

1 Genomic epidemiology of COVID-19 in care homes in the 2 East of England

3
4
5 William L. Hamilton^{*1,2,+}, Gerry Tonkin-Hill^{*3}, Emily Smith⁴, Dinesh Aggarwal^{1,5}, Charlotte J.
6 Houldcroft⁶, Ben Warne^{1,2}, Colin S. Brown⁵, Luke W. Meredith⁶, Myra Hosmillo⁶, Aminu S. Jahun⁶,
7 Martin D. Curran⁷, Surendra Parmar⁷, Laura G. Caller^{6,8}, Sarah L. Caddy⁶, Fahad A. Khokhar¹, Anna
8 Yakovleva⁶, Grant Hall⁶, Theresa Feltwell⁶, Malte L. Pinckert⁶, Iliana Georgana⁶, Yasmin Chaudhry⁶,
9 Nicholas M. Brown^{7,2}, Sónia Gonçalves³, Roberto Amato³, Ewan M. Harrison³, Mathew A. Beale³,
10 Michael Spencer Chapman^{3,9}, David K. Jackson³, Ian Johnston³, Alex Alderton³, John Sillitoe³, Cordelia
11 Langford³, Gordon Dougan¹, Sharon J. Peacock¹, Dominic P. Kwiatkowski³, Ian Goodfellow⁶, M. Estée
12 Török^{1,2,+}, COVID-19 Genomics Consortium UK¹⁰

13
14 * contributed equally

15 + Corresponding authors

17 Affiliations

- 18 1. University of Cambridge, Department of Medicine, Cambridge, UK
- 19 2. Cambridge University Hospitals NHS Foundation Trust, Departments of Infectious Diseases
20 and Microbiology, Cambridge UK
- 21 3. Wellcome Sanger Institute, Hinxton, UK
- 22 4. Cambridgeshire County Council, UK
- 23 5. Public Health England, Colindale, UK
- 24 6. University of Cambridge, Department of Pathology, Division of Virology, Cambridge, UK
- 25 7. Public Health England Clinical Microbiology and Public Health Laboratory, Cambridge UK
- 26 8. The Francis Crick Institute, London, UK
- 27 9. Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS
28 Trust, London, UK
- 29 10. www.cogconsortium.uk

31 Correspondence

32 William Hamilton will.l.hamilton@gmail.com

33 M. Estée Török et317@cam.ac.uk

1 **Word count**

2 Abstract = 150 (max 150)

3 Main text (excluding references or Methods) approx 4,000

4

5 **Key words**

6 COVID-19, SARS-CoV-2, genomics, epidemiology, care home; nursing home; residential home

7

8 **Impact statement**

9 SARS-CoV-2 can spread efficiently within care homes causing COVID-19 outbreaks among residents,
10 who are at increased risk of severe disease, emphasising the importance of stringent infection
11 control in this population.

1 **Abstract**

2 COVID-19 poses a major challenge to care homes, as SARS-CoV-2 is readily transmitted and causes
3 disproportionately severe disease in older people. Here, 1,167 residents from 337 care homes were
4 identified from a dataset of 6,600 COVID-19 cases from the East of England. Older age and being a
5 care home resident were associated with increased mortality. SARS-CoV-2 genomes were available
6 for 700 residents from 292 care homes. By integrating genomic and temporal data, 409 viral clusters
7 within the 292 homes were identified, indicating two different patterns - outbreaks among care
8 home residents and independent introductions with limited onward transmission. Approximately
9 70% of residents in the genomic analysis were admitted to hospital during the study, providing
10 extensive opportunities for transmission between care homes and hospitals. Limiting viral
11 transmission within care homes should be a key target for infection control to reduce COVID-19
12 mortality in this population.

13

14

1 Introduction

2 Care homes are at high risk of experiencing outbreaks of SARS-CoV-2. COVID-19 is associated with
3 higher mortality in older people and those with comorbidities including cardiovascular and
4 respiratory disease (Williamson et al., 2020), making the care home population especially vulnerable.
5 As of week ending 30th June 2020, the United Kingdom (UK) Office for National Statistics (ONS)
6 estimated that 30.2% of all deaths due to COVID-19 (13,417 deaths) in England occurred in care
7 homes, and 63.9% (28,390 deaths) occurred in hospital (ONS, 2020a). Most of the COVID-19 deaths
8 in hospital were in persons aged 65 years and over (86.1%). Deaths due to confirmed COVID-19 from
9 this period may be underestimates due to limitations on diagnostic testing; the ONS estimates that
10 from 28 December 2019 to 12 June 2020, there were 29,393 excess deaths in care homes compared
11 to the expected number based on previous years, of which only two thirds are explained by
12 recorded COVID-19 (ONS, 2020b). To date, SARS-CoV-2 transmission in care homes has not been
13 systematically studied with linkage of epidemiological and genomic data on a large scale.

14

15 Care homes are defined by the Care Quality Commission (CQC), the independent regulator of adult
16 health and social care in England, as “places where personal care and accommodation are provided
17 together” (CQC, 2020a). In 2011, 291,000 people aged 65 or older were living in care homes in
18 England and Wales, representing 3.2% of the total population at this age; 82.5% of the care home
19 population was aged 65 years or older (ONS, 2014). Care homes are known to be high risk settings
20 for infectious diseases, owing to a combination of the underlying vulnerability of residents who are
21 often frail and elderly with multiple comorbidities, the shared living environment with multiple
22 communal spaces, and the high number of interpersonal contacts between residents, staff and
23 visitors in an enclosed space (Curran, 2017; Lansbury et al., 2017; Strausbaugh et al., 2003).
24 Understanding the transmission dynamics of SARS-CoV-2 within care homes is therefore an urgent
25 public health priority.

26

27 Rapid SARS-CoV-2 sequencing combined with detailed epidemiological analysis has been used to
28 trace viral transmission networks in hospital and community-based healthcare settings (Meredith et
29 al., 2020). This study was based in Cambridge University Hospitals (CUH), a secondary care provider
30 and tertiary referral centre in the East of England, UK. The study focused on identifying hospital-
31 acquired and healthcare-associated infections by integrating genomic and epidemiological data with
32 hospital Infection Prevention and Control (IPC) systems. While clusters involving care home residents
33 and healthcare workers were observed, the study was not intended to analyse care home
34 transmission specifically and focused on samples tested at CUH to provide information for IPC on

1 potentially hospital-acquired infections. Previous epidemiological studies of COVID-19 specifically in
2 care homes have been limited in population size, temporal scale and/or the amount of genomic data
3 included (Arons et al., 2020; Burton et al., 2020; Graham et al., 2020; Kemenesi et al., 2020; Quicke
4 et al., 2020). Here, genomic epidemiology is used to investigate viral transmission dynamics in care
5 home residents across the East of England (EoE), the fourth largest of the nine official regions in
6 England (Office for National Statistics, 2011). Several key questions of public health concern are
7 addressed: What is the burden of care home-associated COVID-19 tested in the region? What are
8 the outcomes for care home residents admitted to hospital with COVID-19? Does SARS-CoV-2 spread
9 between care home residents from the same care home via a single introduction and subsequent
10 transmission, or through multiple independent acquisitions of the virus among residents? Finally, is
11 there evidence of viral transmission between care homes and hospitals?

12 **Results**

13 **COVID-19 case numbers from care home and non-care home residents included in the** 14 **study**

15 7,406 SARS-CoV-2 positive samples from 6,600 individuals were identified in the study period (26th
16 February to 10th May 2020) (Figure 1), and care home residency status was determined in 6,413
17 (Figure 1, supplement 1) – the remaining 187 cases had missing address data and care home status
18 could not be determined. The samples were tested at the Public Health England (PHE) Clinical
19 Microbiology and Public Health Laboratory (CMPHL) in Cambridge, which receives samples from
20 across the East of England (EoE). Positive cases came from 37 submitting organisations including
21 regional hospital laboratories and community-based testing services (Supplementary Materials). The
22 proportion of samples coming from different sources changed over the study period (Figure 1,
23 supplement 2). This likely reflects a combination of regional hospitals establishing their own testing
24 facilities, increasing availability of community testing in the UK, and the implementation of national
25 policies that increased the scope of care home testing (Figure 1, supplement 3). Overall, the study
26 population included almost half of the COVID-19 cases diagnosed in the EoE at this time (Public
27 Health England, 2020a), with the remainder being tested at other laboratory sites.

28

29 1,167 / 6,413 (18.2%) of the study population were identified as care home residents from 337 care
30 homes. 193 / 337 (57.3%) care homes were residential homes and 144 / 337 (42.7%) were nursing
31 homes, with the majority located in five counties across EoE: Essex, Hertfordshire, Bedfordshire,
32 Suffolk and Cambridgeshire (Figure 2). This represents around half of the care homes in the East of
33 England which had reported suspected or confirmed COVID-19 outbreaks to PHE as of 11th May

1 2020 (UK government, 2020a). As expected, care home residents were older than non-care home
2 residents (median age 86 years versus 65 years, respectively ($P < 10^{-5}$, Wilcoxon rank sum test))
3 (Table 1). There was a median of 2 cases per care home (range 1-22), with a highly skewed
4 distribution: the 10 care homes (top 3%) with the largest number of cases contained 164 / 1167
5 (14.1%) of all care home cases (Figure 2, supplement 1).

6
7 The epidemic curve for all cases tested at the Cambridge CMPHL peaked in the end of March and
8 early April (Figure 3). Care home residents comprised a greater proportion of cases in late April and
9 May than in March (Figure 3A, Table 2). This may reflect the changing profile of samples submitted
10 to the CMPHL, as more regional hospitals had their own testing capacity and a greater number of
11 samples were submitted from community testing organisations in later weeks. However, a similar
12 trend was observed for patients tested at Cambridge University Hospitals, with the proportion of
13 community-onset care home-associated cases increasing from <5% in March to a peak of 14/49
14 (28.6%) in mid-April (Figure 3B, Table 3). This may suggest that transmission involving care home
15 residents took longer to decline following national lockdown (implemented on 23rd March 2020 in
16 the UK) than transmission in the non-care home general community.

17 **Mortality of COVID-19 infections for care home and non-care home residents tested in** 18 **hospital**

19 464 / 6,600 (7%) individuals with positive COVID-19 tests were patients tested at Cambridge
20 University Hospitals. Richer metadata were available for this subset of patients via the hospital
21 electronic records system. 72 / 464 (15.5%) COVID-19 patients diagnosed at CUH were identified as
22 care home residents (Table 1, Figure 3B), of which <7% were admitted to the intensive care unit
23 (ICU) and 34/72 (47.2%) died within 30 days of their first positive test (precise values not shown
24 where the number of individuals is equal to or below five, to protect patient anonymity). In
25 comparison, amongst non-care home residents, 84 / 392 (21.4%) were admitted to the ICU and 78 /
26 392 (19.9%) died within 30 days of diagnosis. In a logistic regression analysis, older age, care home
27 residency, ICU admission, and lower diagnostic cycle threshold (Ct) values were associated with
28 increased odds of mortality at 30 days from diagnosis (Figure 4, Table 4). The odds of mortality
29 within 30 days of diagnosis did not differ between residents at nursing homes versus residential
30 homes in a separate logistic regression analysis.

31 **Identifying viral clusters within care homes using genomic and epidemiological data**

32 Genome sequence data were available for 700 / 1,167 (60.0%) care home residents from 292 care
33 homes (Figure 3, supplement 1). There was a median of 8 single nucleotide polymorphisms (SNPs)

1 separating care home genomes, compared to 9 for randomly selected non-care home samples
2 ($P=0.95$, Wilcoxon rank sum test) (Figure 5, supplement 2), similar to the EoE region described
3 previously (Meredith et al., 2020). The proportion of viral lineage B.1.1 increased over the study
4 period in both care home residents and non-care home residents (Figure 5, Table 5), consistent with
5 European trends (Alm et al., 2020). With ongoing viral evolution, descendent lineages of B.1 and
6 B.1.1 also rose in frequency and were commonly found in England during the relevant time period.
7 This suggests that the SARS-CoV-2 lineages circulating in care homes were similar to those found
8 across the EoE outside of care homes. Consistent with this, care home and non-care home samples
9 were intermixed across the phylogenetic tree (Figure 6A), suggesting viral transmission could pass
10 between care homes and non-care home settings. No new viral lineages from outside the UK were
11 observed, which may reflect the success of travel restrictions in limiting introductions of new
12 lineages into the general population.

13

14 The ten care homes with the largest number of genomes (top ~3%) contained 102 / 700 (14.6%) of
15 all samples with genomic data available. For several of these ten care homes, all cases clustered
16 closely together on a phylogenetic tree with zero or 1 pairwise SNP differences, consistent with a
17 single “outbreak” spreading within the care home (where an outbreak is defined as two or more
18 cases linked in time or place (McAuslane and Morgan, 2014)) (Figure 6 and Figure 6, supplement 1).
19 By contrast, several care homes were “polyphyletic”, with cases distributed across the phylogenetic
20 tree and higher pairwise SNP difference counts between samples, consistent with multiple
21 independent introductions of the virus among residents.

22

23 The probability of two cases having linked transmission in an epidemiologically meaningful
24 timeframe (for example direct transmission or within one or two intermediate hosts – likely the
25 maximum practical limit for investigating the source of infection for a positive case) is a function of
26 several factors. These include the pairwise genetic differences between viruses and their
27 phylogenetic relatedness, the time difference between cases, and the opportunities for infection
28 between people (for example, the frequency, duration and extent of close contact). For this
29 continuous probability distribution, a pragmatic cut-off was used of $\geq 15\%$ likelihood that samples
30 were connected by ≤ 2 intermediate hosts, using a previously published algorithm called *transcluster*
31 (Stimson et al., 2019), adjusted for SARS-CoV-2 (Methods). Each care home was considered as a
32 separate microcosm of transmission and the number of viral clusters per care home was estimated,
33 with separate clusters implying distinct acquisition events among residents.

