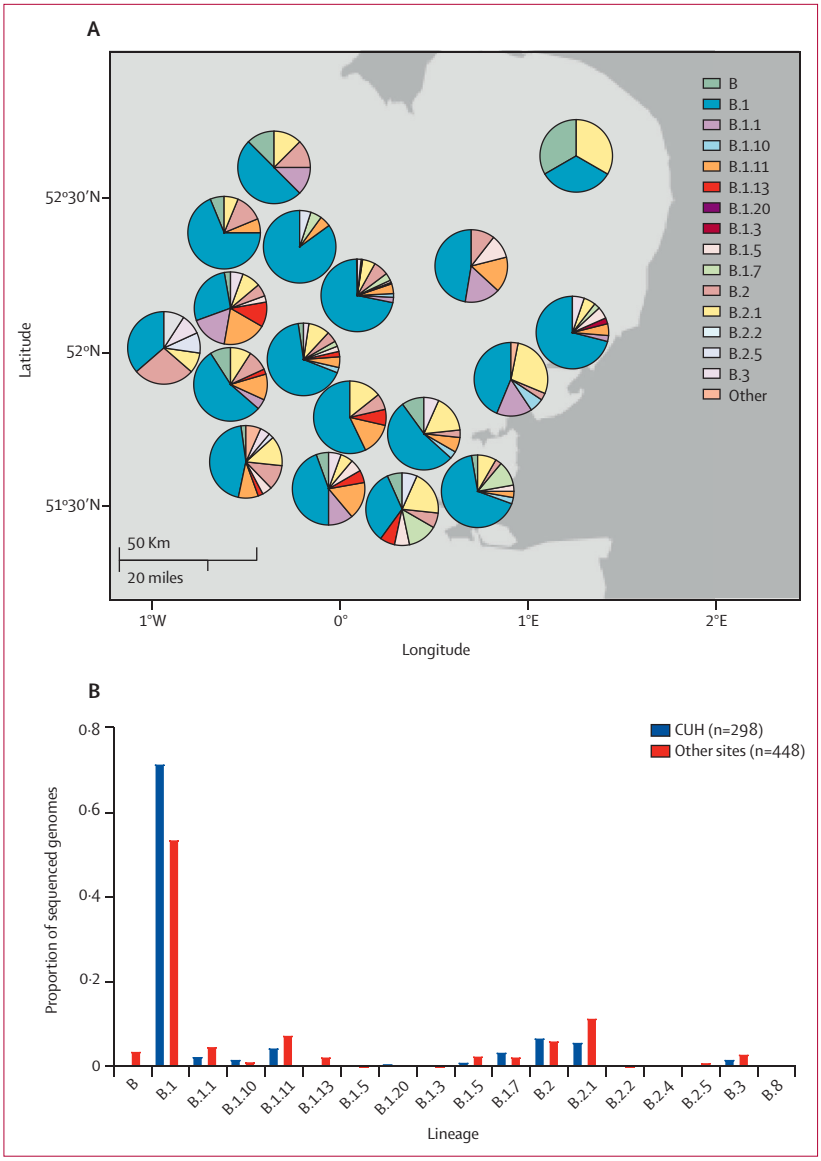


***Figure 3:*** **SARS-CoV-2 lineages identified in the East of England and CUH**
Map of the East of England region (A) showing the breakdown of SARS-CoV-2 lineages by collecting hospital site. Lineage B.1 was the most prevalent lineage throughout the region (B). SARS-CoV-2=severe acute respiratory syndrome coronavirus 2. CUH=Cambridge University Hospitals National Health Service Foundation Trust.

Transmission events involving health-care workers were identified in both hospital and community settings, such as a cluster including several paramedics.

