medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

William L. Hamilton*[1,2,+], Gerry Tonkin-Hill*[3], Emily Smith[4], Dinesh Aggarwal[1,5], Charlotte J. Houldcroft[6], Ben Warne[1,2], Colin S. Brown[5], Luke W. Meredith[6], Myra Hosmillo[6], Aminu S. Jahun[6], Martin D. Curran[7], Surendra Parmar[7], Laura G. Caller[6,8], Sarah L. Caddy[6], Fahad A. Khokhar[1], Anna Yakovleva[6], Grant Hall[6], Theresa Feltwell[6], Malte L. Pinckert[6], Iliana Georgana[6], Yasmin Chaudhry[6], Nicholas M. Brown[7,2], Sónia Gonçalves[3], Roberto Amato[3], Ewan M. Harrison[3], Mathew A. Beale[3], Michael Spencer Chapman[3,9], David K. Jackson[3], Ian Johnston[3], Alex Alderton[3], John Sillitoe[3], Cordelia Langford[3], Gordon Dougan[1], Sharon J. Peacock[1], Dominic P. Kwiatkowski[3], Ian Goodfellow[6], M. Estée Török[1,2,+], COVID-19 Genomics Consortium UK[10]

* contributed equally

+ Corresponding authors

Affiliations

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

Correspondence

William Hamilton will.l.hamilton@gmail.com

M. Estée Török et317@cam.ac.uk

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Word count

Abstract = 150 (max 150)

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

Key words

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

Impact statement

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Abstract

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Introduction

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

Results

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

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

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

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

Identifying viral clusters within care homes using genomic and epidemiological data

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

This clustering method identified 409 transmission clusters from 292 care homes (median 1 cluster per care home, range 1-4). Within each cluster, 673 / 775 (86.8%) of pairwise links had zero or 1 pairwise SNP differences (maximum 4), and 756 / 775 (97.5%) were sampled <14 days apart (maximum 22 days) (Figure 7 Supplements 4-5). Clusters had a tighter distribution of sampling dates than for the total cases within each care home, as expected. For the 170 care homes with 2 or more cases with genomic data, there was a median of 9 (IQR: 4 – 15) days from the first case to the last case within each care home, compared with a median of zero (IQR 0-5) days from the first case to the last case of each cluster ($P < 10^{-5}$, Wilcoxon rank sum test). The median and interquartile range for pairwise date differences between all samples within each cluster is shown in Figure 7, Supplement 6. Transmission networks for the ten care homes with the largest number of genomes are shown in Figure 7A, indicating linked transmission clusters among residents based on the model assumptions and probability threshold (full dataset shown in Figure 7, Supplement 1). Consistent with the phylogeny shown in Figure 6A, some care homes contained a single transmission cluster involving multiple cases (e.g. CARE0314), while others comprised multiple independent clusters (e.g. CARE0061) (Table 6). While care homes frequently had more than one introduction of the virus among residents (i.e. >1 cluster), there was typically a single dominant cluster responsible for the majority of cases within each care home. Of the 170 care homes with 2 or more residents with genomic data (comprising 578 / 700 (82.6%) care home residents with genomic data), 111 / 170 (65.3%) had a dominant cluster responsible for >50% of all cases in the care home. This rises to 74 / 90 (82.2%) of care homes with three or more residents with genomic data.

The contribution made by genomic data in defining care home clusters was quantified. Without genomic data (or access to more detailed epidemiology such as accommodation sub-structuring within care homes), clustering can only be based on temporal differences between cases. For example, if two groups of COVID-19 cases occur several months apart within a care home they could be inferred to have resulted from (at least) two separate introductions. However, this method cannot account for multiple introductions occurring around the same time, as may happen when community transmission is high. To quantify the impact made by adding genomic data, which can distinguish between genetically dissimilar viruses introduced at similar times, the *transcluster* algorithm was repeated using the same parameters as for the main analysis but assuming all genomes were identical. This yielded 316 clusters – 23% fewer than the 409 clusters yielded when incorporating genomics. This suggests that genomics makes a significant contribution to defining viral clusters; without genomic data, cluster sizes may be over-estimated and the number of separate viral introductions under-estimated. This is illustrated by care home CARE0263, in which all

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

12 residents tested positive within 3 days of each-other, but these are divided into three separate clusters by the *transcluster* algorithm (one dominant cluster of 9 cases, one cluster of 2 cases and a single separate case (Table 6)); this is consistent with the phylogeny shown in Figure 6A, with samples split into three branches along the tree. Without genomic data, the three clusters in CARE0263 would have been impossible to distinguish.

Links between care homes and hospitals

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

Viral clusters linking care home residents and healthcare workers

Potential transmission networks involving care home residents and healthcare workers (HCW) were investigated for people tested at CUH (HCW data were not available outside of CUH). This analysis comprised 54 care home residents tested at CUH and 76 HCW with genomic data available. Clusters were defined using the same method as for the care home resident analysis (described above), but allowing HCW to belong to clusters from multiple care homes, so residents from several care homes could be linked to the same HCW. 38 / 54 (70.4%) care home residents had possible links with HCW using this relaxed threshold. However, on review of the medical records we could only identify strong epidemiological links for 14 / 54 (26.0%) residents from two care home clusters, CARE0063 and CARE0114. The CARE0063 cluster has been described previously (Meredith et al., 2020) and includes care home residents, a carer from that same care home and another from an unknown care home, paramedics and people living with the above. The CARE0114 cluster comprises several care home residents and acute medical staff working at CUH who cared for at least one of the residents. The *transcluster* method does not assign probabilities for directionality of transmission and cannot determine precise person-to-person transmission chains. While all residents from a care home

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

cluster may link to a given HCW, in reality the resident-HCW transmission event may have only involved one of the residents from that cluster, so the proportion of residents with links to HCW may be inflated. Nonetheless, these data show that two care home clusters involved HCW, one based mainly in the community and the other with hospital-based staff at CUH.

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

Discussion

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

A smaller proportion of care home residents were admitted to ICU compared with people who were not from care homes. What treatments a patient receives, including the invasive treatments provided in intensive care, are complex and individualised decisions based on risk-benefit assessments involving patients, their families and carers, and healthcare professionals (ICS, 2020; NICE, 2020). Of note, non-invasive respiratory support (such as continuous positive airway pressure, high-flow nasal oxygen therapy and non-invasive ventilation) are routinely provided outside ICU in many UK centres. Despite care home residents being at higher risk of severe COVID-19, and being under-represented in ICU, admission to ICU was still correlated with significantly increased mortality. This is likely because patients admitted to ICU have more severe disease, typically requiring more intensive treatments such as organ support.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

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

The majority of residents had hospital contact during the study period, indicating substantial opportunity for infections to pass between care homes and hospitals in either direction. A third of patients were discharged from hospital within 7 days of their first positive test, and thus were

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

potentially infectious at discharge. We identified transmission clusters that would be consistent with COVID-19 spread between care home residents and HCW, based both in the community and in hospitals. A previous study found that working across different homes was associated with higher SARS-CoV-2 positivity among staff (Ladhani et al., 2020). Limiting the spread of COVID-19 between care home residents, HCW and hospitals is a therefore another key target for infection control and prevention.