34

1 This clustering method identified 409 transmission clusters from 292 care homes (median 1 cluster
2 per care home, range 1-4). Within each cluster, 673 / 775 (86.8%) of pairwise links had zero or 1
3 pairwise SNP differences (maximum 4), and 756 / 775 (97.5%) were sampled <14 days apart
4 (maximum 22 days) (Figure 7, supplements 4-5). Clusters had a smaller distribution of sampling
5 dates than for the total cases within each care home, as expected (Figure 7, supplement 6). For the
6 170 / 292 (58%) care homes with 2 or more cases with genomic data (578 individuals), there was a
7 median of 9 (IQR: 4 – 15) days from the first case to the last case within each care home, up to a
8 maximum of 50 days. In contrast, more clusters comprised only a single individual than for care
9 homes, and for the 133 / 409 (33%) clusters with 2 or more cases with genomic data (424
10 individuals), there was a median of 5 (IQR: 1 – 11) days from the first case to the last case within
11 each cluster, up to a maximum of 22 days ($P < 10^{-5}$, Wilcoxon rank sum test comparing date
12 differences for care homes vs clusters with 2 or more samples; comparison shown in Figure 7,
13 supplement 6). The median and interquartile range for pairwise date differences between all
14 samples within each cluster is shown in Figure 7, supplement 7, and the date ranges for all care
15 homes and clusters is in Supplementary Materials.

16
17 Transmission networks for the ten care homes with the largest number of genomes are shown in
18 Figure 7A, indicating linked transmission clusters among residents based on the model assumptions
19 and probability threshold (full dataset shown in Figure 7, supplement 1). Consistent with the
20 phylogeny shown in Figure 6A, some care homes contained a single transmission cluster involving
21 multiple cases (e.g. CARE0314), while others comprised multiple independent clusters (e.g.
22 CARE0061) (Table 6). While care homes frequently had more than one introduction of the virus
23 among residents (i.e. >1 cluster), there was typically a single dominant cluster responsible for the
24 majority of cases within each care home. Of the 170 care homes with 2 or more residents with
25 genomic data (comprising 578 / 700 (82.6%) care home residents with genomic data), 111 / 170
26 (65.3%) had a dominant cluster responsible for >50% of all cases in the care home. This rises to 74 /
27 90 (82.2%) of care homes with three or more residents with genomic data.

28
29 The contribution made by genomic data in defining care home clusters was quantified. Without
30 genomic data (or access to more detailed epidemiology such as accommodation sub-structuring
31 within care homes), clustering can only be based on temporal differences between cases. For
32 example, if two groups of COVID-19 cases occur several months apart within a care home they could
33 be inferred to have resulted from (at least) two separate introductions. However, this method
34 cannot account for multiple introductions occurring around the same time, as may happen when

1 community transmission is high. To quantify the impact made by adding genomic data, which can
2 distinguish between genetically dissimilar viruses introduced at similar times, the *transcluster*
3 algorithm was repeated using the same parameters as for the main analysis but assuming all
4 genomes were identical. This yielded 316 clusters – 23% fewer than the 409 clusters yielded when
5 incorporating genomics. This suggests that genomics makes a significant contribution to defining
6 viral clusters; without genomic data, cluster sizes may be over-estimated and the number of
7 separate viral introductions under-estimated. This is illustrated by care home CARE0263, in which all
8 12 residents tested positive within 3 days of each-other, but these are divided into three separate
9 clusters by the *transcluster* algorithm (one dominant cluster of 9 cases, one cluster of 2 cases and a
10 single separate case (Table 6)); this is consistent with the phylogeny shown in Figure 6A, with
11 samples split into three branches along the tree. Without genomic data, the three clusters in
12 CARE0263 would have been impossible to distinguish.

13 **Links between care homes and hospitals**

14 Links between care homes and hospitals were investigated for the 700 care home residents with
15 genomic data available. 694 / 700 (99%) of the care home residents with genomic data had NHS
16 numbers available, which were linked to national hospital admissions data (Methods) (Table 7). 470
17 / 694 (67.7%) care home residents had at least one hospital admission within the study period, and
18 398 / 694 (57.3%) were deemed to have been admitted to hospital with COVID-19 (i.e. their first
19 positive sample was taken within 2 days prior to admission up to 7 days post-admission). 40 / 694
20 (5.8%) cases were categorised as suspected hospital-acquired COVID-19 infections, defined as first
21 positive test being 7 days or more after their hospital admission date and prior to their discharge
22 date (N=13) or within 7 days following their hospital discharge (N=27) (Table 7). 230 / 694 (33.1%)
23 individuals were discharged from hospital within 7 days of their first positive test, and thus could
24 potentially have been infectious at the time of hospital discharge (Byrne et al., 2020).

25 **Viral clusters linking care home residents and healthcare workers**

26 Potential transmission networks involving care home residents and healthcare workers (HCW) were
27 investigated for people tested at CUH (HCW data were not available outside of CUH). This analysis
28 comprised 54 care home residents tested at CUH and 76 HCW with genomic data available. Clusters
29 were defined using the same method as for the care home resident analysis (described above), but
30 allowing HCW to belong to clusters from multiple care homes, so residents from several care homes
31 could be linked to the same HCW. 38 / 54 (70.4%) care home residents had possible links with HCW
32 using this relaxed threshold. However, on review of the medical records we could only identify

1 strong epidemiological links for 14 / 54 (26.0%) residents from two care home clusters, CARE0063
2 and CARE0114. The CARE0063 cluster has been described previously (Meredith et al., 2020) and
3 includes care home residents, a carer from that same care home and another from an unknown care
4 home, paramedics and people living with the above. The CARE0114 cluster comprises several care
5 home residents and acute medical staff working at CUH who cared for at least one of the residents.
6 The *transcluster* method does not assign probabilities for directionality of transmission and cannot
7 determine precise person-to-person transmission chains. While all residents from a care home
8 cluster may link to a given HCW, in reality the resident-HCW transmission event may have only
9 involved one of the residents from that cluster, so the proportion of residents with links to HCW may
10 be inflated. Nonetheless, these data show that two care home clusters involved HCW, one based
11 mainly in the community and the other with hospital-based staff at CUH.

12

13 Residents from a third care home, CARE0273, also had strong transmission links to the paramedics
14 and carers involved in the CARE0063 cluster. These two care homes are within 1 kilometre of each-
15 other and the cases cluster together on the phylogenetic tree, raising the possibility of shared
16 transmission between them. A plausible transmission network connecting the residents at these two
17 care homes and the shared HCWs could be made with at most zero SNPs and three days between
18 sampled cases (Figure 7B); these links are in the top 1.1% of all pairwise transmission probabilities
19 inferred using the *transcluster* algorithm. However, without confirmatory epidemiological data this
20 interpretation remains speculative.

21 **Discussion**

22 The genomic epidemiology of SARS-CoV-2 in care homes in the East of England was investigated.
23 Care home residents comprised a large fraction of COVID-19 diagnoses in the “first wave” of the
24 pandemic in this region: up to a quarter of patients in the peak weeks of late March and early April
25 tested at CUH were admitted from care homes. Older age and being from a care home were
26 correlated with each other and were both associated with significantly increased odds of mortality
27 within 30 days of diagnosis. Care home residents thus bore a high burden of COVID-19 infections and
28 mortality.

29

30 A smaller proportion of care home residents were admitted to ICU compared with people who were
31 not from care homes. What treatments a patient receives, including the invasive treatments
32 provided in intensive care, are complex and individualised decisions based on risk-benefit
33 assessments involving patients, their families and carers, and healthcare professionals (ICS, 2020;

1 NICE, 2020). Of note, non-invasive respiratory support (such as continuous positive airway pressure,
2 high-flow nasal oxygen therapy and non-invasive ventilation) are routinely provided outside ICU in
3 many UK centres. Despite care home residents being at higher risk of severe COVID-19, and being
4 under-represented in ICU, admission to ICU was still correlated with significantly increased mortality.
5 This is likely because patients admitted to ICU have more severe disease, typically requiring more
6 intensive treatments such as organ support.

7
8 Viral clusters were defined within each care home by integrating temporal and genetic differences
9 between cases. This provides a “high resolution” picture of viral transmission; without genomic data,
10 separate introductions of the virus occurring around the same time are impossible to distinguish.
11 Care homes frequently experienced “outbreaks” of multiple cases within clusters (the largest of
12 which had >10 residents), consistent with substantial person-to-person transmission taking place
13 within care homes. Care homes also frequently had multiple distinct clusters (up to 4), consistent
14 with independent acquisitions of COVID-19 among residents – however, a single dominant cluster
15 usually comprised the majority of samples within each care home. The majority of care home
16 residents in the genomic analysis did not acquire COVID-19 in hospital. In the context of a national
17 lockdown, the most likely location they acquired their infection was the care home. The high
18 frequency of care home outbreaks may reflect the underlying vulnerability of this population to
19 COVID-19 and the challenges of infection control in care homes. In contrast, the UK as a whole had
20 an average of 2.37 people per household in 2019 (ONS, 2019a) and in the East region only 2.2% of
21 households were made up of two or more unrelated adults (6.2% in London) (ONS, 2019b).

22
23 These findings emphasise the importance of limiting viral transmission within care homes in order to
24 prevent outbreaks. Given there is increasing evidence for asymptomatic and presymptomatic
25 transmission of SARS-CoV-2 (Arons et al., 2020; Goldberg et al., 2020; He et al., 2020), isolating
26 residents or staff when they develop symptoms is not sufficient to prevent within-care home spread
27 once the virus has entered the care home. Certain measures may be required on an ongoing basis
28 within care homes when there is sustained community transmission, even when no outbreak is
29 suspected (at least until the morbidity and mortality of the virus in older people has been reduced
30 substantially through vaccination or treatments). These may include use of appropriate Personal
31 Protective Equipment (PPE) for staff and visitors (including visiting healthcare professionals and
32 friends and family), rigorous hand hygiene, social distancing, and making use of larger, well-
33 ventilated rooms for social interactions or socialising outdoors, providing that this is practical and
34 safe (N. R. Jones et al., 2020). This is consistent with current national guidance for care homes in

1 England (Public Health England, 2020b; UK government, 2020b). Face coverings for residents
2 themselves when interacting socially in communal indoor areas could be considered, if acceptable to
3 residents.

4
5 The majority of residents had hospital contact during the study period, indicating substantial
6 opportunity for infections to pass between care homes and hospitals in either direction. A third of
7 patients were discharged from hospital within 7 days of their first positive test, and thus were
8 potentially infectious at discharge. We identified transmission clusters that would be consistent with
9 COVID-19 spread between care home residents and HCW, based both in the community and in
10 hospitals. A previous study found that working across different homes was associated with higher
11 SARS-CoV-2 positivity among staff (Ladhani et al., 2020). Limiting the spread of COVID-19 between
12 care home residents, HCW and hospitals is a therefore another key target for infection control and
13 prevention.

14
15 There are several limitations to this study. First, not all of the COVID-19 cases from the East of
16 England have been included. Serology data suggest that 10.5% of all residents in care homes for
17 people aged 65 and older in England had been infected with SARS-CoV-2 by early June, the majority
18 of whom were asymptomatic (UK government, 2020c). The Cambridge CMPHL did not receive all of
19 the samples tested from the region; national data indicate around half of the COVID-19 cases
20 reported from EoE during the study were included. Viral sequence data were not available for 40%
21 of care home residents, as a result of missing samples, mismatches between sequences and
22 metadata, genomes not passing quality control filtering using a stringent threshold (<10% missing
23 calls), or sequences being unavailable at the time of data extraction. Viral cluster sizes may therefore
24 be underestimated.

25
26 Second, the nature of diagnostic testing sites changed during the study period as regional hospitals
27 developed their own in-house testing capacity and community testing laboratories were set up.
28 “Pillar 2” testing in the UK was outsourced to high-throughput laboratories during April 2020 and
29 performed an increasing proportion of community testing. It is possible that some care home
30 residents from the same care home could have been tested through different routes, with
31 symptomatic cases more likely to be tested in “Pillar 1” via the CMPHL (and included in this dataset),
32 and asymptomatic screening occurring more via the Pillar 2 laboratories. However, most care homes
33 in EoE only began systematic screening after the end of our study following the introduction of the
34 UK care home testing portal on 11th May 2020. Moreover, the *transcluster* algorithm allows for

1 “missing links” within a cluster (the threshold used assumed a $\geq 15\%$ probability of infections being
2 linked within ≤ 2 intermediate hosts), reducing the impact of missing care home cases on defined
3 clusters. The changing profile of COVID-19 testing in the UK between March and May 2020 should
4 therefore be factored into all interpretations of COVID-19 epidemiology from that period.

5

6 Third, defining who is a care home resident from large electronic healthcare records is challenging
7 and, despite substantial efforts (described in Methods), some care home residents may have been
8 missed. Using pre-defined coding such as care home CQC registration numbers when patients are
9 booked into hospital systems, rather than free-text data entry, would help considerably with care
10 home surveillance. Multiple rounds of electronic searches and manual inspection were undertaken
11 to identify as many care home residents as possible, and every care home resident included was
12 cross-referenced against a CQC database of registered care homes in England. The care homes
13 included for analysis should therefore be accurate.

14

15 Fourth, low viral sequence diversity limits the power of genomics to infer transmission clusters.
16 Between-care home transmission was not investigated specifically because, unlike within-care home
17 cases, opportunities for transfer of SARS-CoV-2 between care homes cannot be assumed or inferred
18 from the data. This could be assessed in a dedicated prospective study gathering epidemiological
19 data on between-care home contacts. Even within care homes, it is possible some genetically similar
20 viruses are from unconnected introduction events. However, incorporating genomic data is more
21 accurate for excluding linked transmission than if only temporal data are available. Genomics can
22 thus be used to “rule out” cases as being part of a linked cluster if the genetic difference is greater
23 than would be expected given the viral mutation rate. This could be practically informative for care
24 homes (along with other organisations at risk of COVID-19 outbreaks like factories (Middleton et al.,
25 2020)), with implications for infection control procedures. Directionality of person-to-person
26 transmission cannot be inferred from the *transcluster* algorithm. Inferring the likelihood of
27 transmission direction between pairs of individuals requires integration with multiple forms of
28 epidemiological data, yielding a probabilistic estimate (Illingworth et al., 2020).

29

30 In conclusion, care homes represent a major burden of COVID-19 morbidity and mortality, with
31 transmission events introducing SARS-CoV-2 into care homes and subsequent transmission within
32 them. Genomic data can be used in outbreak investigations to define viral clusters; this is critically
33 dependent on integration with epidemiological data. The cut-offs we used for defining care home
34 clusters were pragmatic but plausible given current understanding of the biology and epidemiology

1 of SARS-CoV-2. Such cut-offs can be helpful for producing understandable outputs for biological and
2 public health interpretation (MacFadden et al., 2018; Stimson et al., 2019), and for focusing
3 investigations with limited public health resources. Future work will need to prospectively integrate
4 genomic and epidemiological data to rapidly identify viral clusters, thus enabling deployment of
5 infection control and public health interventions in real time.