In hospital cluster 1, six surgical patients on ward A (an area that housed no known patients with COVID-19) were diagnosed with COVID-19 (figure 5). All cases had been on the ward for more than 7 days before their specimen collection date and were considered as probable hospital-acquired infections. Genomic data were available for all cases; five differed by zero SNPs and one by a single SNP, consistent with ward-based transmission events within a few days. The genomic evidence supported the infection

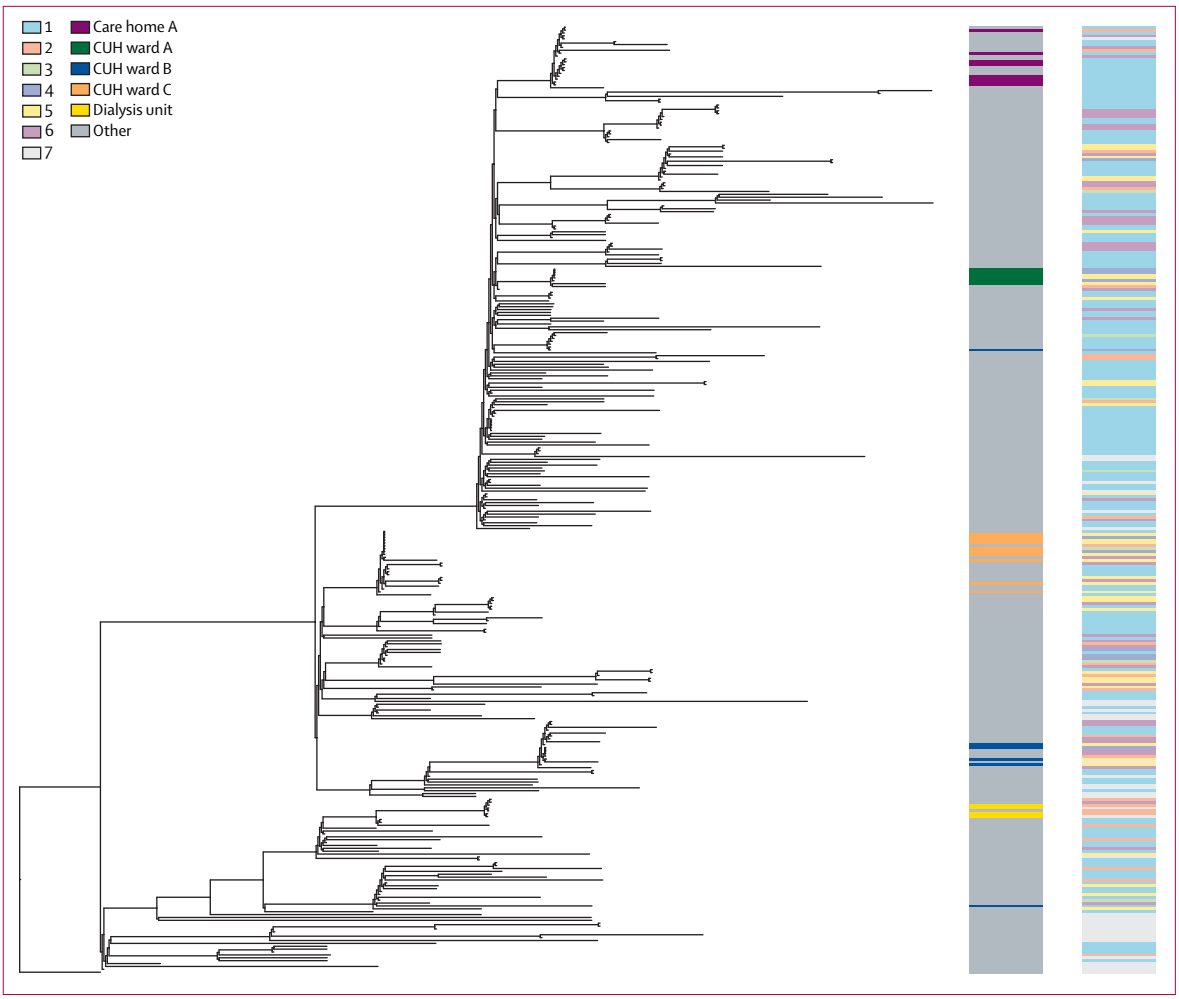


***Figure 4:*** **Phylogenetic tree of SARS-CoV-2 genomes from CUH**

Phylogenetic tree of 299 SARS-CoV-2 genomes form CUH and 30 international genomes. The tree is rooted on a December, 2019, genome from Wuhan, China. The left-hand column highlights several hospital and community associated clusters in different colours. Wards A, B, and C all had clusters of hospital-acquired infections cases with viruses of less than two single nucleotide polymorphisms different. Eight ward C cases are contained within one of the largest clusters of identical viruses. Genomic clusters containing the cases for wards B and C, the dialysis unit, and care home A all included health-care workers. The right hand column shows the classification of infection, also shown in figure 2. Classification of infection: (1) community onset, community associated; (2) community onset, suspected health-care associated; (3) hospital onset, indeterminate health-care associated; (4) hospital onset, suspected health-care associated; (5) hospital onset, health-care associated; (6) health-care worker; (7) unable to determine or missing. See appendix pp 23–24 for GISAID identification codes of included reference genomes, and appendix p 13 for their position on the East of England phylogenetic tree. SARS-CoV-2=severe acute respiratory syndrome coronavirus 2. CUH=Cambridge University Hospitals National Health Service Foundation Trust.

control decision to close the ward to new admissions and enhance surveillance of all patients who had contact with this clinical area.

In hospital cluster 2, four transplant patients on ward B (an area that housed no known patients with COVID-19) were diagnosed with the virus between April 1 and April 20, 2020 (appendix p 21). A fifth patient, who had been discharged from the ward within the past few days, presented to the emergency department with COVID-19 symptoms. Genomic analysis revealed that all five cases had identical genomes. Three health-care workers were found to have identical genomes in the same cluster as the ward B cases; two had worked on ward B, one of whom had professional contact with the other health-care workers. These findings led to a review of infection control and PPE procedures for staff and patients in the transplant service.

Patients on renal dialysis are among the most susceptible to COVID-19 (with up to 19% mortality).[20] Most dialysis units have challenging infection control arrangements, consisting of large open rooms with no barriers between patients. In the dialysis unit cluster, six patients with end-stage renal failure were diagnosed with COVID-19 between April 1 and April 20, 2020, testing positive in several locations including the emergency department and an acute admissions ward (appendix p 21). Their viral genomes were identical, and epidemiological investigation revealed that they dialysed at the same outpatient unit on the same days of the week. This information suggests