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

Second, the nature of diagnostic testing sites changed during the study period as regional hospitals developed their own in-house testing capacity and community testing laboratories were set up. “Pillar 2” testing in the UK was outsourced to high-throughput laboratories during April 2020 and performed an increasing proportion of community testing. It is possible that some care home residents from the same care home could have been tested through different routes, with symptomatic cases more likely to be tested in “Pillar 1” via the CMPHL (and included in this dataset), and asymptomatic screening occurring more via the Pillar 2 laboratories. However, most care homes in EoE only began systematic screening after the end of our study following the introduction of the UK care home testing portal on 11th May 2020. Moreover, the *transcluster* algorithm allows for “missing links” within a cluster (the threshold used assumed a $\geq$15% probability of infections being linked within $\leq$2 intermediate hosts), reducing the impact of missing care home cases on defined clusters. The changing profile of COVID-19 testing in the UK between March and May 2020 should therefore be factored into all interpretations of COVID-19 epidemiology from that period.

Third, defining who is a care home resident from large electronic healthcare records is challenging and, despite substantial efforts (described in Methods), some care home residents may have been

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

missed. Using pre-defined coding such as care home CQC registration numbers when patients are booked into hospital systems, rather than free-text data entry, would help considerably with care home surveillance. Multiple rounds of electronic searches and manual inspection were undertaken to identify as many care home residents as possible, and every care home resident included was cross-referenced against a CQC database of registered care homes in England. The care homes included for analysis should therefore be accurate.

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

In conclusion, care homes represent a major burden of COVID-19 morbidity and mortality, with transmission events introducing SARS-CoV-2 into care homes and subsequent transmission within them. Genomic data can be used in outbreak investigations to define viral clusters; this is critically dependent on integration with epidemiological data. The cut-offs we used for defining care home clusters were pragmatic but plausible given current understanding of the biology and epidemiology of SARS-CoV-2. Such cut-offs can be helpful for producing understandable outputs for biological and public health interpretation (MacFadden et al., 2018; Stimson et al., 2019), and for focusing investigations with limited public health resources. Future work will need to prospectively integrate genomic and epidemiological data to rapidly identify viral clusters, thus enabling deployment of infection control and public health interventions in real time.

Methods

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Study overview

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

Diagnostic testing, metadata collection and genome sequencing

For details on diagnostic testing, patient metadata collection and nanopore genome sequencing see (Meredith et al., 2020). Briefly, CMPHL used an in-house generated and validated one-step RT q-PCR assay detecting a 222-bp region of the RdRp genes, along with an MS2 bacteriophage internal extraction control, using the Rotorgene™ PCR instrument. Samples that generated a Ct value $\leq$36 were considered positive. The study aimed to sequence all samples which tested SARS-CoV-2 PCR positive at the CMPHL during the study period. Sequencing of every positive diagnostic sample could not be performed, however, for the following reasons: (i) sample unavailability (e.g. diagnostic samples being lost or discarded before they could be collected by the sequencing team); (ii) labelling errors when assigning sequencing codes (which resulted in specimens being discarded); or (iii) metadata mismatches (if the sample did not match to a metadata record downloaded from the hospital electronic patient records system). Samples were either sequenced on site using Oxford Nanopore Technologies or transported to the Wellcome Sanger Institute for Illumina sequencing.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

Sequencing at WSI used Illumina technology. cDNA was generated from SARS-CoV-2 viral nucleic acid extracts and subsequently amplified to produce 400nt amplicons tiling the viral genome using V3 nCov-2019 primers (ARTIC). This was followed by Illumina library generation using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs Inc, Cat. No. E7645L). Libraries were amplified with KAPA HiFi Ready Mix (Kapa Biosystems, Cat. No. 07958927001) and uniquely indexed with a 100 µM i5 and i7 primer mix (50 µM each) (Integrated DNA Technologies) to allow multiplexing of up to 384 SARS-CoV-2 viral extracts into one sequencing pool. The PCR products were pooled in equal volume and purified with an AMPure XP workflow (Beckman Coulter, Cat. No. A63880). The purified pool was quantified by qPCR (Illumina Library Quantitation Complete kit, Cat. No. KK4824) and sequenced on one lane of an Illumina NovaSeq SP flow cell (Illumina Inc, NovaSeq 6000 SP Reagent Kit v1.5 (500 cycles), Cat. No. 20028402), with XP workflow (Illumina Inc, NovaSeq XP 2 lane kit v1.5, Cat. No. 20043130). Genomes were generated for each library's sequencing data using bwa mem (Li, 2013) for alignment with MN908947.3 (Wu et al., 2020) as reference, samtools (Li et al., 2009) for pileup and ivar (Grubaugh et al., 2019) for trimming and consensus generation, all orchestrated by the ncov2019-artic-nf pipeline (Bull, 2020).

The WSI sequencing workflow also uses negative controls and the pass rate to date related to negative controls is 90%. Sequencing read counts are considered after a clipping and minimum

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

```
snp-sites -v alignment_file.aln
```

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

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

Sample selection

As described in (Meredith et al., 2020), patient metadata were downloaded daily from the electronic medical record system (Epic Systems, Verona, WI, USA) and metadata manipulations were performed in R (v 3.6.2) using the *tidyverse* packages (v 1.3.0) installed on CUH computers. Positive

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

samples were collected and assigned either for nanopore sequencing on site (focusing on CUH samples and a randomised selection of EoE samples), or sent to WSI for Illumina sequencing. Metadata were uploaded weekly to the MRC CLIMB system as part of the COG-UK Consortium. Samples included healthcare workers (HCW) tested in the CUH HCW screening programme (N. K. Jones et al., 2020; Rivett et al., 2020), all of which were nanopore sequenced on site.

Identifying care home residents

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

Step 1: Search terms in patient address fields

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

- "residential home"
- "care home"
- "nursing home"
- "care centre"
- "care hom"
- "nursing hom"
- "residential hom"
- "carehome"

This identified 765 patients as being care home residents.

Step 2: Matching location names to CQC registered care facilities

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

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

*Filtering using the "carehome?" column was based on advice given after correspondence with the CQC.

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

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

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

Step 3: Manual inspection and data clean up

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

Next, address lines for the care home residents were manually inspected and 14 were deemed not to be care home residents. Most of these were due to unrelated locations sharing the same address name as a CQC registered care home. The manual filtering steps thus yielded a care home resident

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

count of 1,147 + 89 – 14 = 1,222. Address line 1 for all 1,222 care home residents was manually inspected and formatted to ensure residents from the same care home had matching terms in this column. This was necessary due to discrepant address entrance formats for identical care homes; without this step, residents from the same care home would be incorrectly assigned to different anonymised care home codes.

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

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

Care home location IDs assigned by the CQC were turned into anonymised codes (format: CARE followed by a 4-digit numeric code). Care homes were classified as "residential homes" or "nursing homes" using the CQC data column "Service type - Care home service with nursing" filtered to "Y" for care homes with nursing, and column "Service type - Care home service without nursing" = "Y" for care homes without nursing ("residential homes"). If both fields were "Y" then the care home was coded as being a nursing home.

Linking care home data to CUH acute medical testing data

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

Statistics

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

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

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

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

To produce the plot of odds ratios shown in Figure 4, the age and diagnostic Ct value continuous variables were transformed into binary categoricals using cut-offs of age $\geq$80 years and Ct value <20. *P*-values below 10^{-5} are not reported.

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

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

- Patient address available

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

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

Care home viral phylogenetics and cluster analysis

Consensus fasta sequences were downloaded from the MRC-CLIMB website (https://www.climb.ac.uk/) (Connor et al., 2016). Genomes were de-duplicated (1 genome per person) and passed through quality control (QC) filtering using the same criteria as in (Meredith et al., 2020): genome size >29Kb, N count <2990 (i.e. >90% coverage). Where there were multiple sequences from the same patient, the sequence passing QC filters that was collected first was used for genomic analysis (closest to the onset of symptoms).