6 **Methods**

7 **Study overview**

8 Data were collected on SARS-CoV-2 positive samples from the East of England, tested at the PHE
9 CMPHL in Cambridge, between 26th February and 10th May 2020. The CMPHL is a PHE diagnostic
10 laboratory that receives samples from across the East of England. The East of England is one of nine
11 official regions in England. In the 2011 census, it had a population of 5,847,000, one of the fastest
12 growing populations in England and Wales and the fourth largest population of the nine official
13 regions (Office for National Statistics, 2011). The most populous cities include Luton, Norwich,
14 Southend-on-Sea, and Peterborough (City Population, 2020). The 10th May was selected as a study
15 end-date because it encompassed the bulk of the “first wave” of the epidemic in the East of England.
16 Furthermore, prior to the 11th May 2020, systematic screening of all residents within care homes
17 was much less common and testing primarily occurred where there was a suspicion of an outbreak.
18 The UK government launched a national care home testing portal on 11th May 2020 (UK government,
19 2020d), in which all care home staff and residents were eligible for testing with priority for homes
20 caring for people aged 65 years or older. Ending the study on 10th May reduces the risk of bias which
21 may be introduced by uneven systematic screening, for example when comparing the population
22 genetics of care home and non-care home samples, if care homes undergo screening while non-care
23 home settings do not. During the study period the scope of testing in hospital, community and care
24 home settings changed several times, as eligibility criteria were modified (Figure 1, supplement 1).
25 When interpreting trends in COVID-19 cases in the UK during this period it is essential to consider
26 the changing capacity and policies surrounding testing.

27 **Diagnostic testing, metadata collection and genome sequencing**

28 For details on diagnostic testing, patient metadata collection and nanopore genome sequencing see
29 (Meredith et al., 2020). Briefly, CMPHL used an in-house generated and validated one-step RT q-PCR
30 assay detecting a 222-bp region of the RdRp genes, along with an MS2 bacteriophage internal
31 extraction control, using the Rotorgene™ PCR instrument. Samples that generated a Ct value ≤ 36
32 were considered positive. The study aimed to sequence all samples which tested SARS-CoV-2 PCR

1 positive at the CMPHL during the study period. Sequencing of every positive diagnostic sample could
2 not be performed, however, for the following reasons: (i) sample unavailability (e.g. diagnostic
3 samples being lost or discarded before they could be collected by the sequencing team); (ii) labelling
4 errors when assigning sequencing codes (which resulted in specimens being discarded); or (iii)
5 metadata mismatches (if the sample did not match to a metadata record downloaded from the
6 hospital electronic patient records system). Samples were either sequenced on site using Oxford
7 Nanopore Technologies or transported to the Wellcome Sanger Institute for Illumina sequencing.

8

9 Samples from Cambridge University Hospitals NHS Foundation Trust (CUH) and a selection of East of
10 England (EoE) samples were sequenced on site to provide rapid information on hospital-acquired
11 infections (Meredith et al., 2020). Nanopore sequencing (Oxford Nanopore Technologies) took place
12 in the Division of Virology, Department of Pathology, University of Cambridge, following the
13 ARTICnetwork V3 protocol and assembled using the ARTICnetwork assembly pipeline. The
14 sequencing workflow involved a directional sample flow as used in a diagnostic laboratory which
15 includes separated pre- and post-PCR areas, with dedicated equipment for each stage of the
16 process. All steps were performed in PCR cabinets which were cleaned using DNA removal solutions
17 and a UV decontamination cycle run after each batch. All sequencing batches included at least one
18 water negative control carried over from the reverse-transcription step. Mapped reads were
19 assessed in real-time during sequencing with RAMPART (Hadfield, 2020) and all data from batches
20 containing a contaminated negative control were discarded before sequence assembly. The
21 remaining EoE samples, where available, were sent to the Wellcome Sanger Institute (WSI) for
22 sequencing.

23

24 Sequencing at WSI used Illumina technology. cDNA was generated from SARS-CoV-2 viral nucleic
25 acid extracts and subsequently amplified to produce 400nt amplicons tiling the viral genome using
26 V3 nCov-2019 primers (ARTIC). This was followed by Illumina library generation using the NEBNext
27 Ultra II DNA Library Prep Kit for Illumina (New England Biolabs Inc, Cat. No. E7645L). Libraries were
28 amplified with KAPA HiFi Ready Mix (Kapa Biosystems, Cat. No. 07958927001) and uniquely indexed
29 with a 100 μ M i5 and i7 primer mix (50 μ M each) (Integrated DNA Technologies) to allow
30 multiplexing of up to 384 SARS-CoV-2 viral extracts into one sequencing pool. The PCR products
31 were pooled in equal volume and purified with an AMPure XP workflow (Beckman Coulter, Cat. No.
32 A63880). The purified pool was quantified by qPCR (Illumina Library Quantitation Complete kit, Cat.
33 No. KK4824) and sequenced on one lane of an Illumina NovaSeq SP flow cell (Illumina Inc, NovaSeq
34 6000 SP Reagent Kit v1.5 (500 cycles), Cat. No. 20028402), with XP workflow (Illumina Inc, NovaSeq

1 XP 2 lane kit v1.5, Cat. No. 20043130). Genomes were generated for each library's sequencing data
2 using bwa mem (Li, 2013) for alignment with MN908947.3 (Wu et al., 2020) as reference, samtools
3 (Li et al., 2009) for pileup and ivar (Grubaugh et al., 2019) for trimming and consensus generation, all
4 orchestrated by the ncov2019-artic-nf pipeline (Bull, 2020).

5
6 The WSI sequencing workflow also uses negative controls and the pass rate to date related to
7 negative controls is 90%. Sequencing read counts are considered after a clipping and minimum
8 alignment length filtering step (corresponding to data which is used to create consensus sequence or
9 variant calls). Such read counts for the samples analysed in this study were typically in the millions
10 (median: 4,497,543). If such read counts for the corresponding negative controls are >100 then the
11 samples are currently failed. This QC procedure was introduced for samples analysed on or after the
12 18th of April. Of the 1,007 samples analysed in this study sequenced at WSI (503 care home
13 residents and 504 non-care home residents), 749 were sequenced once this workflow was
14 established, 242 were sequenced before this but had a negative control and 16 did not have a
15 negative control. If we apply the current criteria then 38 of these earlier samples would have failed
16 ($38/1400 = 2.7\%$ of the analysed samples). 26 of these 38 samples are non-care home samples and
17 12 are from care homes. Of the 12 care home samples ($12/700 = 1.7\%$ total care home genomes
18 analysed), 1 belongs to one of the "top 10" care homes with the largest number of genomes, care
19 home CARE0063, which comprises a single cluster of 12 genomes using the *transcluster* algorithm,
20 described in main text. Thus, the main result of our genomic cluster analysis (that multiple
21 introductions are often observed in care homes, but typically a single dominant cluster causes most
22 of the cases) would not be altered by the small number of early genomes included that would now
23 be excluded by current criteria.

24
25 Sequences were available from both Illumina and Nanopore platforms for eight care home residents
26 included in the study (in all cases the Illumina data were used for the study analysis). In 7/8 cases the
27 sequence pairs were identical. In one case there were two SNP differences between the consensus
28 fasta sequences: C1884T and C16351T; for both SNPs the Illumina sequence matched the reference
29 genome (C) and the nanopore sequence had the alt call (T). These are not included among a list of
30 previously identified sites that are highly homoplasic or have no phylogenetic signal and/or low
31 prevalence (De Maio et al., 2020). The sequence pairs are shown below:

32

Illumina sample - COG-UK ID	Illumina sample - date	Nanopore sample - COG-UK ID	Nanopore sample - date	Pairwise SNP difference
--------------------------------	---------------------------	--------------------------------	---------------------------	-------------------------------

CAMB-761D5	30/03/2020	CAMB-7B088	11/04/2020	zero
CAMB-1AF1F0	30/04/2020	CAMB-1AD8A2	30/04/2020	zero
CAMB-1AE7C2	30/04/2020	CAMB-1AC269	30/04/2020	2
CAMB-80590	09/04/2020	CAMB-789BD	06/04/2020	zero
CAMB-1AB23D	20/04/2020	CAMB-840B9	26/04/2020	zero
CAMB-83AAD	15/04/2020	CAMB-8416B	25/04/2020	zero
CAMB-1ABE2A	21/04/2020	CAMB-8468A	27/04/2020	zero
CAMB-1AB631	21/04/2020	CAMB-1ABF18	27/04/2020	zero

1

2 As with all the sample dates used, the above dates are based on sample collection date where
3 available, with missing data substituted with the date of receipt in the laboratory. SNP differences
4 were identified from a vcf file produced from the alignments using the package *snp-sites* v 2.5.1
5 (Page et al., 2016), command:

6 `snp-sites -v alignment_file.aln`

7

8 In Meredith *et al.* 2020, out of 14 sample pairs sequenced both by Illumina at WSI and nanopore in
9 the University of Cambridge there were zero SNP differences at positions where both sequences had
10 made a call (Meredith et al., 2020). There are several reasons why pairwise comparisons between
11 different sequences from the same individual may not be identical, even if both sequences are
12 produced using the same technology. When the cycle threshold (Ct) of a sample is near the limit of
13 detection sensitivity, and/or RNA is degraded (eg. due to delays between sampling and sequencing
14 at room temperature), it is likely that amplicons that are not as efficiently amplified by the multiplex
15 PCR may have low read coverage, or could be more sensitive to amplification bias. In this case, the
16 samples both had high Ct values: CAMB-1AE7C2 (sequenced by Illumina at WSI) had Ct value of 30
17 and CAMB-1AC269 (nanopore sequenced in Cambridge) had a Ct value of 31. Median Ct value for
18 the 700 care home residents with genomes analysed was 24 (interquartile range: 20-27) (data
19 displayed in Table 1). If an individual is infected with more than one clone at significant frequency, it
20 is also possible for stochastic variation in read counts for the two variants to yield different
21 consensus calls at the variant locus. However, larger studies have systematically evaluated
22 sequencing quality for SARS-CoV-2 between Oxford Nanopore Technology (ONT) and Illumina, and
23 demonstrated highly accurate consensus-level sequence determination (Bull et al., 2020). Given this
24 degree of consensus sequence accuracy, and because *transcluster* uses a transmission probability
25 cut-off based on integrating pairwise SNP and temporal differences (rather than relying solely on a
26 strict SNP cut-off), limited sequencing noise is unlikely to have a substantial impact on the clusters
27 identified.

28

1 COG-UK IDs and GISAID accession numbers for genomes analysed in this study are included in
2 Supplementary Materials, along with a complete author list for the COG-UK consortium.

3 **Sample selection**

4 As described in (Meredith et al., 2020), patient metadata were downloaded daily from the electronic
5 medical record system (Epic Systems, Verona, WI, USA) and metadata manipulations were
6 performed in R (v 3.6.2) using the *tidyverse* packages (v 1.3.0) installed on CUH computers. Positive
7 samples were collected and assigned either for nanopore sequencing on site (focusing on CUH
8 samples and a randomised selection of EoE samples), or sent to WSI for Illumina sequencing.
9 Metadata were uploaded weekly to the MRC CLIMB system as part of the COG-UK Consortium.
10 Samples included healthcare workers (HCW) tested in the CUH HCW screening programme (N. K.
11 Jones et al., 2020; Rivett et al., 2020), all of which were nanopore sequenced on site.

12 **Identifying care home residents**

13 Care home residents were identified using a two-stage data mining approach followed by manual
14 inspection and linking of putative care home addresses to care homes registered to the Care Quality
15 Commission (CQC).

17 *Step 1: Search terms in patient address fields*

18 Patient address lines 1 and 2 were searched for the following list of key phrases (not case sensitive)
19 in their electronic healthcare records; if any phrases were present the patient was labelled as being
20 from a care home:

- 21 · "residential home"
- 22 · "care home"
- 23 · "nursing home"
- 24 · "care centre"
- 25 · "care hom"
- 26 · "nursing hom"
- 27 · "residential hom"
- 28 · "carehome"

29 This identified 765 patients as being care home residents.

30

1 *Step 2: Matching location names to CQC registered care facilities*

2 Many care homes do not have the above list of phrases in their address names. To capture these
3 facilities, we used the publicly available database of care homes registered to the CQC, the
4 independent regulator of health and adult social care in England. All organisations providing
5 accommodation for persons who require nursing or personal care must be registered with the CQC,
6 including care homes with or without nursing care (CQC, 2020b). Details of the CQC registration
7 scope can be found in “The scope of registration (Registration under the Health and Social Care Act
8 2008)”, March 2015, available at this link as of 24th June 2020: [\(CQC, 2015\)](#).

9

10 The file “CQC care directory – with filters (1 June 2020)” was accessed on 23rd June 2020 from the
11 CQC website: (CQC, 2020c), and the following filters were applied:

- 12 - Total facilities in CQC database: N = 49,516
- 13 - “Carehome?” column filtered to “Y”: N = 15,507*
- 14 - Only care homes for which the “Location Postal Code” column matched at least 1 postcode
15 from the dataset of 6,600 patients were included, yielding N = 444 care homes.**
- 16 - Following manual review and consistifying postcodes with the sample metadata, a set of 469
17 CQC registered care homes were included.***

18

19 *Filtering using the “carehome?” column was based on advice given after correspondence with the
20 CQC.

21 ** Requiring CQC registered care homes to match postcodes from the patient dataset minimised the
22 number of “false positives” – patients whose address name matched a CQC registered care home
23 name by coincidence.

24 *** 25 CQC registered care homes were added following manual review of the identified putative
25 care home residents, who had a different postcode documented in the electronic healthcare records
26 for the same care home, yielding the final “CQC EoE care home search set” of 469 care homes.

27

28 We then used the values from the “Location name” column of the filtered CQC dataset (i.e. the care
29 home facility names) as search phrases for address line 1 in the patient database. Any patients with
30 exactly matching phrases were labelled as care home residents. This increased the number of care
31 home residents identified by a further 382 to 1,147, i.e. around one third of care home residents
32 were identified using CQC facility names and would have been missed by relying on generic care
33 home-related search phrases alone.

34

1 *Step 3: Manual inspection and data clean up*

2 Address lines for the non-care home patients were manually inspected; this identified a further 89
3 care home residents. Most of these had not been detected in steps 1 and 2 due to spelling or
4 formatting issues with the patient addresses (e.g. short-hand abbreviations used for care home
5 names, or inclusion of extra details like flat number meaning the string did not match a CQC care
6 home name exactly).