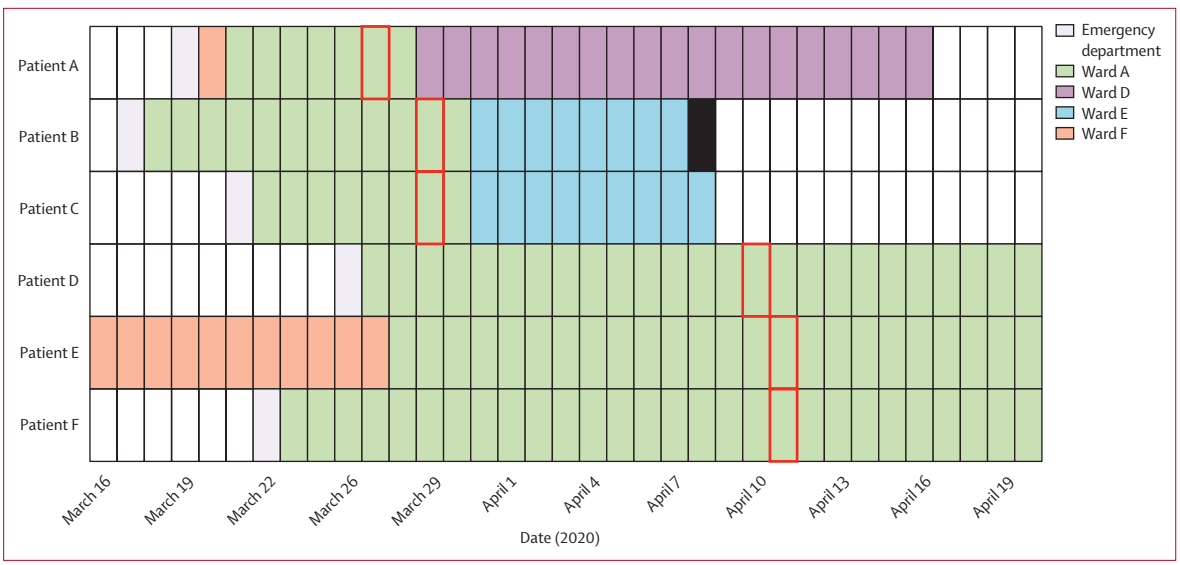


***Figure 5:*** **Epidemiological timeline of ward A cluster**
Six patients on ward A were diagnosed with COVID-19. All had been admitted for more than 7 days before their specimen date and were considered likely to have hospital-acquired infections. The date of the first positive sample collection is shown with a red box and patient death date indicated with a solid black bar. Five of the viral genomes (patients A to E) had zero single nucleotide polymorphism differences between them, and one of the viral genomes (patient F) differed by one single nucleotide polymorphism.

linked transmission of community-onset, health-care associated infections. These findings led to a review of infection control procedures in patients on dialysis and identified shared patient transportation and neighbouring dialysis chairs as risk factors for transmission. Genomics was also useful for ruling out linked transmission. The renal ward (which shares patients with the outpatient dialysis unit) also had a group of COVID-19 cases at around the same time. However, the dialysis unit genomes belonged to lineage B.2 (relatively rare in East of England), whereas the renal ward genomes were the common B.1 lineage, making it very unlikely that infections between the two patient groups were related.

In the community cluster, 18 patients were admitted to CUH with COVID-19 between April 5 and April 21, 2020, with genetically identical viruses. Nine patients were residents at a community care home (care home A). A review of medical records revealed that another patient in this genetic cluster worked in care home A and one was a retired nurse who worked in an unknown care home. Three cases were paramedics and two were nurses (who worked in different wards at CUH but lived with paramedics). The final case did not have any discernible epidemiological links with the others. In summary, this investigation revealed a cluster of cases with evidence of linked transmission coming from either the same care home, the ambulance service, or shared accommodation. None of these associations had been detected by clinicians or infection control.

The information from these combined epidemiological and genomic investigations was fed back to the clinical, infection control, and hospital management teams. This information triggered further investigations into patient isolation, ward cleaning procedures, use of PPE, and staff physical distancing behaviour. Health-care associated infections were also assessed in relation to potential harm caused to patients and recorded in the hospital's patient safety reporting system for follow-up and further action.

## Discussion

The value of real-time viral genome sequencing has been shown in previous epidemics (eg, Ebola virus, measles, Zika virus, and influenza).[21–25] We sought to embed genomic surveillance as part of an active SARS-CoV-2 infection control process in a large UK hospital. A rapid sequencing workflow was established on March 23, 2020, with multiplex PCR-based nanopore sequencing, which has been shown to be effective in a wide range of clinical samples and viral loads.[26] We aimed to sequence all available positive samples from CUH and a selection from each of the East of England regional hospitals submitted to the diagnostic microbiology laboratory, linking with clinical metadata pulled from the hospital electronic patient records system. We also included 37 samples collected as part of a local health-care worker screening programme.[10] In 5 weeks, more than 1000 SARS-CoV-2 genomes had been sequenced including most of the CUH samples from this phase of the epidemic. We applied this system to investigate nosocomial and health-care worker COVID-19 cases at CUH, integrating genomics with epidemiological and clinical data.