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

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

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

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

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

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

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

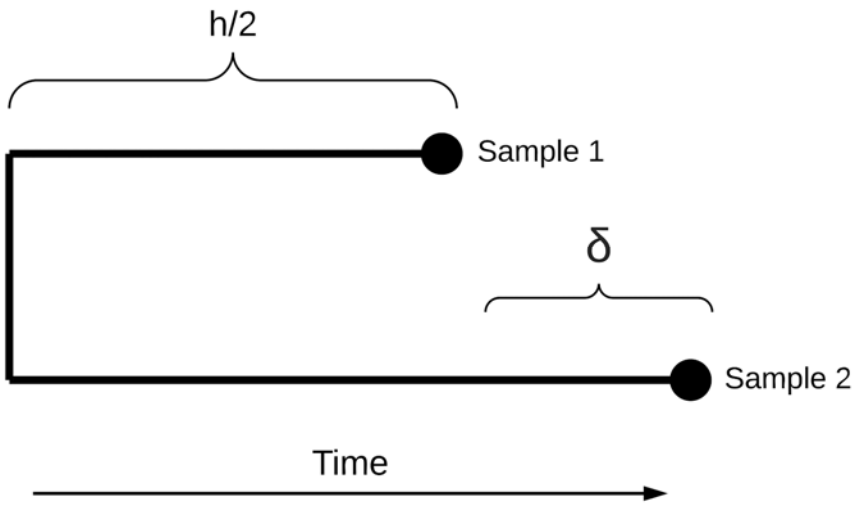
Lineage assignment

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

Clustering

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .



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

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

This can be expressed as the sum:

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

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

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

Investigating hospital admissions for care home residents

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

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

Hospital admission data were parsed to yield the following outputs:

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

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

CUH HCW-Care home resident cluster analysis

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

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

The 130 genomes were aligned using MAFFT and underwent the same cluster analysis using the *transcluster* algorithm as described above. Transmission links between care homes were excluded as were links between HCWs. HCWs could belong to multiple clusters from different care homes to allow for the possibility of a HCW seeding multiple care home infections. 21 clusters involving both care home residents and HCWs were identified. Of the 54 care home residents, 38 had links with HCWs within the 0.15 probability threshold. Medical notes for potential care home resident-HCW transmission pairs were reviewed by author WLH as described in (Meredith et al., 2020), with cases

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

being categorised as strongly linked epidemiologically (e.g. the HCW documented in the care home residents' medical notes); possibly linked (e.g. both working in the hospital at the same time but not in the same wards); or no evidence of an epidemiological link.

Ethics

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

Funding

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

Acknowledgements

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

References

Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-stroh S. 2020. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region , January to June 2020 1–8.

Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MTG, Yeats CA, Grundmann H, Spratt BG, Aanensen DM. 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb genomics* **2**:e000093. doi:10.1099/mgen.0.000093

Arons MM, Hatfield KM, Reddy SC, Kimball A, James A, Jacobs JR, Taylor J, Spicer K, Bardossy AC, Oakley LP, Tanwar S, Dyal JW, Harney J, Chisty Z, Bell M, Methner M, Paul P, Carlson CM,

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

McLaughlin HP, Thornburg N, Tong S, Tamin A, Tao Y, Uehara A, Harcourt J, Clark S, Brostrom-Smith C, Page LC, Kay M, Lewis J, Montgomery P, Stone ND, Clark TA, Honein MA, Duchin JS, Jernigan JA. 2020. Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. *N Engl J Med* **382**:2081–2090. doi:10.1056/NEJMoa2008457

Bull M. 2020. A Nextflow pipeline for running the ARTIC network's fieldbioinformatics tools with a focus on ncov2019. https://github.com/connor-lab/ncov2019-artic-nf

Bull RA, Adikari T, Hammond JM, Stevanovski I, Ferguson JM, Beukers AG, Naing Z, Yeang M, Verich A, Gamaarachichi H, Kim KW, Luciani F, Stelzer-Braid S, Eden J-S, Rawlinson WD, Van Hal SJ. 2020. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *bioRxiv* 2020.08.04.236893.

Burton JK, Bayne G, Evans C, Garbe F, Gorman D, Honhold N, McCormick D, Othieno R, Stevenson J, Swietlik S, Templeton K, Tranter M, Willocks L, Guthrie B. 2020. Evolution and impact of COVID-19 outbreaks in care homes: population analysis in 189 care homes in one geographic region. *medRxiv* 2020.07.09.20149583. doi:10.1101/2020.07.09.20149583

Byrne AW, McEvoy D, Collins AB, Hunt K, Casey M, Barber A, Butler F, Griffin J, Lane EA, McAloon C, O'Brien K, Wall P, Walsh KA, More SJ. 2020. Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. *BMJ Open* **10**:e039856. doi:10.1136/bmjopen-2020-039856

City Population. 2020. East of England (United Kingdom): Counties and Unitary Districts & Settlements - Population Statistics, Charts and Map. https://www.citypopulation.de/en/uk/eastofengland/

COG-UK. 2020. Pangolin COVID-19 Lineage Assigner. https://pangolin.cog-uk.io/

Collier D, Assennato S, Sithole N, Sharrocks K, Ritchie A, Ravji P, Routledge M, Sparkes D, Skittrall J, Warne B, smielewska A, RAMSEY I, Goel N, CURRAN M, ENOCH D, TASSELL R, LINEHAM M, VAGHELA D, LEONG C, Gupta R. 2020. Rapid point of care nucleic acid testing for SARS-CoV-2 in hospitalised patients: a clinical trial and implementation study. *Cell Reports Med* 100062. doi:10.1101/2020.05.31.20114520

Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, Bull MJ, Richardson E, Ismail M, Thompson SE, Kitchen C, Guest M, Bakke M, Sheppard SK, Pallen MJ. 2016. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microb genomics* **2**:e000086. doi:10.1099/mgen.0.000086

Coronavirus testing - GOV.UK. 2020. https://www.gov.uk/government/news/coronavirus-testing

COVID-19 policy tracker | The Health Foundation. 2020. https://www.health.org.uk/news-and-comment/charts-and-infographics/covid-19-policy-tracker

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

CQC. 2020a. Service types | Care Quality Commission. https://www.cqc.org.uk/guidance-providers/regulations-enforcement/service-types#care-homes-without-nursing

CQC. 2020b. What is registration? | Care Quality Commission. https://www.cqc.org.uk/guidance-providers/registration/what-registration

CQC. 2020c. Using CQC data | Care Quality Commission. https://www.cqc.org.uk/about-us/transparency/using-cqc-data

CQC. 2015. The scope of registration.