7

8 Next, address lines for the care home residents were manually inspected and 14 were deemed not
9 to be care home residents. Most of these were due to unrelated locations sharing the same address
10 name as a CQC registered care home. The manual filtering steps thus yielded a care home resident
11 count of $1,147 + 89 - 14 = 1,222$. Address line 1 for all 1,222 care home residents was manually
12 inspected and formatted to ensure residents from the same care home had matching terms in this
13 column. This was necessary due to discrepant address entrance formats for identical care homes;
14 without this step, residents from the same care home would be incorrectly assigned to different
15 anonymised care home codes.

16

17 *Step 4: Linking care home addresses to CQC registered care homes*

18 First line of patient address and postcodes were matched to care home names and postcodes from
19 the CQC EoE care home search set (described above). Any discrepancies (care homes not matching
20 the CQC data) were manually inspected and in the majority of cases the discrepancy could be
21 reconciled (e.g. alternative name or postcode used for the same care home). In 55 cases a “care
22 home” was reclassified to non-care home, either because the address was independent housing with
23 a matching name to a care home by coincidence, or because a care facility was determined by CQC
24 definitions to not be a care home - e.g. several mental health community hospitals, drug
25 rehabilitation centres, and supported living environments were excluded. This yielded the final
26 analysis set of $1,222 - 55 = 1,167$ care home residents, from 337 care homes. All 337 care homes
27 included were therefore linked to CQC data; in two cases the care home had been previously
28 registered but had since been “archived”, and the most recent CQC data for defining whether
29 residential or nursing care was being provided was used.

30

31 Care home location IDs assigned by the CQC were turned into anonymised codes (format: CARE
32 followed by a 4-digit numeric code). Care homes were classified as “residential homes” or “nursing
33 homes” using the CQC data column “Service type - Care home service with nursing” filtered to “Y”
34 for care homes with nursing, and column “Service type - Care home service without nursing” = “Y”

1 for care homes without nursing (“residential homes”). If both fields were “Y” then the care home
2 was coded as being a nursing home.

3 **Linking care home data to CUH acute medical testing data**

4 The dataset of 7,407 PCR-positive samples with metadata were collected prospectively as part of the
5 COG-UK study in Cambridge. Data on CUH acute care testing, including categorisations of whether
6 infections were community- or hospital-acquired (definitions provided in (Meredith et al., 2020)) and
7 data on patient outcomes (mortality at 30 days and ICU admissions), were collected separately as
8 part of CUH and national monitoring. During the study period, 464 patients tested positive for
9 COVID-19 at CUH.

10

11 When merging the metadata collected for COG-UK (including the above care home categorisations)
12 with CUH acute testing data, 71 care home residents tested at CUH were identified. However, there
13 were 23 samples that had tested positive in CUH that were not in the COG-UK dataset. 21/23 of
14 these were tested on the SAMBA platform at CUH (Collier et al., 2020), which is not PCR-based;
15 sequencing was not possible for these samples owing to rapid RNA degradation. For technical
16 reasons, SAMBA results were not included in the data collected prospectively in the Cambridge COG-
17 UK study. The remaining two discrepancies were not captured in the electronic patient record
18 downloads, which likely reflects periods where the download processes and coding methods were
19 being established. Of the 23 missing samples, 20 were community-onset community-associated, two
20 were hospital-onset indeterminate healthcare-associated, and one was a healthcare worker. These
21 are counted as such and depicted with the above categorisations in the CUH epidemic curve shown
22 in Figure 3B. Of the 23 CUH samples missing from the Cambridge COG-UK dataset, one was
23 determined to be a care home resident, bringing the total CUH care home residents analysed to 72.

24 **Statistics**

25 All statistical analyses were performed in R. The logistic regression model used to estimate odds of
26 30-day mortality was coded as follows:

```
27 glm.fit <- glm(mortality_30_days ~ age + sex + care_status + ICU_admission +  
28 diagnostic_ct_value, data=data, family=binomial)  
29 summary(glm.fit)
```

30 Odds ratios and 95% confidence intervals were derived by exponentiating the model coefficients:

```
31 exp(cbind(coef(glm.fit), confint(glm.fit)))
```

1 To produce the plot of odds ratios shown in Figure 4, the age and diagnostic Ct value continuous
2 variables were transformed into binary categoricals using cut-offs of age ≥ 80 years and Ct value < 20 .
3 Wilcoxon rank sum tests performed in R using command format:
4 `wilcox.test(x, y, alternative = "two.sided", conf.level = 0.95)`
5 *P*-values below 10^{-5} are not reported.

6 **Selecting randomised sample of non-care home residents as comparison group**

7 A randomised sample of non-care home residents was selected to use as a control group for
8 comparison of viral lineage composition against the care home residents. Because this group was
9 intended to be representative of non-care home community-acquired transmission, we applied the
10 following inclusion criteria prior to randomisation:

- 11 • Patient address available
- 12 • Not one of the identified care home residents
- 13 • Not a healthcare worker (information only available for people tested at CUH)
- 14 • Not a CUH case of indeterminate, suspected or definite hospital acquired infection
- 15 • Not living in a long-term care facility other than a care home (e.g. mental health
16 hospital, rehabilitation unit, *etc*)
- 17 • Not living in a prison

18 We attempted to have a roughly equivalent representation of nanopore and WSI sequenced
19 samples as present in the care home database. Samples were selected using the R randomisation
20 command `sample_n()` from available genomes in the CLIMB database passing QC filters. Having
21 identified 698 samples, any cases with matching addresses that had been excluded were added to
22 yield the final set of 700 non-care home genomes for comparison. Of the 700 non-care home
23 samples included, we note that there were five instances of pairs of samples sharing the same
24 address; in all five cases the pairwise SNP difference was zero or 1, and in 4/5 cases the people
25 shared the same surname. This non-care home comparison set is not part of the care home viral
26 cluster analysis performed using the *transcluster* algorithm.

27 **Care home viral phylogenetics and cluster analysis**

28 Consensus fasta sequences were downloaded from the MRC-CLIMB website
29 (<https://www.climb.ac.uk/>) (Connor et al., 2016). Genomes were de-duplicated (1 genome per
30 person) and passed through quality control (QC) filtering using the same criteria as in (Meredith et
31 al., 2020): genome size $> 29\text{Kb}$, N count < 2990 (i.e. $> 90\%$ coverage). Where there were multiple

1 sequences from the same patient, the sequence passing QC filters that was collected first was used
2 for genomic analysis (closest to the onset of symptoms).

3

4 The 700 de-duplicated viral genomes from care home residents passing QC were aligned using
5 MAFFT (v 7.458) (Kato and Standley, 2013) with default settings. Command:

```
6 "/PATH/mafft" --retree 2 --inputorder "multi_fasta_filename.fasta" > "alignment_filename"
```

7

8 A SNP difference matrix was produced from the alignment using *snp-dists* v 0.7.0 (Seemann, 2020)
9 installed in a conda environment, run with the following command:

```
10 snp-dists -c alignment_filename.aln > snp_diff_matrix_filename.csv
```

11

12 The SNP difference matrix was manipulated in R using the *Matrix* and *tidyverse* packages to generate
13 the SNP difference histogram and boxplots.

14

15 Phylogenetic trees were generated using IQ-TREE (v 1.6.12 built Aug 15 2019). An alignment was
16 generated as above including a reference genome from Wuhan, China, collected December 2019
17 and used to root the tree (GISAID ID: EPI_ISL_402123). The IQ-TREE Model Finder Plus option was
18 used (Kalyaanamoorthy et al., 2017) which searches from a database of available nucleotide
19 substitution models and selects the best fit to the analysis, command line:

```
20 ~/PATH/iqtree -s alignment_filename -m MFP
```

21

22 The best-fit nucleotide substitution model according to BIC was GTR+F+R2. The tree shown in this
23 manuscript was produced using the GTR+F+R2 model with the ultrafast bootstrap option (Hoang et
24 al., 2018) run through 1,000 iterations to estimate branch support values, using command:

```
25 ~/PATH/iqtree -s alignment_filename -m GTR+F+R2 -bb 1000
```

26

27 Newick trees were manipulated in *FigTree* (v 1.4.4) to root on the Wuhan sample and put in
28 increasing node order. Trees were visualised initially using the microreact online tool (Argimón et al.,
29 2016), and Figure 6A was produced in R using *ggtree* (v 2.0.4) (Yu et al., 2017).

30

31 For the phylogenetic tree of all samples in the study (Figure 6, supplement 1), consensus fasta files
32 were downloaded from the COG-UK database (<https://www.cogconsortium.uk/data/>) accessed
33 01/12/2020. The same QC filtering described above was applied (genome size >29Kb, N count
34 <2990). Sequences passing QC were linked by their COG-UK IDs to individuals from this study. Of the

1 6,600 people in the study, 1,167 had been identified as care home residents and 700 / 1,167 (60.0%)
2 had genomes available that passed QC at time of the main analysis, leaving 5,246 non-care home
3 residents (187 were undetermined). 3,745 / 5,246 (71.4%) non-care home residents had genomes
4 available that passed QC (including the 700 randomly sub-sampled non-care home residents
5 described above). A multiple sequence alignment was produced in MAFFT and phylogenetic tree
6 produced using IQTREE, command line:

7 `iqtree -s alignment_all.aln -m GTR+F -nt AUTO -ntmax 16 -mem 16G -bb 1000`

8 The tree was manipulated in *FigTree* (v 1.4.4) and Figure 6 supplement 1 was produced in R using the
9 *ggtree* package as with Figure 6.

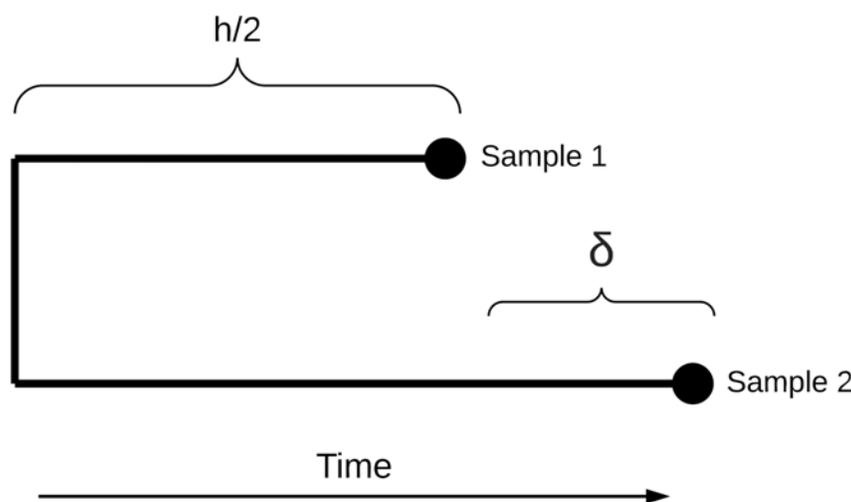
10 **Lineage assignment**

11 Viral lineages were assigned using the Pangolin COVID-19 Lineage Assigner web utility (COG-UK,
12 2020). Analysis was performed with Pangolin (Andrew Rambaut et al., 2020) version 1.1.14, lineages
13 version 2020-05-19-2. Contextual information about lineages was taken from (A Rambaut et al.,
14 2020), accessed 24/07/2020.

15 **Clustering**

16 Clusters were produced using an implementation of the *transcluster* algorithm (Stimson et al., 2019;
17 Tonkin-Hill, 2020). Instead of targeting the number of SNPs separating two genomes, the
18 *transcluster* algorithm proposes a probabilistic alternative which estimates the number of
19 intermediate transmission events separating two sampled genomes. The method takes into account
20 both genetic SNP distance as well as the time at which each sample was taken. The approach models
21 both the SNP distance and the number of intermediate hosts as a Poisson process. Using a
22 predefined evolutionary rate as well as an estimate of the generation time (the time between
23 transmission events), the method infers the distribution of the number of intermediate hosts
24 separating two samples.

25



1

2

3 Briefly, let N be the SNP distance separating two genomes and δ the time difference between when
 4 the samples were taken. We would like to estimate h , the time between the infection times of the
 5 two samples. The number of SNPs per unit time can be modelled as a Poisson process with
 6 evolutionary rate λ . Similarly, we assume the rate β at which the pathogen jumps to a new host is
 7 constant resulting in another Poisson process for the number of intermediate hosts given h and δ .
 8 We are thus interested in the probability that there are k intermediate hosts given N and δ which,
 9 following the derivation in Stimson *et al.*, 2019, can be written as:

$$P(k|N, \delta) = \int_{h=0}^{\infty} \mathcal{L}(h|N, \delta) P(k|h) dh$$

10

11

12 This can be expressed as the sum:

13

$$P(k|N, \delta) = \frac{\lambda^{N+1} \beta^k (n+k)!}{e^{\delta \beta n! k!} \sum_{i=0}^N \frac{(\lambda \delta)^i}{i!}} \sum_{i=0}^{N+k} \frac{\delta^{N+k-i}}{(N+k-i)! (\lambda + \beta)^{i+1}}$$

14

15

16 The implementation of *transcluster* assumed a viral mutation rate of 1e-3 substitutions/site/year
 17 (Fauver *et al.*, 2020) and generation time of five days, approximated by previous estimates of the
 18 serial interval of SARS-CoV-2 (He *et al.*, 2020; Zhang *et al.*, 2020). Days between first positive
 19 sampling date for pairs of individuals was used as a proxy for generation time. As above, where
 20 collection date was missing, the date the sample was received in the Cambridge PHE laboratory was

1 used. The resulting pairwise transmission probabilities were used to generate a pairwise distance
2 matrix and clustering was performed using single linkage hierarchical clustering with the R *hclust*
3 function. Links were only considered if they involved residents from the same care home; thus, the
4 largest theoretical number of clusters in this analysis would be 700 (every individual is their own
5 distinct cluster), and the smallest would be 292 (one cluster for each care home).

6
7 The relationship between the probability of infections being linked by ≤ 2 intermediate hosts and the
8 resulting number of care home clusters was explored. A higher threshold leads to more care home
9 clusters, with greater likelihood of linked transmission within each cluster than when using a lower
10 threshold. A pragmatic cut-off of $\leq 15\%$ probability was selected, yielding 409 clusters. The majority
11 of pairwise comparisons within clusters were zero or 1 SNP different and < 14 days apart.

12
13 For 16/700 (2.3%) genomes, the sample that produced the analysed sequence was not the first
14 positive test for that individual in the dataset. This could have occurred if the first positive test was
15 not sequenced, or the sequencing failed or did not pass QC filters. This could theoretically lead to
16 different clustering outcomes, if two cases were counted as further apart temporally than they really
17 were from the date of first positive swab. To ensure this had not biased our findings, the *transcluster*
18 analysis was re-run with identical thresholds using the date of first positive test for each individual
19 (keeping the same genomes). There was no change in the number of clusters identified ($n=409$).