We examined the diversity in SARS-CoV-2 at CUH and found that overall genetic diversity was low and

reflected the pattern seen in the East of England as a whole, with most viruses belonging to the B.1 lineage. We identified a median of eight (range 0–24) SNP differences between viruses, and 4·5% of pairwise comparisons between CUH genomes had zero to one SNP differences. Given the virus' mutation rate and infectious timeframe, cases might share linked transmission within a few days or weeks if there are fewer than approximately two SNP differences. We applied a more stringent definition of genetic clusters with zero SNP differences, and despite detailed epidemiological data, we found no identifiable connection between 22% of pairwise comparisons within clusters. This finding reflects the low genetic diversity in circulating SARS-CoV-2 during the study, and emphasises the need for in-depth epidemiological analysis to unravel potential transmission networks. The ability of genomics to resolve transmission events might increase as the virus evolves and accumulates greater diversity. Genetics can be used more confidently to rule out transmission—eg, if viruses from two patients suspected as being linked belong to different lineages.

We investigated groups of patients and health-care workers at CUH in response to queries from the clinical and infection control teams. Using ward location data for patients and health-care workers, we analysed epidemiological data to establish if there had been ward-based contact. We compared the genomes of patients and health-care workers in the suspected groups with those of other patients at CUH and in hospitals in the East of England to examine relatedness. This approach enabled us to add supporting evidence or to refute linked transmission between patients and health-care workers (eg, adding confidence to our assessment of whether an infection was hospital or community acquired). Ruling out transmission was useful as a mechanism to monitor and target infection control measures (eg, showing that viruses on the renal ward belonged to a distinct lineage from those in the outpatient dialysis unit). Genomic analysis also enabled us to identify cryptic transmission (ie, additional cases that were not initially suspected to be linked to the original ward or health-care worker clusters, including multiple instances of patients with identical viruses from the same care homes).

These cases illustrate the power of combining rapid genomic and epidemiological analyses in near real time. By contrast with previous studies,[27–30] we reported results of our investigations to the clinical, infection control, and management teams on a weekly basis, thus enabling them to respond to this information and act accordingly within the timeframe of ongoing ward outbreaks. The genomic data informed reviews of patient placement and isolation procedures, assessment of PPE use, and staff break arrangements, supporting us to better focus efforts at a time of unprecedented demand on infection control teams. Finally, these analyses are being used to inform existing patient safety review processes within our hospital, including investigations related to hospital-onset COVID-19 in which the patient has come to harm.

Our study highlights the importance of understanding SARS-CoV-2 transmission within health-care settings in managing the pandemic. The transmission networks that we identified were complex, involving patients and health-care workers in both hospital and community settings such as care homes, outpatient units, and ambulance services, which have been poorly studied. Of note, we have identified transmission events on nine wards that were considered green (ie, no known patients with COVID-19) at the start of each cluster. Although there were strong epidemiological and genomic associations between cases, the mechanism and direction of transmissions within these clusters are unclear. The role of asymptomatic intermediates, fomites (including PPE), and the environment are not well understood and require further investigation. During the timeframe of this study, several infection-control interventions were implemented across the hospital, as well as national public health measures to reduce community spread. Understanding the interaction between such interventions and nosocomial transmission are complex (especially in the context of the comparably long incubation period for SARS-CoV-2 relative to other respiratory viruses), but essential in enabling health-care providers to safely deliver existing services in the context of a pandemic.