Curran ET. 2017. Infection outbreaks in care homes: prevention and management, Nursing Times. https://www.nursingtimes.net/clinical-archive/infection-control/infection-outbreaks-in-care-homes-prevention-and-management-14-08-2017/

De Maio N, Walker C, Borges R, Weilguny L, Slodkowicz G, Goldman N. 2020. Issues with SARS-CoV-2 sequencing data - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology - Virological. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473

Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, Watts AG, Vogels CBF, Brito AF, Alpert T, Muyombwe A, Razeq J, Downing R, Cheemarla NR, Wyllie AL, Kalinich CC, Ott IM, Quick J, Loman NJ, Neugebauer KM, Greninger AL, Jerome KR, Roychoudhury P, Xie H, Shrestha L, Huang ML, Pitzer VE, Iwasaki A, Omer SB, Khan K, Bogoch II, Martinello RA, Foxman EF, Landry ML, Neher RA, Ko AI, Grubaugh ND. 2020. Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* **181**:990–996. doi:10.1016/j.cell.2020.04.021

Goldberg SA, Lennerz J, Klompas M, Mark E, Pierce VM, Thompson RW, Pu CT, Ritterhouse LL, Dighe A, Rosenberg ES, Grabowski DC. 2020. Presymptomatic Transmission of Severe Acute Respiratory Syndrome Coronavirus 2 Among Residents and Staff at a Skilled Nursing Facility: Results of Real-time Polymerase Chain Reaction and Serologic Testing. *Clin Infect Dis*. doi:10.1093/cid/ciaa991

Graham NSN, Junghans C, Downes R, Sendall C, Lai H, McKirdy A, Elliott P, Howard R, Wingfield D, Priestman M, Ciechonska M, Cameron L, Storch M, Crone MA, Freemont PS, Randell P, McLaren R, Lang N, Ladhani S, Sanderson F, Sharp DJ. 2020. SARS-CoV-2 infection, clinical features and outcome of COVID-19 in United Kingdom nursing homes. *J Infect*. doi:10.1016/j.jinf.2020.05.073

Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, Tan AL, Paul LM, Brackney DE, Grewal S, Gurfield N, Van Rompay KKA, Isern S, Michael SF, Coffey LL, Loman NJ, Andersen KG. 2019. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* **20**:1–19. doi:10.1186/s13059-018-1618-7

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Hadfield J. 2020. artic-network/rampart: Read Assignment, Mapping, and Phylogenetic Analysis in Real Time. https://github.com/artic-network/rampart

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, Chen W, Hu F, Zhang Q, Zhong M, Wu Y, Zhao L, Zhang F, Cowling BJ, Li F, Leung GM. 2020. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* **26**:672–675. doi:10.1038/s41591-020-0869-5

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular biology and evolution. *Mol Biol Evol* **35**:518–522. doi:10.5281/zenodo.854445

ICS. 2020. Assessing whether COVID-19 patients will benefit from critical care, and an objective approach to capacity challenges. https://www.rcplondon.ac.uk/file/20726/download

Illingworth CJR, Hamilton WL, Jackson C, Popay A, Meredith L, Houldcroft CJ, Hosmillo M, Jahun A, Routledge M, Warne B, Caller L, Caddy S, Yakovleva A, Hall G, Khokhar FA, Feltwell T, Pinckert ML, Georgana I, Chaudhry Y, Curran M, Parmar S, Sparkes D, Rivett L, Jones NK, Sridhar S, Forrest S, Dymond T, Grainger K, Workman C, Gkrania-Klotsas E, Brown NM, Weekes MP, Baker S, Peacock SJ, Gouliouris T, Goodfellow I, De Angelis D, Török ME. 2020. A2B-COVID: A method for evaluating potential SARS-CoV-2 transmission events. *medRxiv* 2020.10.26.20219642.

Jones NK, Rivett L, Sparkes D, Forrest S, Sridhar S, Young J, Pereira-Dias J, Cormie C, Gill H, Reynolds N, Wantoch M, Routledge M, Warne B, Levy J, Jiménez WDC, Samad FNB, McNicholas C, Ferris M, Gray J, Gill M, Curran MD, Fuller S, Chaudhry A, Shaw A, Bradley JR, Hannon GJ, Goodfellow IG, Dougan G, Smith KGC, Lehner PJ, Wright G, Matheson NJ, Baker S, Weekes MP. 2020. Effective control of sars-cov-2 transmission between healthcare workers during a period of diminished community prevalence of covid-19. *Elife* **9**:1–10. doi:10.7554/eLife.59391

Jones NR, Qureshi ZU, Temple RJ, Larwood JPJ, Greenhalgh T, Bourouiba L. 2020. Two metres or one: what is the evidence for physical distancing in covid-19? *BMJ* **370 :m3223**. doi:https://doi.org/10.1136/bmj.m3223

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**:587–589. doi:10.1038/nmeth.4285

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**:772–780. doi:10.1093/molbev/mst010

Kemenesi G, Kornya L, Tóth GE, Kurucz K, Zeghbib S, Somogyi BA, Zöldi V, Urbán P, Herczeg R, Jakab F. 2020. Nursing homes and the elderly regarding the COVID-19 pandemic: situation report from Hungary. *GeroScience* **2**. doi:10.1007/s11357-020-00195-z

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Ladhani SN, Chow JY, Janarthanan R, Fok J, Crawley-Boevey E, Vusirikala A, Fernandez E, Perez MS, Tang S, Dun-Campbell K, Wynne-Evans E, Bell A, Patel B, Amin-Chowdhury Z, Aiano F, Paranthaman K, Ma T, Saavedra-Campos M, Myers R, Ellis J, Lackenby A, Gopal R, Patel M, Chand M, Brown K, Hopkins S, Consortium CG, Shetty N, Zambon M, Ramsay ME. 2020. Increased risk of SARS-CoV-2 infection in staff working across different care homes: enhanced CoVID-19 outbreak investigations in London care Homes. *J Infect* **81**:621–624. doi:10.1016/j.jinf.2020.07.027

Lansbury LE, Brown CS, Nguyen-Van-Tam JS. 2017. Influenza in long-term care facilities. *Influenza Other Respi Viruses* **11**:356–366. doi:10.1111/irv.12464

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM **00**:1–3.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079. doi:10.1093/bioinformatics/btp352

MacFadden DR, McGeer A, Athey T, Perusini S, Olsha R, Li A, Eshaghi A, Gubbay JB, Hanage WP. 2018. Use of genome sequencing to define institutional influenza outbreaks, Toronto, Ontario, Canada, 2014-15. *Emerg Infect Dis* **24**:492–497. doi:10.3201/eid2403.171499

McAuslane H, Morgan D. 2014. Communicable Disease Outbreak Management. Operational guidance. *Public Heal Engl* 1–66.

Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, Curran MD, Parmar S, Caller LG, Caddy SL, Khokhar FA, Yakovleva A, Hall G, Feltwell T, Forrest S, Sridhar S, Weekes MP, Baker S, Brown N, Moore E, Popay A, Roddick I, Reacher M, Gouliouris T, Peacock SJ, Dougan G, Török ME, Goodfellow I. 2020. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis* **0**. doi:10.1016/S1473-3099(20)30562-4

Middleton J, Reintjes R, Lopes H. 2020. Meat plants-a new front line in the covid-19 pandemic. *BMJ* **370**:1–2. doi:10.1136/bmj.m2716

NICE. 2020. NICE guideline NG159. COVID-19 rapid guideline: critical care in adults. 2. Admission to critical care.

Office for National Statistics. 2011. 2011 Census - Population and Household Estimates for England and Wales, March 2011. *Natl Census* 37.

ONS. 2020a. Deaths involving COVID-19, England and Wales - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsinvolvingcovid19englandandwales/deathsoccurringinjune2020

ONS. 2020b. Deaths involving COVID-19 in the care sector, England and Wales - Office for National

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/deathsinvolvingcovid19inthecaresectorenglandandwales/deathsoccurringupto12june2020andregisteredupto20june2020provisional

ONS. 2019a. Families and households in the UK - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2019

ONS. 2019b. Families and households - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/datasets/familiesandhouseholdsfamiliesandhouseholds

ONS. 2014. Changes in the Older Resident Care Home Population between 2001 and 2011 - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing/articles/changesintheolderresidentcarehomepopulationbetween2001and2011/2014-08-01

Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb genomics* **2**:e000056. doi:10.1099/mgen.0.000056

Public Health England. 2020a. Weekly Coronavirus Disease 2019 (COVID-19) Surveillance Report Confirmed cases in England. Year: 2020, Week: 20. *Summ COVID-19 Surveill Syst*.