20
21 To maintain study participant anonymity, care home residency status cannot be released publicly
22 linked to their COG-UK genome codes. However, an anonymised version of the same dataset
23 analysed in this study, with COG-UK sequence codes replaced by anonymised sample codes, can be
24 accessed via GitHub at <https://github.com/gtonkinhill/SC2-care-homes-anonymised>. This includes all
25 code and anonymised input data to reproduce the transmission analysis. Further discussion on data
26 release is provided in Supplementary Materials.

27 **Investigating hospital admissions for care home residents**

28 Hospital Episode Statistics (HES) data from 26th February to 10th May 2020 were linked to cases from
29 this study using matching NHS numbers. The data were accessed by the Public Health England
30 Healthcare Associated Infections (HCAI) division via the PHE Data Lake. This was possible for 694/700
31 (99%) of the care home residents with genomes available (used in the cluster analysis); six cases
32 could not be linked to admission data due to missing NHS numbers in the study metadata.

1 Hospital admission coding included transfer of care between medical units as separate admissions.
2 These were condensed into single admissions if the time interval between the preceding discharge
3 and the following admission was less than or equal to 1 day. i.e. an admission had to occur 2 days or
4 more after the preceding discharge to be counted as a new admission.

5 Hospital admission data were parsed to yield the following outputs:

- 6 • COVID-19 related hospital admission: first positive test date was -2 to +7 days inclusive from
7 a hospital admission date
- 8 • Suspected hospital acquired: first positive test date was +7 days from a hospital admission to
9 +7 days from a hospital discharge, inclusive. The people testing positive in the community
10 within 7 days of discharge from hospital are categorised as, “community onset, suspected
11 hospital acquired”; the people testing positive after 7 days from admission but before their
12 discharge are categorised as, “hospital onset, suspected hospital acquired”.
- 13 • For the 6 individuals with no NHS number, we assumed they were not discharged within 7
14 days of a positive test.

15 For the care home residents with community-onset, suspected hospital-acquired infections, the
16 number of days the patient had been admitted to hospital prior to their positive test was calculated.

17 **CUH HCW-Care home resident cluster analysis**

18 The analysis of transmission between healthcare workers (HCW) and care home residents focused
19 on CUH cases, where the richest metadata was available including HCW status.

20

21 Of 6,600 PCR-positive patients, 91 had been identified as HCW. 74 of these were from the CUH HCW
22 screening programme (which includes symptomatic, asymptomatic and household contact arms) (N.
23 K. Jones et al., 2020; Rivett et al., 2020) and 17 had presented acutely to CUH medical services, and
24 been identified as HCW during their initial medical clerking and subsequent note reviews. Of the 91
25 HCW, 76 had genomes available for analysis (breakdown: 56 samples identified through the CUH
26 HCW screening programme, 9 CUH HCW who presented to acute medical services at CUH, and 11
27 HCW from community settings (paramedics and care home workers) that had been flagged as HCW
28 through admission clerkings). Of 464 CUH cases in the study period, 72 were care home residents
29 (described above) and 54 of these had available genomes for analysis. The total combined analysis
30 set of CUH HCW and care home residents was therefore $76+54 = 130$.

31

1 The 130 genomes were aligned using MAFFT and underwent the same cluster analysis using the
2 *transcluster* algorithm as described above. Transmission links between care homes were excluded as
3 were links between HCWs. HCWs could belong to multiple clusters from different care homes to
4 allow for the possibility of a HCW seeding multiple care home infections. 21 clusters involving both
5 care home residents and HCWs were identified. Of the 54 care home residents, 38 had links with
6 HCWs within the 0.15 probability threshold. Medical notes for potential care home resident-HCW
7 transmission pairs were reviewed by author WLH as described in (Meredith et al., 2020), with cases
8 being categorised as strongly linked epidemiologically (e.g. the HCW documented in the care home
9 residents' medical notes); possibly linked (e.g. both working in the hospital at the same time but not
10 in the same wards); or no evidence of an epidemiological link.

11 **Ethics**

12 This study was conducted as part of surveillance for COVID-19 infections under the auspices of
13 Section 251 of the NHS Act 2006. It therefore did not require individual patient consent or ethical
14 approval. The COG-UK study protocol was approved by the Public Health England Research Ethics
15 Governance Group (reference: R&D NR0195).

16 **Funding**

17 This work was funded by COG-UK (supported by the Medical Research Council (MRC) part of UK
18 Research & Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome
19 Research Limited, operating as the Wellcome Sanger Institute); the Wellcome Trust; the Academy of
20 Medical Sciences; the Health Foundation; and the Cambridge NIHR Biomedical Research Centre.

21 **Acknowledgements**

22 We gratefully acknowledge the invaluable contributions of all members of the Wellcome Sanger
23 Institute Covid-19 Surveillance Team (www.sanger.ac.uk/covid-team) who have supported this
24 project. We would also like to thank Nick Donnelly for advice with statistical analyses, and the Public
25 Health England Hospital Acquired Infection (HCAI) division, in particular Rebecca Guy and Mehdi
26 Minaji, for assistance accessing hospital admission data for this study.

27 **References**

28 Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-stroh S. 2020. Geographical and
29 temporal distribution of SARS-CoV-2 clades in the WHO European Region , January to June

- 1 2020 1–8.
- 2 Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MTG, Yeats CA,
3 Grundmann H, Spratt BG, Aanensen DM. 2016. Microreact: visualizing and sharing data for
4 genomic epidemiology and phylogeography. *Microb genomics* **2**:e000093.
5 doi:10.1099/mgen.0.000093
- 6 Arons MM, Hatfield KM, Reddy SC, Kimball A, James A, Jacobs JR, Taylor J, Spicer K, Bardossy AC,
7 Oakley LP, Tanwar S, Dyal JW, Harney J, Chisty Z, Bell M, Methner M, Paul P, Carlson CM,
8 McLaughlin HP, Thornburg N, Tong S, Tamin A, Tao Y, Uehara A, Harcourt J, Clark S, Brostrom-
9 Smith C, Page LC, Kay M, Lewis J, Montgomery P, Stone ND, Clark TA, Honein MA, Duchin JS,
10 Jernigan JA. 2020. Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing
11 facility. *N Engl J Med* **382**:2081–2090. doi:10.1056/NEJMoa2008457
- 12 Bull M. 2020. A Nextflow pipeline for running the ARTIC network’s fieldbioinformatics tools with a
13 focus on ncov2019. <https://github.com/connor-lab/ncov2019-artic-nf>
- 14 Bull RA, Adikari T, Hammond JM, Stevanovski I, Ferguson JM, Beukers AG, Naing Z, Yeang M, Verich
15 A, Gamaarachichi H, Kim KW, Luciani F, Stelzer-Braid S, Eden J-S, Rawlinson WD, Van Hal SJ.
16 2020. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *bioRxiv*
17 2020.08.04.236893.
- 18 Burton JK, Bayne G, Evans C, Garbe F, Gorman D, Honhold N, McCormick D, Othieno R, Stevenson J,
19 Swietlik S, Templeton K, Tranter M, Willocks L, Guthrie B. 2020. Evolution and impact of COVID-
20 19 outbreaks in care homes: population analysis in 189 care homes in one geographic region.
21 *medRxiv* 2020.07.09.20149583. doi:10.1101/2020.07.09.20149583
- 22 Byrne AW, McEvoy D, Collins AB, Hunt K, Casey M, Barber A, Butler F, Griffin J, Lane EA, McAloon C,
23 O’Brien K, Wall P, Walsh KA, More SJ. 2020. Inferred duration of infectious period of SARS-CoV-
24 2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic
25 COVID-19 cases. *BMJ Open* **10**:e039856. doi:10.1136/bmjopen-2020-039856
- 26 City Population. 2020. East of England (United Kingdom): Counties and Unitary Districts &
27 Settlements - Population Statistics, Charts and Map.
28 <https://www.citypopulation.de/en/uk/eastofengland/>
- 29 COG-UK. 2020. Pangolin COVID-19 Lineage Assigner. <https://pangolin.cog-uk.io/>
- 30 Collier D, Assennato S, Sithole N, Sharrocks K, Ritchie A, Ravji P, Routledge M, Sparkes D, Skittrall J,
31 Warne B, smielewska A, RAMSEY I, Goel N, CURRAN M, ENOCH D, TASSELL R, LINEHAM M,
32 VAGHELA D, LEONG C, Gupta R. 2020. Rapid point of care nucleic acid testing for SARS-CoV-2 in
33 hospitalised patients: a clinical trial and implementation study. *Cell Reports Med* 100062.
34 doi:10.1101/2020.05.31.20114520

- 1 Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, Bull MJ, Richardson E, Ismail
2 M, Thompson SE, Kitchen C, Guest M, Bakke M, Sheppard SK, Pallen MJ. 2016. CLIMB (the
3 Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical
4 microbiology community. *Microb genomics* **2**:e000086. doi:10.1099/mgen.0.000086
- 5 Coronavirus testing - GOV.UK. 2020. <https://www.gov.uk/government/news/coronavirus-testing>
6 COVID-19 policy tracker | The Health Foundation. 2020. [https://www.health.org.uk/news-and-](https://www.health.org.uk/news-and-comment/charts-and-infographics/covid-19-policy-tracker)
7 [comment/charts-and-infographics/covid-19-policy-tracker](https://www.health.org.uk/news-and-comment/charts-and-infographics/covid-19-policy-tracker)
- 8 CQC. 2020a. Service types | Care Quality Commission. [https://www.cqc.org.uk/guidance-](https://www.cqc.org.uk/guidance-providers/regulations-enforcement/service-types#care-homes-without-nursing)
9 [providers/regulations-enforcement/service-types#care-homes-without-nursing](https://www.cqc.org.uk/guidance-providers/regulations-enforcement/service-types#care-homes-without-nursing)
- 10 CQC. 2020b. What is registration? | Care Quality Commission. [https://www.cqc.org.uk/guidance-](https://www.cqc.org.uk/guidance-providers/registration/what-registration)
11 [providers/registration/what-registration](https://www.cqc.org.uk/guidance-providers/registration/what-registration)
- 12 CQC. 2020c. Using CQC data | Care Quality Commission. [https://www.cqc.org.uk/about-](https://www.cqc.org.uk/about-us/transparency/using-cqc-data)
13 [us/transparency/using-cqc-data](https://www.cqc.org.uk/about-us/transparency/using-cqc-data)
- 14 CQC. 2015. The scope of registration.
- 15 Curran ET. 2017. Infection outbreaks in care homes: prevention and management, *Nursing Times*.
16 [https://www.nursingtimes.net/clinical-archive/infection-control/infection-outbreaks-in-care-](https://www.nursingtimes.net/clinical-archive/infection-control/infection-outbreaks-in-care-homes-prevention-and-management-14-08-2017/)
17 [homes-prevention-and-management-14-08-2017/](https://www.nursingtimes.net/clinical-archive/infection-control/infection-outbreaks-in-care-homes-prevention-and-management-14-08-2017/)
- 18 De Maio N, Walker C, Borges R, Weiglun L, Slodkowitz G, Goldman N. 2020. Issues with SARS-CoV-2
19 sequencing data - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology - *Virological*.
20 <https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473>
- 21 Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, Watts AG, Vogels CBF, Brito AF, Alpert T,
22 Muyombwe A, Razeq J, Downing R, Cheemarla NR, Wyllie AL, Kalinich CC, Ott IM, Quick J,
23 Loman NJ, Neugebauer KM, Greninger AL, Jerome KR, Roychoudhury P, Xie H, Shrestha L,
24 Huang ML, Pitzer VE, Iwasaki A, Omer SB, Khan K, Bogoch II, Martinello RA, Foxman EF, Landry
25 ML, Neher RA, Ko AI, Grubaugh ND. 2020. Coast-to-Coast Spread of SARS-CoV-2 during the
26 Early Epidemic in the United States. *Cell* **181**:990–996. doi:10.1016/j.cell.2020.04.021
- 27 Goldberg SA, Lennerz J, Klompas M, Mark E, Pierce VM, Thompson RW, Pu CT, Ritterhouse LL, Dighe
28 A, Rosenberg ES, Grabowski DC. 2020. Presymptomatic Transmission of Severe Acute
29 Respiratory Syndrome Coronavirus 2 Among Residents and Staff at a Skilled Nursing Facility:
30 Results of Real-time Polymerase Chain Reaction and Serologic Testing. *Clin Infect Dis*.
31 doi:10.1093/cid/ciaa991
- 32 Graham NSN, Junghans C, Downes R, Sendall C, Lai H, McKirdy A, Elliott P, Howard R, Wingfield D,
33 Priestman M, Ciechonska M, Cameron L, Storch M, Crone MA, Freemont PS, Randell P,
34 McLaren R, Lang N, Ladhani S, Sanderson F, Sharp DJ. 2020. SARS-CoV-2 infection, clinical