We acknowledge several limitations to our study. Firstly, we were unable to sequence all genomes from samples that were collected during the study period. We might therefore have missed the opportunity to investigate all potential transmission events. 166 (17%) of 1000 sequenced genomes did not pass our quality filtering. This finding reflects the stringent coverage threshold used (90%), the desire not to bias sample selection by sequencing only low Ct samples, and the nature of the diagnostic material used for sequencing. We only had main ward location data for health-care workers and could have overlooked potential epidemiological links with patients with whom they had contact on other wards or within shared communal areas. Due to its low genetic diversity, highly similar genomes could not be used definitively to infer meaningfully linked transmission events without supportive epidemiological data. However, our experience indicates that further investigation of genomic clusters with highly similar genomes can uncover previously unknown epidemiological links. Furthermore, we were able to use rapidly generated genomic data to investigate health-care associated infections within the hospital setting. Similar approaches could be applied in future studies to assess infections in health-care workers and community settings such as care homes. As the practical challenges associated with implementing real-time genome sequencing during epidemics are overcome, unlocking the real power of genomic epidemiology will require its integration with clinical and public health systems to support decision making on local, national, and international scales. This

implementation might be of particular benefit in supporting the UK Government's test, track, and trace initiative, enabling a more targeted approach to disease control.

#### Contributors

LWM contributed to the data collection, data analysis, data interpretation, figures, and tables. WLH, BW, and CJH contributed to the data collection, data analysis, data interpretation, figures, tables, literature review, and writing. MH, ASJ, MDC, SP, LGC, SLC, FAK, AY, GH, TF, SF, SS, MPW, SB, NB, EM, and TG contributed to the data collection, data analysis, and data interpretation. AP, IR, and MR contributed to the data collection, data analysis, data interpretation, and figures. SJP and GD contributed to the writing of the manuscript. MET and IG designed and supervised the study and contributed to the data collection, data analysis, data interpretation, literature review, and wrote the first draft of the manuscript. All authors reviewed and approved the final manuscript.

#### Declaration of interests

CJH, SLC, MPW, and SB report grants from Wellcome. GH reports grants from Rotary International. MET reports grants from Academy of Medical Sciences, Health Foundation, grants from Medical Research Council, grants and non-financial support from National Institute of Health Research, during the conduct of the study. MET has published books with Oxford University Press and receives royalty payments from them, personal fees from the Wellcome Sanger Institute, personal fees from University of Cambridge, outside the submitted work. IG reports grants from Wellcome; and grants from Medical Research Council part of UK Research & Innovation. All other authors declare no competing interests.

#### Acknowledgments

We acknowledge the assistance of the laboratory staff of Public Health England Clinical Microbiology and Public Health Laboratory for processing the diagnostic samples; the clinical teams (infectious diseases, microbiology, virology, infection control) at CUH for their assistance with the investigation of health-care associated infections; and the Wellcome Sanger Institute for sequencing 14 samples included in this study. We are grateful to Richard Smith (Head of Patient Safety, CUH), Lucy Rivett, Dominic Sparkes, Nick K Jones, and Matthew Routledge (for providing health-care worker data), and Anthony Underwood (Centre for Genomic Pathogen Surveillance) for helpful discussions and advice. This work was funded by COVID-19 Genomics UK, which is supported by funding from the Medical Research Council part of UK Research and Innovation, the National Institute of Health Research, and Genome Research, operating as the Wellcome Sanger Institute. It was also supported by the Wellcome (Senior Fellowship 207498/Z/17/Z and ARTIC Network Collaborative Award 206298/B/17/Z to IG, Collaborative Award 204870/Z/16/Z supporting CJH, Senior Research Fellowship to SGB 215515/Z/19/Z, Senior Clinical Research Fellowship 108070/Z/15/Z to MPW), the Academy of Medical Sciences and the Health Foundation (Clinician Scientist Fellowship to MET), and the National Institute for Health Research Cambridge Biomedical Research Centre at the CUH (BW, GD, MET). The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute of Health Research, or the Department of Health and Social Care.

#### References

1 Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020; **579:** 270–73.

2 Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. *Nat Med* 2020; **26:** 450–52.