Public Health England. 2020b. COVID-19 Personal protective equipment (PPE) – resource for care workers working in care homes during sustained COVID-19 transmission in England.

Quicke K, Gallichote E, Sexton N, Young M, Janich A, Gahm G, Carlton EJ, Ehrhart N, Ebel GD. 2020. Longitudinal Surveillance for SARS-CoV-2 RNA Among Asymptomatic Staff in Five Colorado Skilled Nursing Facilities: Epidemiologic, Virologic and Sequence Analysis. *medRxiv* 2020.06.08.20125989. doi:10.1101/2020.06.08.20125989

Rambaut A, Holmes E, O'Toole Á, Hill V, McCrone J, Ruis C, du Plessis L, Pybus O. 2020. SARS-CoV-2 lineages. https://cov-lineages.org/descriptions.html

Rambaut Andrew, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. doi:10.1038/s41564-020-0770-5

Rivett L, Sridhar S, Sparkes D, Routledge M, Jones NK, Forrest S, Young J, Pereira-Dias J, Hamilton WL, Ferris M, Torok ME, Meredith L, Curran M, Fuller S, Chaudhry A, Shaw A, Samworth RJ, Bradley JR, Dougan G, Smith KGC, Lehner PJ, Matheson NJ, Wright G, Goodfellow I, Baker S, Weekes MP. 2020. Screening of healthcare workers for SARS-CoV-2 highlights the role of asymptomatic

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

carriage in COVID-19 transmission. *Elife* **9**:1–20. doi:10.7554/eLife.58728

Seemann T. 2020. tseemann/snp-dists: Pairwise SNP distance matrix from a FASTA sequence alignment. https://github.com/tseemann/snp-dists

Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. 2019. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. *Mol Biol Evol* **36**:587–603. doi:10.1093/molbev/msy242

Strausbaugh LJ, Sukumar SR, Joseph CL. 2003. Infectious disease outbreaks in nursing homes: An unappreciated hazard for frail elderly persons. *Clin Infect Dis* **36**:870–876. doi:10.1086/368197

Tonkin-Hill G. 2020. gtonkinhill/fasttranscluster. https://github.com/gtonkinhill/fasttranscluster

UK government. 2020a. COVID-19: number of outbreaks in care homes - management information - GOV.UK. https://www.gov.uk/government/statistical-data-sets/covid-19-number-of-outbreaks-in-care-homes-management-information

UK government. 2020b. Update on policies for visiting arrangements in care homes. https://www.gov.uk/government/publications/visiting-care-homes-during-coronavirus/update-on-policies-for-visiting-arrangements-in-care-homes#section-4

UK government. 2020c. Vivaldi 1: COVID-19 care homes study report - GOV.UK. https://www.gov.uk/government/publications/vivaldi-1-coronavirus-covid-19-care-homes-study-report/vivaldi-1-covid-19-care-homes-study-report

UK government. 2020d. Get coronavirus tests for a care home - GOV.UK. https://www.gov.uk/apply-coronavirus-test-care-home

Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, Curtis HJ, Mehrkar A, Evans D, Inglesby P, Cockburn J, McDonald HI, MacKenna B, Tomlinson L, Douglas IJ, Rentsch CT, Mathur R, Wong AYS, Grieve R, Harrison D, Forbes H, Schultze A, Croker R, Parry J, Hester F, Harper S, Perera R, Evans SJW, Smeeth L, Goldacre B. 2020. OpenSAFELY: factors associated with COVID-19 death in 17 million patients. *Nature*. doi:10.1038/s41586-020-2521-4

Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. 2020. A new coronavirus associated with human respiratory disease in China. *Nature* **579**:265–269. doi:10.1038/s41586-020-2008-3

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

Zhang J, Litvinova M, Wang W, Wang Y, Deng X, Chen Xinghui, Li M, Zheng W, Yi L, Chen Xinhua, Wu Q, Liang Y, Wang X, Yang J, Sun K, Longini IM, Halloran ME, Wu P, Cowling BJ, Merler S, Viboud

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

C, Vespignani A, Ajelli M, Yu H. 2020. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. *Lancet Infect Dis* **20**:793–802. doi:10.1016/S1473-3099(20)30230-9

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Main display items – legends

Figure 1

Study flow diagram

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 1, supplement 1

Flow diagram for identifying care homes from Cambridge-COGUK metadata

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 1, supplement 2

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 1, supplement 3

UK care home testing policy timeline

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 2

Care home locations by county, showing nursing and residential homes

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 2, supplement 1

Distribution of cases per care home

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 3

Epidemic curves for EoE and CUH showing care home residents

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 3, supplement 1

Care home residents per week showing genome sequencing site

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 4

Odds ratios for mortality at 30 days

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 4, supplement 1

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 5

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 5, Supplement 1

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 5, supplement 2

Distribution of pairwise SNP differences between care home samples

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 6

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 7

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 7, supplement 1

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 7, supplement 2

Histogram of pairwise transmission probabilities between care home samples

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 7, supplement 3

Transmission probability threshold vs number of care home clusters

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 7, supplement 4

Pairwise SNP difference distribution between samples within clusters

Within each cluster, 673 / 775 (86.8%) of pairwise links that had a ≥15% probability of transmission with ≤2 intermediate hosts had 0 or 1 pairwise SNP differences (maximum 4).

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 7, supplement 5

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

Within each cluster, 756 / 775 (97.5%) of pairwise links that had a ≥15% probability of transmission with ≤2 intermediate hosts cases were sampled <14 days apart (maximum 22 days).

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 7, Supplement 6

Pairwise date difference distribution between samples within each cluster

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Table 1

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%)

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Table 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%

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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%)

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Table 4

Variable	OR	95% CI low	95% CI high	*P* value
Age >= 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