- 1 features and outcome of COVID-19 in United Kingdom nursing homes. *J Infect.*
2 doi:10.1016/j.jinf.2020.05.073
- 3 Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, Tan AL, Paul LM,
4 Brackney DE, Grewal S, Gurfield N, Van Rompay KKA, Isern S, Michael SF, Coffey LL, Loman NJ,
5 Andersen KG. 2019. An amplicon-based sequencing framework for accurately measuring
6 intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* **20**:1–19. doi:10.1186/s13059-
7 018-1618-7
- 8 Hadfield J. 2020. artic-network/rampart: Read Assignment, Mapping, and Phylogenetic Analysis in
9 Real Time. <https://github.com/artic-network/rampart>
- 10 He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, Lau YC, Wong JY, Guan Y, Tan X, Mo X, Chen Y, Liao B,
11 Chen W, Hu F, Zhang Q, Zhong M, Wu Y, Zhao L, Zhang F, Cowling BJ, Li F, Leung GM. 2020.
12 Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* **26**:672–675.
13 doi:10.1038/s41591-020-0869-5
- 14 Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast
15 Bootstrap Approximation. *Molecular biology and evolution. Mol Biol Evol* **35**:518–522.
16 doi:10.5281/zenodo.854445
- 17 ICS. 2020. Assessing whether COVID-19 patients will benefit from critical care, and an objective
18 approach to capacity challenges. <https://www.rcplondon.ac.uk/file/20726/download>
- 19 Illingworth CJR, Hamilton WL, Jackson C, Popay A, Meredith L, Houldcroft CJ, Hosmillo M, Jahun A,
20 Routledge M, Warne B, Caller L, Caddy S, Yakovleva A, Hall G, Khokhar FA, Feltwell T, Pinckert
21 ML, Georgana I, Chaudhry Y, Curran M, Parmar S, Sparkes D, Rivett L, Jones NK, Sridhar S,
22 Forrest S, Dymond T, Grainger K, Workman C, Gkrania-Klotsas E, Brown NM, Weekes MP, Baker
23 S, Peacock SJ, Gouliouris T, Goodfellow I, De Angelis D, Török ME. 2020. A2B-COVID: A method
24 for evaluating potential SARS-CoV-2 transmission events. *medRxiv* 2020.10.26.20219642.
- 25 Jones NK, Rivett L, Sparkes D, Forrest S, Sridhar S, Young J, Pereira-Dias J, Cormie C, Gill H, Reynolds
26 N, Wantoch M, Routledge M, Warne B, Levy J, Jiménez WDC, Samad FNB, McNicholas C, Ferris
27 M, Gray J, Gill M, Curran MD, Fuller S, Chaudhry A, Shaw A, Bradley JR, Hannon GJ, Goodfellow
28 IG, Dougan G, Smith KGC, Lehner PJ, Wright G, Matheson NJ, Baker S, Weekes MP. 2020.
29 Effective control of sars-cov-2 transmission between healthcare workers during a period of
30 diminished community prevalence of covid-19. *Elife* **9**:1–10. doi:10.7554/eLife.59391
- 31 Jones NR, Qureshi ZU, Temple RJ, Larwood JPI, Greenhalgh T, Bourouiba L. 2020. Two metres or one:
32 what is the evidence for physical distancing in covid-19? *BMJ* **370** :m3223.
33 doi:<https://doi.org/10.1136/bmj.m3223>
- 34 Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermini LS. 2017. ModelFinder: Fast model

- 1 selection for accurate phylogenetic estimates. *Nat Methods* **14**:587–589.
- 2 doi:10.1038/nmeth.4285
- 3 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements
4 in performance and usability. *Mol Biol Evol* **30**:772–780. doi:10.1093/molbev/mst010
- 5 Kemenesi G, Kornya L, Tóth GE, Kurucz K, Zeghib S, Somogyi BA, Zöldi V, Urbán P, Herczeg R, Jakab
6 F. 2020. Nursing homes and the elderly regarding the COVID-19 pandemic: situation report
7 from Hungary. *GeroScience* **2**. doi:10.1007/s11357-020-00195-z
- 8 Ladhani SN, Chow JY, Janarthanan R, Fok J, Crawley-Boevey E, Vusirikala A, Fernandez E, Perez MS,
9 Tang S, Dun-Campbell K, Wynne-Evans E, Bell A, Patel B, Amin-Chowdhury Z, Aiano F,
10 Paranthaman K, Ma T, Saavedra-Campos M, Myers R, Ellis J, Lackenby A, Gopal R, Patel M,
11 Chand M, Brown K, Hopkins S, Consortium CG, Shetty N, Zambon M, Ramsay ME. 2020.
12 Increased risk of SARS-CoV-2 infection in staff working across different care homes: enhanced
13 CoVID-19 outbreak investigations in London care Homes. *J Infect* **81**:621–624.
14 doi:10.1016/j.jinf.2020.07.027
- 15 Lansbury LE, Brown CS, Nguyen-Van-Tam JS. 2017. Influenza in long-term care facilities. *Influenza
16 Other Respi Viruses* **11**:356–366. doi:10.1111/irv.12464
- 17 Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM **00**:1–3.
- 18 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The
19 Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079.
20 doi:10.1093/bioinformatics/btp352
- 21 MacFadden DR, McGeer A, Athey T, Perusini S, Olsha R, Li A, Eshaghi A, Gubbay JB, Hanage WP.
22 2018. Use of genome sequencing to define institutional influenza outbreaks, Toronto, Ontario,
23 Canada, 2014-15. *Emerg Infect Dis* **24**:492–497. doi:10.3201/eid2403.171499
- 24 McAuslane H, Morgan D. 2014. Communicable Disease Outbreak Management. Operational
25 guidance. *Public Heal Engl* 1–66.
- 26 Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, Curran MD, Parmar S,
27 Caller LG, Caddy SL, Khokhar FA, Yakovleva A, Hall G, Feltwell T, Forrest S, Sridhar S, Weekes
28 MP, Baker S, Brown N, Moore E, Popay A, Roddick I, Reacher M, Gouliouris T, Peacock SJ,
29 Dougan G, Török ME, Goodfellow I. 2020. Rapid implementation of SARS-CoV-2 sequencing to
30 investigate cases of health-care associated COVID-19: a prospective genomic surveillance
31 study. *Lancet Infect Dis* **0**. doi:10.1016/S1473-3099(20)30562-4
- 32 Middleton J, Reintjes R, Lopes H. 2020. Meat plants-a new front line in the covid-19 pandemic. *BMJ*
33 **370**:1–2. doi:10.1136/bmj.m2716
- 34 NICE. 2020. NICE guideline NG159. COVID-19 rapid guideline: critical care in adults. 2. Admission to

- 1 critical care.
- 2 Office for National Statistics. 2011. 2011 Census - Population and Household Estimates for England
3 and Wales, March 2011. *Natl Census* 37.
- 4 ONS. 2020a. Deaths involving COVID-19, England and Wales - Office for National Statistics.
5 [https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/b](https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsinvolvingcovid19englandandwales/deathsoccurringinjune2020)
6 [ulletins/deathsinvolvingcovid19englandandwales/deathsoccurringinjune2020](https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsinvolvingcovid19englandandwales/deathsoccurringinjune2020)
- 7 ONS. 2020b. Deaths involving COVID-19 in the care sector, England and Wales - Office for National
8 Statistics.
9 [https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/a](https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/deathsinvolvingcovid19inthecaresectorenglandandwales/deathsoccurringupto12june2020andregisteredupto20june2020provisional)
10 [rticles/deathsinvolvingcovid19inthecaresectorenglandandwales/deathsoccurringupto12june20](https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/deathsinvolvingcovid19inthecaresectorenglandandwales/deathsoccurringupto12june2020andregisteredupto20june2020provisional)
11 [20andregisteredupto20june2020provisional](https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/deathsinvolvingcovid19inthecaresectorenglandandwales/deathsoccurringupto12june2020andregisteredupto20june2020provisional)
- 12 ONS. 2019a. Families and households in the UK - Office for National Statistics.
13 [https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/](https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2019)
14 [bulletins/familiesandhouseholds/2019](https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2019)
- 15 ONS. 2019b. Families and households - Office for National Statistics.
16 [https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/](https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/datasets/familiesandhouseholds)
17 [datasets/familiesandhouseholds](https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/datasets/familiesandhouseholds)
- 18 ONS. 2014. Changes in the Older Resident Care Home Population between 2001 and 2011 - Office
19 for National Statistics.
20 [https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing/a](https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing/articles/changesintheolderresidentcarehomepopulationbetween2001and2011/2014-08-01)
21 [rticles/changesintheolderresidentcarehomepopulationbetween2001and2011/2014-08-01](https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing/articles/changesintheolderresidentcarehomepopulationbetween2001and2011/2014-08-01)
- 22 Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid
23 efficient extraction of SNPs from multi-FASTA alignments. *Microb genomics* 2:e000056.
24 doi:10.1099/mgen.0.000056
- 25 Public Health England. 2020a. Weekly Coronavirus Disease 2019 (COVID-19) Surveillance Report
26 Confirmed cases in England. Year: 2020, Week: 20. *Summ COVID-19 Surveill Syst*.
- 27 Public Health England. 2020b. COVID-19 Personal protective equipment (PPE) – resource for care
28 workers working in care homes during sustained COVID-19 transmission in England.
- 29 Quicke K, Gallichote E, Sexton N, Young M, Janich A, Gahm G, Carlton EJ, Ehrhart N, Ebel GD. 2020.
30 Longitudinal Surveillance for SARS-CoV-2 RNA Among Asymptomatic Staff in Five Colorado
31 Skilled Nursing Facilities: Epidemiologic, Virologic and Sequence Analysis. *medRxiv*
32 2020.06.08.20125989. doi:10.1101/2020.06.08.20125989
- 33 Rambaut A, Holmes E, O’Toole Á, Hill V, McCrone J, Ruis C, du Plessis L, Pybus O. 2020. SARS-CoV-2
34 lineages. <https://cov-lineages.org/descriptions.html>

- 1 Rambaut Andrew, Holmes EC, O’Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG. 2020. A
2 dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat*
3 *Microbiol.* doi:10.1038/s41564-020-0770-5
- 4 Rivett L, Sridhar S, Sparkes D, Routledge M, Jones NK, Forrest S, Young J, Pereira-Dias J, Hamilton WL,
5 Ferris M, Torok ME, Meredith L, Curran M, Fuller S, Chaudhry A, Shaw A, Samworth RJ, Bradley
6 JR, Dougan G, Smith KGC, Lehner PJ, Matheson NJ, Wright G, Goodfellow I, Baker S, Weekes
7 MP. 2020. Screening of healthcare workers for SARS-CoV-2 highlights the role of asymptomatic
8 carriage in COVID-19 transmission. *Elife* 9:1–20. doi:10.7554/eLife.58728
- 9 Seemann T. 2020. tseemann/snp-dists: Pairwise SNP distance matrix from a FASTA sequence
10 alignment. <https://github.com/tseemann/snp-dists>
- 11 Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. 2019. Beyond the SNP Threshold:
12 Identifying Outbreak Clusters Using Inferred Transmissions. *Mol Biol Evol* 36:587–603.
13 doi:10.1093/molbev/msy242
- 14 Strausbaugh LJ, Sukumar SR, Joseph CL. 2003. Infectious disease outbreaks in nursing homes: An
15 unappreciated hazard for frail elderly persons. *Clin Infect Dis* 36:870–876. doi:10.1086/368197
- 16 Tonkin-Hill G. 2020. gtonkinhill/fasttranscluster. <https://github.com/gtonkinhill/fasttranscluster>
- 17 UK government. 2020a. COVID-19: number of outbreaks in care homes - management information -
18 GOV.UK. [https://www.gov.uk/government/statistical-data-sets/covid-19-number-of-](https://www.gov.uk/government/statistical-data-sets/covid-19-number-of-outbreaks-in-care-homes-management-information)
19 [outbreaks-in-care-homes-management-information](https://www.gov.uk/government/statistical-data-sets/covid-19-number-of-outbreaks-in-care-homes-management-information)
- 20 UK government. 2020b. Update on policies for visiting arrangements in care homes.
21 [https://www.gov.uk/government/publications/visiting-care-homes-during-](https://www.gov.uk/government/publications/visiting-care-homes-during-coronavirus/update-on-policies-for-visiting-arrangements-in-care-homes#section-4)
22 [coronavirus/update-on-policies-for-visiting-arrangements-in-care-homes#section-4](https://www.gov.uk/government/publications/visiting-care-homes-during-coronavirus/update-on-policies-for-visiting-arrangements-in-care-homes#section-4)
- 23 UK government. 2020c. Vivaldi 1: COVID-19 care homes study report - GOV.UK.
24 [https://www.gov.uk/government/publications/vivaldi-1-coronavirus-covid-19-care-homes-](https://www.gov.uk/government/publications/vivaldi-1-coronavirus-covid-19-care-homes-study-report/vivaldi-1-covid-19-care-homes-study-report)
25 [study-report/vivaldi-1-covid-19-care-homes-study-report](https://www.gov.uk/government/publications/vivaldi-1-coronavirus-covid-19-care-homes-study-report/vivaldi-1-covid-19-care-homes-study-report)
- 26 UK government. 2020d. Get coronavirus tests for a care home - GOV.UK. [https://www.gov.uk/apply-](https://www.gov.uk/apply-coronavirus-test-care-home)
27 [coronavirus-test-care-home](https://www.gov.uk/apply-coronavirus-test-care-home)
- 28 Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, Curtis HJ, Mehrkar A, Evans D,
29 Inglesby P, Cockburn J, McDonald HI, MacKenna B, Tomlinson L, Douglas IJ, Rentsch CT, Mathur
30 R, Wong AYS, Grieve R, Harrison D, Forbes H, Schultze A, Croker R, Parry J, Hester F, Harper S,
31 Perera R, Evans SJW, Smeeth L, Goldacre B. 2020. OpenSAFELY: factors associated with COVID-
32 19 death in 17 million patients. *Nature*. doi:10.1038/s41586-020-2521-4
- 33 Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL,
34 Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. 2020. A new coronavirus

1 associated with human respiratory disease in China. *Nature* **579**:265–269. doi:10.1038/s41586-
2 020-2008-3

3 Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. Ggtree: an R Package for Visualization and
4 Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods*
5 *Ecol Evol* **8**:28–36. doi:10.1111/2041-210X.12628

6 Zhang J, Litvinova M, Wang W, Wang Y, Deng X, Chen Xinghui, Li M, Zheng W, Yi L, Chen Xinhua, Wu
7 Q, Liang Y, Wang X, Yang J, Sun K, Longini IM, Halloran ME, Wu P, Cowling BJ, Merler S, Viboud
8 C, Vespignani A, Ajelli M, Yu H. 2020. Evolving epidemiology and transmission dynamics of
9 coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study.
10 *Lancet Infect Dis* **20**:793–802. doi:10.1016/S1473-3099(20)30230-9

11

12

1 **Main display items – legends**

2 **Figure 1**

3 **Study flow diagram**

4 Out of 6,600 patients testing positive in the Cambridge Microbiology Public Health Laboratory
5 (CMPHL) during the study period, 1,167 were identified as being care home residents from 337 care
6 homes. (The methodology for assigning care home status is described in main text and Figure 1
7 supplement 1.) Out of 1,297 samples from 1,167 care home residents, 286 samples were assigned
8 for nanopore sequencing on site and 833 samples for sequencing at the Wellcome Sanger Institute
9 (WSI). Of these, 258 and 533 sequences were available and downloaded from the MRC-CLIMB server
10 at the time of running the analysis, respectively. Of these available genomes, 224 and 522 passed
11 sequencing quality control thresholds (described in Methods), respectively. This yielded the final
12 analysis set of 700 high-coverage genomes from care home residents (representing 292 care
13 homes): 197 genomes sequenced on site by nanopore and 503 sequences at WSI by Illumina. * 193
14 care homes were registered with the CQC as being residential homes without nursing care, referred
15 to as “residential homes” in main text, and 144 had nursing care available, referred to as “nursing
16 homes”. ** Samples were selected for nanopore sequencing on site if they were inpatients or
17 healthcare workers at Cambridge University Hospitals NHS Foundation Trust (CUH), where we
18 prioritised rapid turnaround time to investigate hospital-acquired infections, plus a randomised
19 selection of other East of England samples to provide broader genomic context to the CUH cases.
20 The remaining samples not selected for nanopore sequencing on site, where available, were sent to
21 WSI for sequencing.