3 Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. Temporal signal and the phylodynamic threshold of SARS-CoV-2. *bioRxiv* 2020; published online May 4. DOI:10.1101/2020.05.04.077735 (preprint).

4 Grubaugh ND, Ladner JT, Lemey P, et al. Tracking virus outbreaks in the twenty-first century. *Nat Microbiol* 2019; **4:** 10–19.

5 Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat Rev Genet* 2018; **19:** 9–20.

6 Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome sequencing. *Nat Rev Microbiol* 2017; **15:** 183–92.

7 Grubaugh ND, Ladner JT, Kraemer MUG, et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* 2017; **546:** 401–05.

8 Wang D, Hu B, Hu C, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* 2020; **323:** 1061–69.

9 WHO. Coronavirus disease 2019 (COVID-19) Situation Report–82. 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200411-sitrep-82-covid-19.pdf (accessed May 8, 2020).

10 Rivett L, Sridhar S, Sparkes D, et al. Screening of healthcare workers for SARS-CoV-2 highlights the role of asymptomatic carriage in COVID-19 transmission. *eLife* 2020; **9:** e58728.

11 Quick J. https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/ (accessed May 6, 2020).

12 Quick J. nCoV-2019 sequencing protocol. 2020. https://www.protocols.io/view/ncov-2019-sequencing-protocol-v2-bdp7i5rn (accessed May 6, 2020).

13 Loman N, Rowe W, Rambaut A. nCoV-2019 novel coronavirus bioinformatics protocol. 2020. https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html (accessed May 6, 2020).

14 Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013; **30:** 772–80.

15 Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015; **32:** 268–74.

16 Argimón S, Abudahab K, Goater RJE, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* 2016; **2:** e000093.

17 Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017; **8:** 28–36.

18 Rambaut A, Holmes EC, Hill V, et al. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv* 2020; published online April 19. DOI:2020.04.17.046086 (preprint).

19 Consortium C-U. COVID-19 Genomics UK (COG-UK) Consortium Weekly Report #1 23rd March 2020. 2020. https://www.cogconsortium.uk/wp-content/uploads/2020/04/14-COG-UK-Weekly-Report-23rd-March-20204.pdf (accessed May 7, 2020).

20 Association TR. COVID-19 data. 2020. https://renal.org/covid-19/data/ (accessed May 8, 2020).

21 Arias A, Watson SJ, Asogun D, et al. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evol* 2016; **2:** vew016.

22 Quick J, Loman NJ, Duraffour S, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature* 2016; **530:** 228–32.

23 Dudas G, Carvalho LM, Bedford T, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* 2017; **544:** 309–15.

24 Gardy JL, Naus M, Amlani A, et al. Whole-genome sequencing of measles virus genotypes H1 and D8 during outbreaks of infection following the 2010 Olympic Winter Games reveals viral transmission routes. *J Infect Dis* 2015; **212:** 1574–78.

25 MacFadden DR, McGeer A, Athey T, et al. Use of genome sequencing to define institutional influenza outbreaks, Toronto, Ontario, Canada, 2014–15. *Emerg Infect Dis* 2018; **24:** 492–97.

26 Quick J, Grubaugh ND, Pullan ST, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc* 2017; **12:** 1261–76.

27 Köser CU, Holden MTG, Ellington MJ, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* 2012; **366:** 2267–75.

28 Harris SR, Cartwright EJP, Török ME, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect Dis* 2013; **13:** 130–36.

29 Coll F, Harrison EM, Toleman MS, et al. Longitudinal genomic surveillance of MRSA in the UK reveals transmission patterns in hospitals and the community. *Sci Transl Med* 2017; **9:** eaak9745.

30 Houldcroft CJ, Roy S, Morfopoulou S, et al. Use of whole-genome sequencing of adenovirus in immunocompromised pediatric patients to identify nosocomial transmission and mixed-genotype infection. *J Infect Dis* 2018; **218:** 1261–71.