Odds ratios for mortality at 30 days

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Table 5

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%

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Table 7

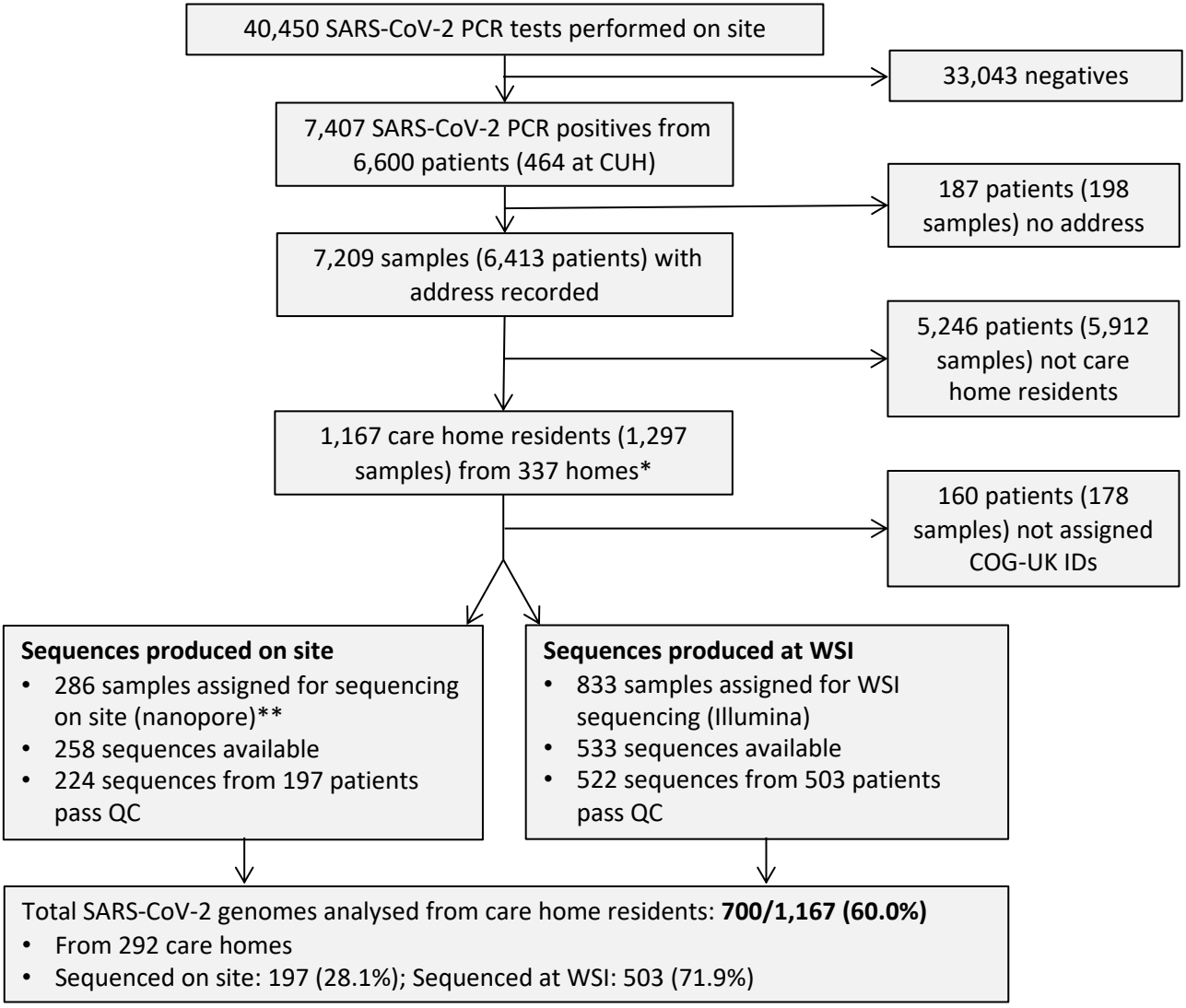
Category	Counts (%)
Care home residents with genomic data	700
Care home residents with genomic data that could be linked to hospitalisation data	694/700 (99.1%)
Hospitalised during study period	470/694 (67.7%)
Hospitalised due to COVID-19	398/694 (57.3%)
Suspected hospital-acquired COVID-19	40/694 (5.76%)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

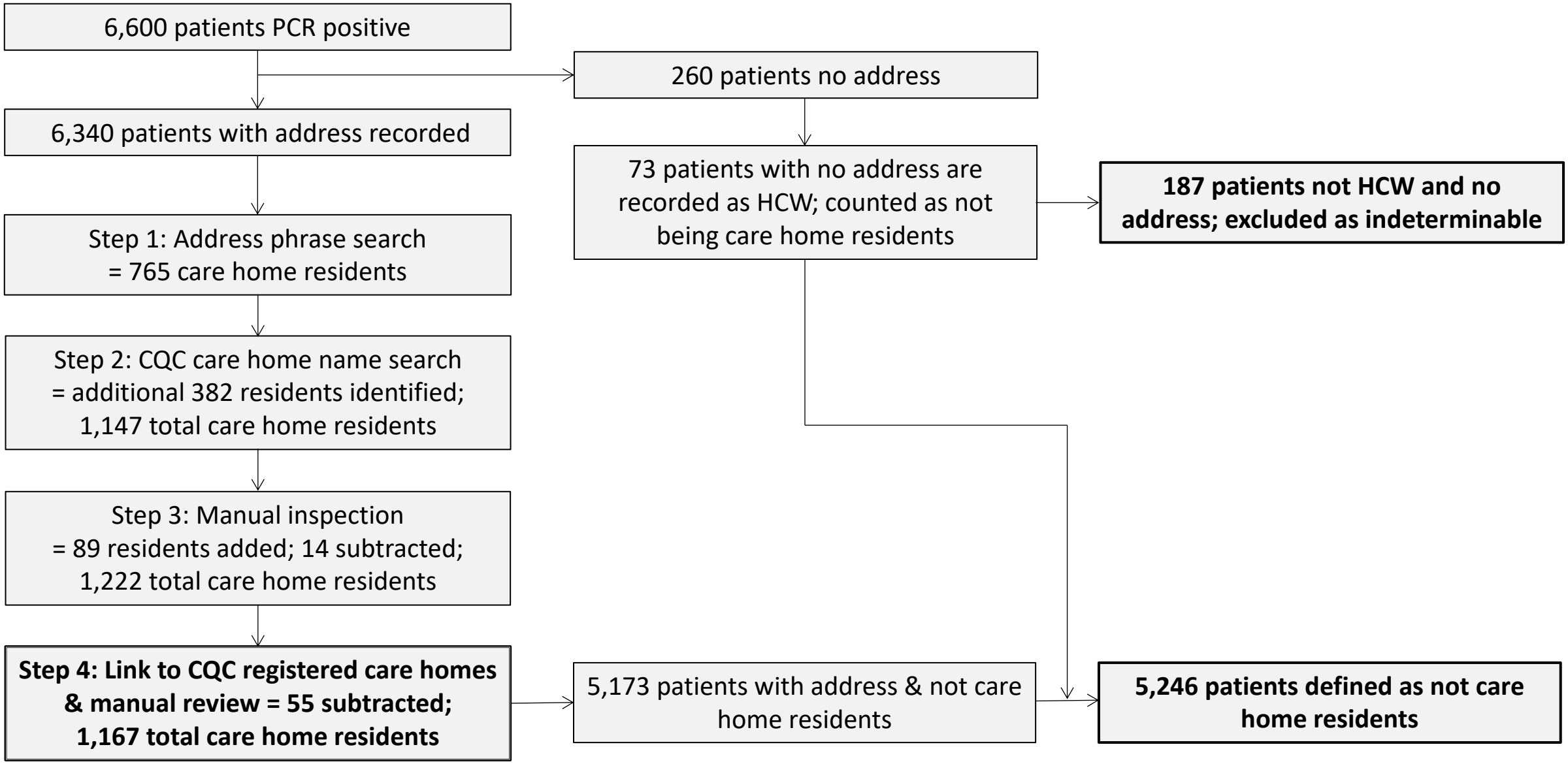
Hospitalisation data for the 700 care home residents with genomic data available