22

23

24

1 **Figure 1, supplement 1**

2 **Flow diagram for identifying care homes from Cambridge-COGUK metadata**

3 Steps for identifying care home residents (further details in Methods). First, the address field in the
4 patient electronic healthcare records was searched for matching terms indicating a care home (e.g.
5 “care home”, “nursing home”, *etc*). Second, the patient address field was searched for matching
6 terms from a list of care home names registered to the Care Quality Commission (CQC). The resulting
7 list was manually inspected and every care home included in the study was linked to a registered
8 CQC care home. CQC coding of whether the care home had nursing care available was used (referred
9 to as “nursing homes” if nursing care was available and “residential homes” if not). If the address
10 information was incomplete (no postcode and/or no address line) then the case was excluded as
11 impossible to determine whether or not the patient was from a care home, unless the person was
12 known to be a healthcare worker (HCW), in which case it was assumed they were not a care home
13 resident. This process yielded the final result of 1,167 care home residents from 337 care homes;
14 5,246 individuals that were not care home residents, and 187 individuals that were indeterminable.

15

16

1 **Figure 1, supplement 2**

2 **Breakdown of main organisations submitting samples to Cambridge PHE Laboratory over study**
3 **period per week**

4 Only showing sites that submitted samples from >50 people with positive test results over study
5 period, otherwise counted as “Other”. To maintain patient anonymity, per time interval only
6 showing sites that submitted samples from >5 people with positive test results (otherwise counted
7 as “Other”). Data prior to 16 March is amalgamated due to low sample numbers. Note that over the
8 course of the study, some sites changed testing provider from CMPHL as further testing sites
9 became available around the region. This explains some of the variation in the relative proportion of
10 cases submitted from each site. The numbers reported here do not necessarily reflect total case
11 numbers for each hospital or submitting organisation, as tests may have been performed elsewhere
12 or metadata not collected in this study; the numbers are included purely to indicate where the
13 samples included in this study originated from.

14

15

1 **Figure 1, supplement 3**

2 **UK care home testing policy timeline**

- 3 1. 31st January - first recorded case of covid-19 in the UK
- 4 2. 26th February - first case of COVID-19 in the East of England; start date of this study
- 5 3. 12th March – individuals in the community advised to self-isolate for 7 days, without testing. Testing
- 6 only offered to care homes in the context of a suspected outbreak
- 7 4. 23rd March - UK lockdown officially begins
- 8 5. 15th April – action plan announced to test all symptomatic residents in care homes, plus testing of all
- 9 residents prior to admission to care home from hospital
- 10 6. 29th April – testing guidance amended to reflect that asymptomatic as well symptomatic residents and
- 11 staff in care homes may need to be tested as part of an outbreak
- 12 7. Policy for COVID-19 testing prior to discharge to care homes instigated 16th April:
- 13 [https://www.gov.uk/government/publications/coronavirus-covid-19-adult-social-care-action-](https://www.gov.uk/government/publications/coronavirus-covid-19-adult-social-care-action-plan/covid-19-our-action-plan-for-adult-social-care)
- 14 [plan/covid-19-our-action-plan-for-adult-social-care](https://www.gov.uk/government/publications/coronavirus-covid-19-adult-social-care-action-plan/covid-19-our-action-plan-for-adult-social-care)
- 15 8. 10th May - end date of this study
- 16 9. 11th May – national whole care home testing portal (offering a single test to all staff and residents)
- 17 goes live for care homes with residents aged 65 years and over or dementia patients
- 18 10. 8th June – national whole care home testing portal extends eligibility to care homes with residents
- 19 aged under 65 years
- 20 11. 3rd July – announcement that regular asymptomatic testing for care home staff and residents will be
- 21 rolled out through the national whole care home testing portal in July for homes with residents aged
- 22 over 65 years or dementia patients

23 References: (“Coronavirus testing - GOV.UK,” 2020, “COVID-19 policy tracker | The Health Foundation,” 2020)

24
25

1 **Figure 2**

2 **Care home locations by county, showing nursing and residential homes**

3 Only showing the five counties with the largest number of cases (all >25) to preserve patient
4 anonymity. Definitions of “nursing home” and “residential home” are based on Care Quality
5 Commission (CQC) information on whether nursing care is or is not present. If no nursing care is
6 available the home is classified as a residential home. If the care home offers nursing care (including
7 if it can offer both nursing and residential care) then the home is classified as a nursing home.

8

9

1 **Figure 2, supplement 1**

2 **Distribution of cases per care home**

3 The number of positive cases per care home was highly skewed, such that a relatively small number
4 of care homes contributed a large proportion of cases (right-hand side of the plot). Plot produced
5 with R package *ggplot2* using `geom_histogram` with `binwidth=1`.

6

1 **Figure 3**

2 **Epidemic curves for EoE and CUH showing care home residents**

3 Number of positive cases per week over the study period for different infection sources, for all
4 samples tested from EoE at the Cambridge PHE laboratory (A), or those tested at CUH acute medical
5 services (B). Peak of the epidemic for samples tested at the Cambridge PHE laboratory and CUH
6 acute medical services were weeks commencing 30th March and 6th April, respectively. UK
7 lockdown started 23rd March 2020. In both settings a prolonged right-hand “tail” was observed as
8 case numbers gradually fell. The relative proportion of cases admitted from care homes increased
9 over this period for both sample sets, while the contribution of general community cases fell more
10 quickly. However, interpreting these trends is confounded by the changing profile of COVID-19
11 testing nationally and regionally. If the patient address was missing, and they were not a HCW, then
12 the care home status was undetermined. CAI = Community Acquired Infection; EoE = East of
13 England; HAI = Hospital Acquired Infection; HCW = Healthcare Worker; “Other” mainly comprise
14 inpatient transfers from other hospitals to CUH for which metadata was lacking to determine the
15 infection category. CAI was considered “healthcare-associated” if there had been healthcare contact
16 within 14 days of first positive swab. The three categories of HAI were defined based on the
17 difference in days between admission and first positive swab, reflecting increasing likelihood of
18 hospital acquisition: indeterminate = 3-6 days; suspected 7-14 days; definite >14 days (as used in
19 (Meredith et al., 2020)).

20

21

1 **Figure 3, supplement 1**

2 **Care home residents per week showing genome sequencing site**

3 Plot shows total care home residents testing positive per week over the study period, showing
4 number of care home residents with genomes included in the study broken down by sequencing
5 location (on site in the Department of Pathology, Division of Virology or at the Wellcome Sanger
6 Institute).

7
8

1 **Figure 4**

2 **Odds ratios for mortality at 30 days**

3 Logistic regression analysis showing odds of death at 30 days (with 95% confidence intervals) for five
4 available metadata variables: patient sex, age (here categorised as ≥ 80 years), whether they were a
5 care home resident, the diagnostic Ct value (here categorised as < 20), and whether they were
6 admitted to the intensive care unit. Overall there were 116 deaths within 30 days of diagnosis (out
7 of 464 CUH patients). ICU = intensive care unit. Ct = Cycle threshold for diagnostic PCR.

8

9

10

1 **Figure 4, supplement 1**

2 **Pairwise comparisons of mortality at 30 days, age and whether the person was a care home**
3 **resident**

4 Each plot compares two of these three variables to visualise cross-associations, and the data are
5 divided in each case into individuals that died (yellow) or survived (blue). The plot was produced
6 using `GGally::ggpairs()`.

7

1 **Figure 5**

2 **Viral lineage compositions in care home and non-care home samples**

3 Plots showing the ratios of SARS-CoV-2 viral lineages for 700 care home resident genomes (A) and a
4 randomly selected subset of 700 non-care home residents (B). The proportion of lineage B.1.1
5 increased over the study period in both care home and non-care home residents. Lineages defined
6 using *pangolin*. Data also presented in Table 5.

7

8

1 **Figure 5, supplement 1**

2 **Viral lineage compositions in care home and non-care home samples by count**

3 Plots showing the counts of SARS-CoV-2 viral lineages for 700 care home resident genomes (A) and a
4 randomly selected subset of 700 non-care home residents (B). Lineages defined using *pangolin*. Data
5 also presented in Table 5.

6
7

1 **Figure 5, supplement 2**

2 **Distribution of pairwise SNP differences between care home samples**

3 Pairwise SNP differences between the 700 care home residents (244,650 comparisons). There was a
4 median of 8 single nucleotide polymorphisms (SNPs) separating care home genomes (interquartile
5 range, IQR 6 – 12 , range 0 – 29), compared to 9 (IQR 5 – 13, range 0 – 28) for randomly selected
6 non-care home samples ($P=0.95$, Wilcoxon rank sum test).

7

1 **Figure 6**

2 **Care home clustering on viral phylogenetic tree and within-care home pairwise SNP differences**

3 A. Phylogenetic tree of 1,400 East of England SARS-CoV-2 genomes rooted on a sample from Wuhan,
4 China, collected December 2019, including 700 care home residents and 700 randomly selected non-
5 care home residents. The colour bar (right) indicates whether samples were from care home
6 residents (blue) or non-care home residents (grey). Samples from the ten care homes with the
7 largest number of genomes are highlighted by coloured circles on branch tips. A magnified subtree
8 of the branch containing all 18 samples from care home CARE0314 is shown to the left. These
9 genomes were all either identical or differed by one SNP from the most common genome in this
10 cluster. Two non-care home genomes are also present in this group. Across the dataset, viruses from
11 care home residents and people not living in care homes are phylogenetically intermixed, consistent
12 with viral transmission between these two settings. B. Distributions of pairwise SNP differences for
13 the ten care homes with the largest number of genomes (same samples as highlighted in the branch
14 tips of panel A). Numbers above each box indicate the number of genomes present from that care
15 home. Among the ten care homes with the largest number of genomes, some clustered closely on
16 the phylogenetic tree with low pairwise SNP differences (e.g. CARE0063, CARE0264, CARE0314); in
17 contrast, some care homes were distributed across the tree with higher pairwise SNP differences
18 (e.g. CARE0061, CARE0151, CARE0173, CARE0263). Clusters within each care home were defined
19 using integrated genomic and temporal data using the *transcluster* algorithm and are shown in
20 Figure 7.

21

22

1 **Figure 6, supplement 1**

2 **Phylogenetic tree of all available genomes highlighting care home and non-care home samples**

3 Of the 6,600 individuals in the study, 1,167 were identified as care home residents and 5,246 were
4 not care home residents (187 were undetermined). 700 / 1,167 (60.0%) care home residents had
5 genomes available that passed quality control (QC) filtering at time of analysis. 3,745 / 5,246 (71.4%)
6 non-care home residents had genomes available and passing the same QC filtering at time of
7 analysis, accessed from the COG-UK public database (<https://www.cogconsortium.uk/data/>). This
8 tree comprises all 700 care home and 3,745 non-care home genomes from the study (total 4,445
9 samples), rooted on a 2019 genome from Wuhan, China. As with Figure 6, the colour bar (right)
10 indicates whether samples were from care home residents (blue) or non-care home residents (grey).
11 Samples from the ten care homes with the largest number of genomes are highlighted by coloured
12 circles on branch tips. This supports the findings shown in Figure 6 using the randomly selected sub-
13 sample of non-care home samples, (1) that care home genomes were phylogenetically intermixed
14 with non-care home genomes (consistent with transmission between care homes and outside of
15 care homes), and (2) that, using the 10 care homes with the largest number of samples as examples,
16 some care homes were monophyletic (such as CARE0314) while others were polyphyletic (such as
17 CARE0061). Even for polyphyletic care homes (implying multiple independent introductions of the
18 virus among residents), the majority of samples were usually attributable to a single dominant
19 cluster (described further in main text).

1 **Figure 7**

2 **Visualisations of SARS-CoV-2 clusters among care home residents**

3 Transmission networks were produced using a derivative of the *transcluster* algorithm, which
4 incorporates pairwise date and genetic differences to estimate the probability of cases being
5 connected within a defined number of intermediate hosts. Clusters were defined using a probability
6 threshold of $\geq 15\%$ for cases being linked by ≤ 2 intermediate hosts (further details in Methods). A.
7 Transmission clusters for the ten care homes with the largest number of care home residents with
8 available genomes. Consistent with Figure 6, several of the ten care homes with the largest number
9 of genomes comprised single transmission clusters (e.g. CARE0314), while others contained two or
10 more clusters consistent with multiple independent transmission sources among the residents.
11 These data alone do not indicate where the residents acquired their infections, and hospital-
12 acquired infections for some of the clusters is a possibility alongside multiple introductions into the
13 same care homes. B. Visualisation of transmission links between residents of two nearby carehomes
14 and a group of healthcare workers (HCW). Two care homes, CARE0063 (blue) and CARE0273
15 (orange), each had strong transmission links identified with the *transcluster* algorithm to a group of
16 HCW (green). The HCW comprised paramedics and care home carers – one working at CARE0063
17 and the other working at an unknown care home. We do not have confirmatory epidemiological
18 data available, but this raises the possibility of the cases sharing a linked transmission network.
19
20

1 **Figure 7, supplement 1**

2 **Transmission network diagrams for all care homes with 2 or more cases with genomic data**

3 Transmission networks were produced using a derivative of the *transcluster* algorithm, which
4 incorporates pairwise date and genetic differences to estimate the probability of cases being
5 connected within a defined number of intermediate hosts. Clusters were defined using a probability
6 threshold of $\geq 15\%$ for cases being linked by ≤ 2 intermediate hosts (further details in Methods). This
7 figure displays data from all care homes with ≥ 2 samples with genomic data.

1 **Figure 7, supplement 2**

2 **Histogram of pairwise transmission probabilities between care home samples**

3 Histogram of the pairwise probabilities for cases being connected by ≤ 2 intermediate hosts for all
4 700 care home residents as inferred by the *transcluster* algorithm, with vertical red line at 0.15
5 showing the cutoff used to identify care home clusters in our analysis. Note the data gaps along the
6 x-axis reflect the inherent discontinuity of the input datasets, measured in days and SNP differences
7 between cases.

8

1 **Figure 7, supplement 3**

2 **Transmission probability threshold vs number of care home clusters**

3 The *transcluster* algorithm computes the likelihood of two samples being linked within a given
4 number of intermediate hosts, based on the date and genetic differences between samples
5 (assuming a given serial interval and mutation rate, further details in Methods). Changing the
6 probability threshold used to define clusters changes the number of clusters defined, with a higher
7 threshold yielding more clusters (and higher likelihood of transmission within each cluster). The
8 dataset analysed contained 700 genomes from residents in 292 care homes, and we treated each
9 care home separately as microcosms of potential infection networks. Therefore, the highest
10 theoretical number of clusters is 700, if every genome were its own cluster; and the lowest possible
11 number of clusters is 292, if every person within each care home was part of the same cluster. The
12 cut-off used ($\geq 15\%$ probability of transmission with ≤ 2 intermediate hosts) is indicated by the red
13 vertical line. This is arbitrary, and was selected 1) because the distribution of pairwise SNP and date
14 differences within resulting clusters appeared reasonable (Figure 7, supplement 4 and 5) and
15 because of a “jump” in the number of clusters occurring at that point.

16

17

18

1 **Figure 7, supplement 4**

2 **Pairwise SNP difference distribution between samples within clusters**

3 Within each cluster, 673 / 775 (86.8%) of pairwise links that had a $\geq 15\%$ probability of transmission

4 with ≤ 2 intermediate hosts had 0 or 1 pairwise SNP differences (maximum 4).

5

1 **Figure 7, supplement 5**

2 **Pairwise date difference distribution between samples within clusters, aggregated across dataset**

3 Within each cluster, 756 / 775 (97.5%) of pairwise links that had a $\geq 15\%$ probability of transmission

4 with ≤ 2 intermediate hosts cases were sampled < 14 days apart (maximum 22 days).

5

6

1 **Figure 7, supplement 6**

2 **Distributions of date ranges (from first to last sampling dates) for care homes vs clusters**

3 Date ranges were calculated by subtracting the date of the first sample from the last sample for each
4 care home (left) or cluster (right). Care homes and clusters were only included in this analysis if there
5 were ≥ 2 samples with available genomic data in that care home or cluster. 170 / 292 (58%) care
6 homes had 2 or more cases with genomic data (578 individuals), compared with 133 / 409 (33%)
7 clusters (424 individuals). Using these datasets, there was a median of 9 days (IQR: 4 – 15, range: 0 –
8 50) from the first case to the last case within each care home, compared with 5 days (IQR: 1 – 11,
9 range: 0 – 22) from the first case to the last case within each cluster ($P = 9.2e-06$, Wilcoxon rank sum
10 test). As expected, the *transcluster* algorithm produces clusters with a narrower and smaller date
11 range between samples than for the care homes as a whole. Collection date was used for sample
12 dates; if collection date was missing then receive date in the laboratory was used instead.

1 **Figure 7, supplement 7**

2 **Pairwise date difference distribution between samples within each cluster**

3 Boxplots indicate the median and interquartile ranges for the number of days separating samples
4 found to be within the same transmission cluster by the *transcluster* algorithm. The boxplots are
5 overlaid with points representing the underlying transmission links. Larger points are used to
6 represent cases where many transmission links within a cluster are separated by the same number
7 of days.

1 **Table 1**

2

Variable	Care home residents (all)	Non-care home residents (all)	Care home residents with genomes
Number (%)	1167/6413 (18.2%)	5246/6413 (81.8%)	700/1167 (60%)
Female (%)	624/1167 (53.5%)	2338/5246 (44.6%)	363/700 (51.9%)
Male (%)	543/1167 (46.5%)	2908/5246 (55.4%)	337/700 (48.1%)
Age in years (median, IQR, range)	86 (IQR: 79-90, range: 30-100)	65 (IQR: 48-80, range: 0-100)	86 (IQR: 78-90, range: 42-99)
Diagnostic Ct value	26 (IQR: 22-29)	25 (IQR: 21-29)	24 (IQR: 20-27)
Tested at CUH (%)	72/464 (15.5%)	392/464 (84.5%)	54/72 (75%)
CUH patient admitted to ICU (%)	<5/72 (<7%)	84/392 (21.4%)	<5/54 (<9%)
CUH patient 30-day mortality (%)	34/72 (47.2%)	78/392 (19.9%)	23/54 (42.6%)
Number of care homes	337	-	292
Cases/ care home (median, IQR, range)	2 (IQR: 1-5, range: 1-22)	-	2 (IQR: 1-3, range: 1-18)
Care homes with ≥ 5 cases	85/337 (25.2%)	-	32/292 (11%)

3

4 **Epidemiological characteristics of care home and non-care home residents with COVID-19 included**
 5 **in the study**

6 The total sample set for this study comprised 6,600 individuals. Of these, care home residency status
 7 could be established for 6,413 (97.2%). 1,167/6,413 (18.2%) individuals were identified as being care
 8 home residents, of which 700/1,167 (60.0%) had genomic data available that passed quality control
 9 filtering and were used for identifying care home clusters using the *transcluster* algorithm (described
 10 in Methods and main text). The subset of individuals (464/6,600, 7.03%) that were tested at
 11 Cambridge University Hospitals (CUH) had richer metadata available and were used for analysing
 12 intensive care unit (ICU) admissions and 30-day mortality after first positive test, shown here. Not
 13 showing precise values where the number of cases is equal to or less than five individuals, to
 14 preserve patient anonymity. Ct = Cycle threshold; CUH = Cambridge University Hospitals; ICU =
 15 Intensive Care Unit; IQR = interquartile range.

16

1 **Table 2**

2

Week commencing	Care home resident	Not determined	Not care home resident	Weekly total	Care home resident (%)
24-Feb	0	0	≤5	≤5	0.0%
02-Mar	0	0	31	31	0.0%
09-Mar	10	6	149	165	6.1%
16-Mar	25	6	364	395	6.3%
23-Mar	60	26	852	938	6.4%
30-Mar	126	35	1235	1396	9.0%
06-Apr	162	43	1064	1269	12.8%
13-Apr	154	31	540	725	21.2%
20-Apr	247	16	415	678	36.4%
27-Apr	198	16	393	607	32.6%
04-May	185	8	199	392	47.2%

3

4 **Case numbers from care homes and non-care home residents per week for full dataset tested at**
5 **Cambridge CMPHL**

6 Data plotted in Figure 3A of main text, showing case numbers for care homes, non-care homes, and
7 undetermined, for all EoE samples tested at CMPHL. The proportion of COVID-19 cases from care
8 home residents increased in April and May; however, this may reflect the changing profile of
9 samples submitted to the Cambridge CMPHL rather than underlying epidemiological trends.

10

11

1 **Table 3**

Week	Total weekly COVID-19 cases	Community acquired, care home-associated (%)
09-Mar	12	0 (0%)
16-Mar	24	0 (0%)
23-Mar	75	≤5 (≤7%)
30-Mar	96	≤5 (≤5.2%)
06-Apr	99	14 (14.1%)
13-Apr	49	14 (28.6%)
20-Apr	41	10 (24.4%)
27-Apr	41	9 (22.0%)
04-May	27	6 (22.2%)

2

3 **Proportion of community acquired, care home-associated COVID-19 infections tested at**
4 **Cambridge University Hospitals**

5 The proportion of community onset, care home-associated COVID-19 infections tested at Cambridge
6 University Hospitals (CUH) peaked in mid to late April. Total cases shows the total number of new
7 COVID-19 cases diagnosed at CUH that week. “Community acquired” was defined as first positive
8 test ≤48 hours from admission and no healthcare contact within the previous 14 days. Not showing
9 precise values if number of patients is less than or equal to 5 to preserve patient anonymity.

10

11

1 **Table 4**

2

Variable	OR	95% CI low	95% CI high	P value
Age \geq 80	6.6	3.7	12.0	2.46E-10
Sex	1.5	0.9	2.6	1.30E-01
Care resident status	3.0	1.6	5.7	9.22E-04
ICU admission	3.9	2.1	7.5	3.02E-05
Ct value $<$ 20	2.9	1.6	5.3	5.04E-04

3

4 **Odds ratios for mortality at 30 days**

5 Logistic regression analysis of odds of mortality at 30 days. Age \geq 80 years, being a care home
6 resident, being admitted to ICU and Ct $<$ 20 were significantly associated with increased odds of death
7 at 30 days post-diagnosis ($P<$ 0.05). OR = Odds Ratios. CI = Confidence Interval. ICU = intensive care
8 unit. Ct = Cycle threshold for diagnostic PCR.

9

10

1 **Table 5**

2

Care home status	Early	Late	% change
Care home resident	6/47 (12.8%)	155/286 (54.2%)	+ 41.40%
Not care home resident	39/173 (22.5%)	50/96 (52.1%)	+ 29.50%

3

4 **Proportion of care home and non-care home samples that were lineage B.1.1**

5 The proportion of lineage B.1.1 (defined using the Pangolin tool) increased from earlier to later
6 sampling weeks, for both care home and non-care home samples. Data based on the 700 care home
7 residents with genomic data available and 700 randomly selected non-care home samples. “Early”
8 was defined as the period from the start of the study (26th February 2020) to 29th March 2020. “Late”
9 was defined as 20th April 2020 to the end of the study (10th May 2020).

1 **Table 6**

Care home code	Sample count	Age (median, IQR, range)	Ct values (median, IQR, range)	Cluster count	Major cluster count	Care home date range (days)	Cluster date range (days, sample count)
CARE0032	7	87 (IQR: 81-91, range: 56-93)	23 (IQR: 22-24, range: 14-26)	2	6/7 (85.7%)	39	0 days, n=1 10 days, n=6
CARE0061	10	88.5 (IQR: 87-92.2, range: 84-97)	23 (IQR: 21.2-26.5, range: 12-33)	4	7/10 (70%)	38	0 days, n=1 22 days, n=7 0 days, n=1 0 days, n=1
CARE0063	12	74.5 (IQR: 67.8-81, range: 42-94)	23 (IQR: 20.8-27, range: 14-30)	2	11/12 (91.7%)	21	18 days, n=11 0 days, n=1
CARE0097	7	90 (IQR: 82.5-92, range: 73-95)	23 (IQR: 20.5-24, range: 17-27)	2	6/7 (85.7%)	28	0 days, n=1 14 days, n=6
CARE0151	7	81 (IQR: 77-89, range: 69-96)	20 (IQR: 19-25.5, range: 17-30)	4	4/7 (57.1%)	20	0 days, n=1 0 days, n=4 0 days, n=1 0 days, n=1
CARE0173	7	81 (IQR: 77.5-94, range: 71-95)	19 (IQR: 17.5-26, range: 15-27)	3	3/7 (42.9%)	21	0 days, n=1 3 days, n=3 0 days, n=3
CARE0263	12	85.5 (IQR: 81.8-90.5, range: 69-97)	19.5 (IQR: 18.5-24.8, range: 14-29)	3	9/12 (75%)	3	3 days, n=9 0 days, n=2 0 days, n=1
CARE0264	9	91 (IQR: 82-95, range: 73-96)	26 (IQR: 25-27, range: 18-29)	1	9/9 (100%)	14	14 days, n=9
CARE0277	13	84 (IQR: 82-89, range: 71-94)	26 (IQR: 24-27, range: 23-29)	2	12/13 (92.3%)	13	13 days, n=12 0 days, n=1
CARE0314	18	87.5 (IQR: 81.2-90.8, range: 74-97)	24 (IQR: 22.2-26, range: 14-29)	1	18/18 (100%)	5	5 days, n=18

2

3 **Outbreak characteristics for 10 care homes with the largest number of SARS-CoV-2 genomes**

4 Epidemiological characteristics of the 10 care homes with the largest number of genomes are
5 shown. Collectively these comprised 102 cases (102/700 (14%) of the total number of care home
6 cases with genomic data available). “Cluster count” refers to the number of SARS-CoV-2 clusters
7 within each care home defined by *transcluster* (described in Methods and main text). “Major cluster
8 count” shows the count for the dominant cluster (with the largest number of cases) and its
9 percentage contribution to total case numbers for each care home. “Care home date range”
10 indicates the number of days from first sample to last sample date for residents from each care
11 home. “Cluster date range” indicates the number of days from first sample to last sample date for
12 residents from each cluster within that care home, as defined by the *transcluster* algorithm, also
13 showing the sample count (n) for each cluster. Sampling dates used collection date if known, or

1 receive date in the diagnostic laboratory if collection date was unknown. The date range for each
2 care home is typically larger than the date range for clusters within care homes, except for single-
3 cluster care homes like CARE0314. This is consistent with the *transcluster* algorithm defining groups
4 of cases occurring closer together in time. While the care homes frequently had more than one
5 introduction of the virus among residents (i.e. >1 clusters), there was usually a single dominant
6 cluster responsible for the majority of cases. Individual counts of males and females for each care
7 home are not shown as this generally gave counts of less than five, risking patient anonymity.
8 Overall there were 59/102 (57.8%) females for these 10 care homes.

9
10

1 **Table 7**

Category	Counts (%)
Care home residents with genomic data	700
Care home residents with genomic data that could be linked to hospitalisation data	694/700 (99.1%)
Hospitalised during study period	470/694 (67.7%)
Hospitalised due to COVID-19	398/694 (57.3%)
Suspected hospital-acquired COVID-19	40/694 (5.76%)
Discharged within 7 days of positive test	230/694 (33.1%)

2

3 **Hospitalisation data for the 700 care home residents with genomic data available**

4 700/1,167 (60.0%) care home residents identified in the study had genomic data available and were
5 used to define care home SARS-CoV-2 clusters. We investigated the proportions of these care home
6 residents that were hospitalised and may have acquired their infections through interactions with
7 hospitals. This was possible for 694/700 (99.1%) individuals who had NHS numbers documented that
8 could be linked with national hospitalisation data. Being hospitalised due to COVID-19 was defined
9 as the date of first positive sampling being within 2 days prior to admission up to 7 days post-
10 admission. Suspected hospital-acquired COVID-19 infections were defined as first positive test being
11 7 days or more after hospital admission date and prior to discharge date (N=13) or within 7 days
12 following hospital discharge (N=27). Of the latter group, 10 individuals were admitted to hospital
13 and discharged on the same day prior to their positive test, 9 were admitted for 1 to 7 days, and 8
14 had been admitted for greater than 7 days.

15

- 1 **Supplementary File 1.** Supplementary materials for “Genomic epidemiology of
- 2 COVID-19 in care homes in the East of England”
- 3