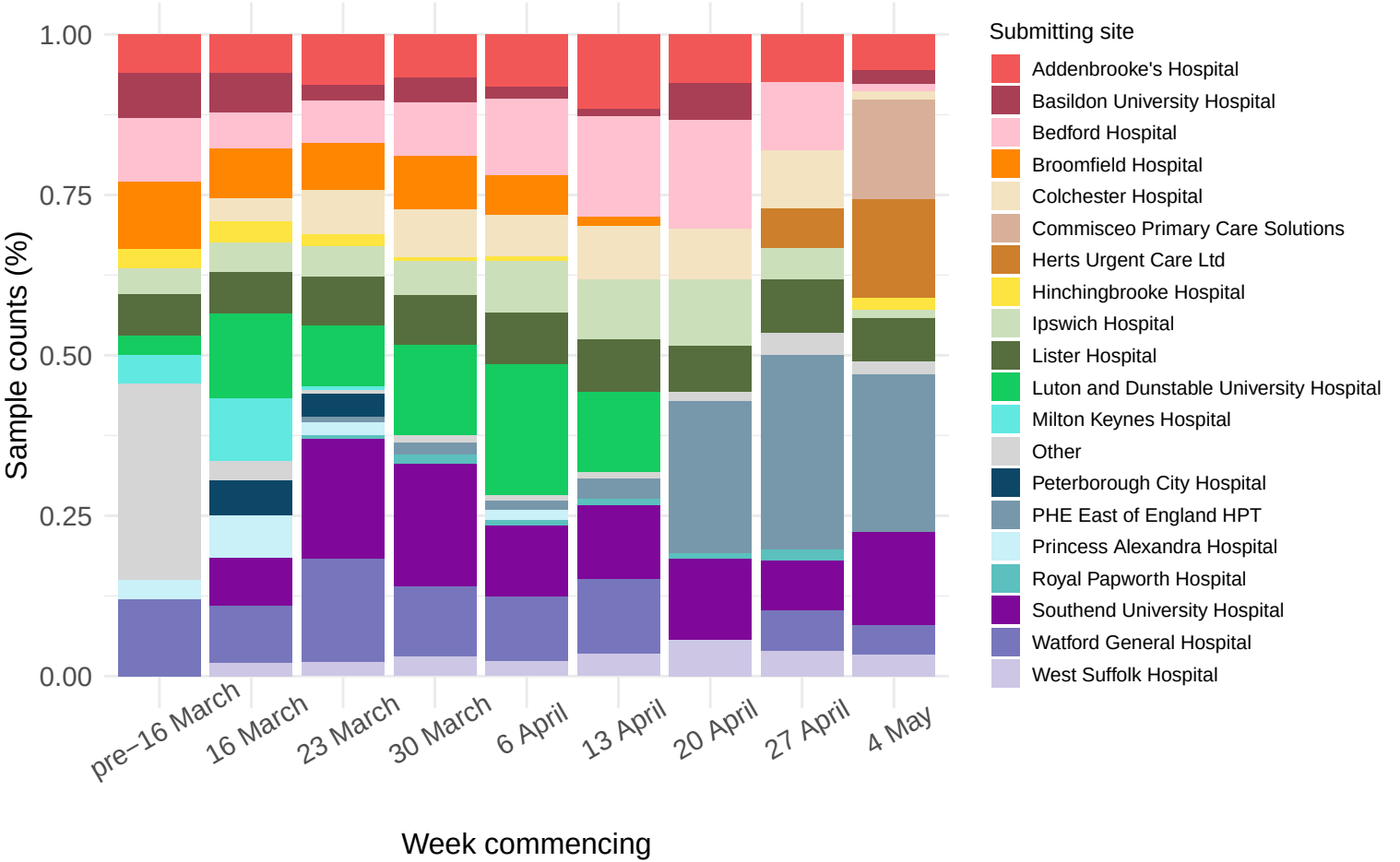
700/1,167 (60.0%) care home residents identified in the study had genomic data available and were used to define care home SARS-CoV-2 clusters. We investigated the proportions of these care home residents that were hospitalised and may have acquired their infections through interactions with hospitals.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

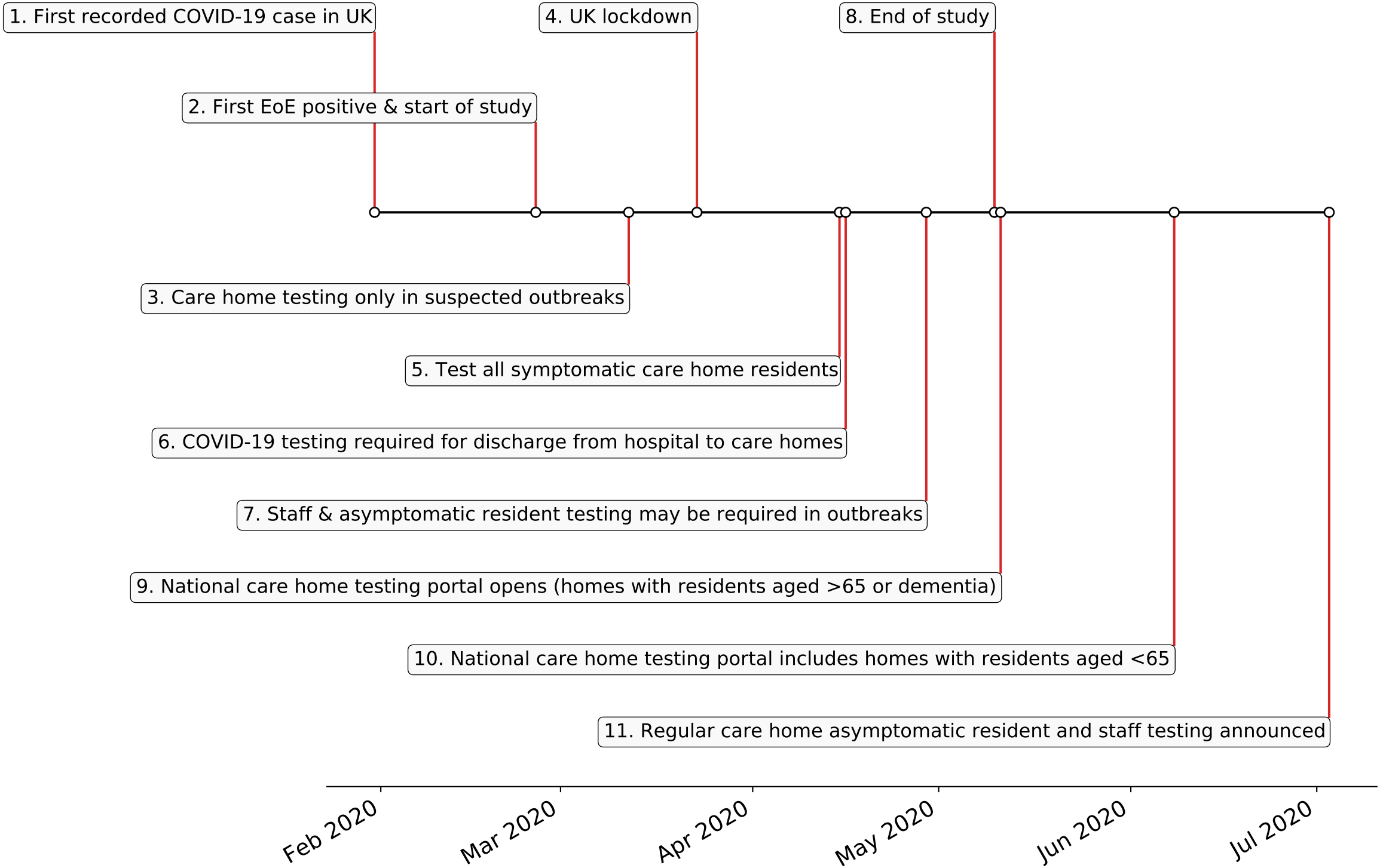
medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

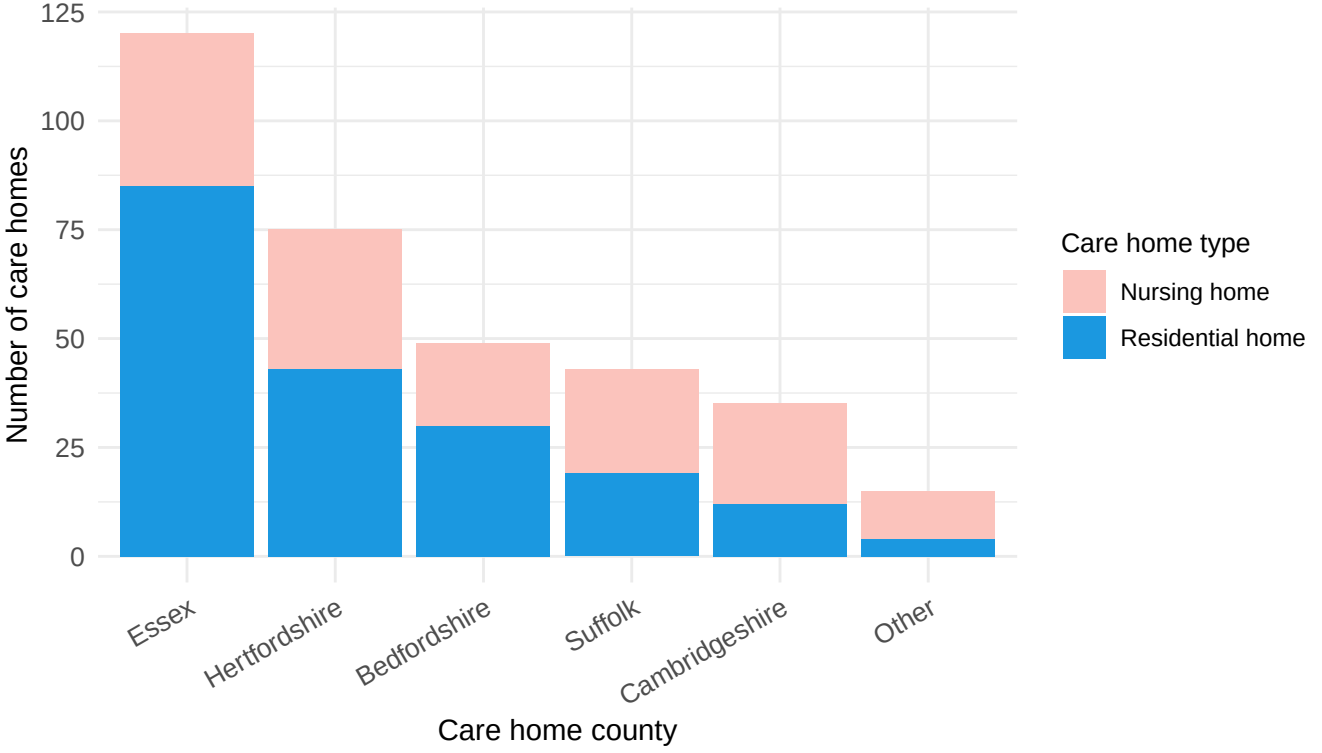
Supplementary File 1. Supplementary materials for "Genomic epidemiology of COVID-19 in care homes in the East of England"

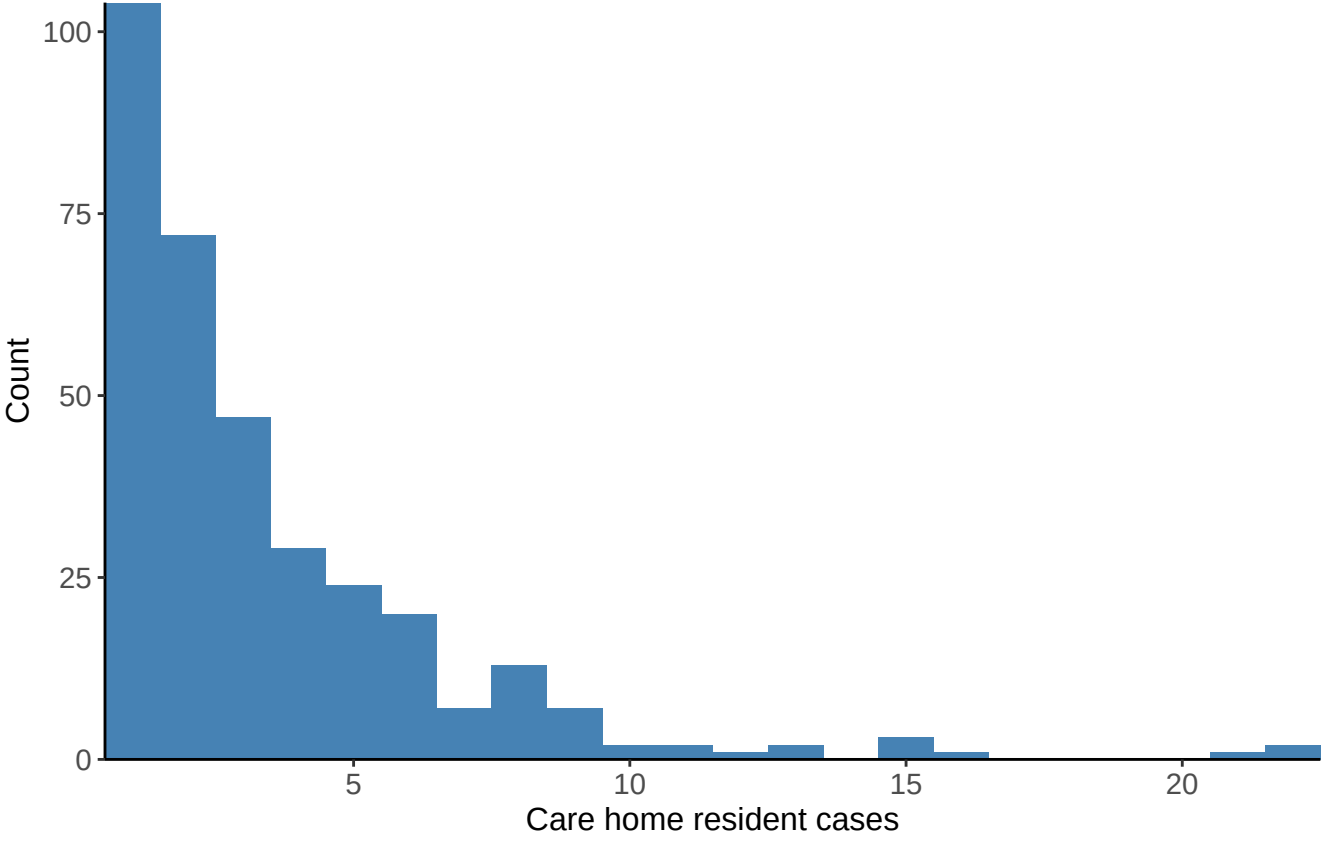

40,450 SARS-CoV-2 PCR tests performed on site
33,043 negatives
7,407 SARS-CoV-2 PCR positives from 6,600 patients (464 at CUH)
187 patients (198 samples) no address
7,209 samples (6,413 patients) with address recorded
5,246 patients (5,912 samples) not care home residents
1,167 care home residents (1,297 samples) from 337 homes*
160 patients (178 samples) not assigned COG-UK IDs
Sequences produced on site
• 286 samples assigned for sequencing on site (nanopore)**
• 258 sequences available
• 224 sequences from 197 patients pass QC
Sequences produced at WSI
• 833 samples assigned for WSI sequencing (Illumina)
• 533 sequences available
• 522 sequences from 503 patients pass QC
Total SARS-CoV-2 genomes analysed from care home residents: 700/1,167 (60.0%)
• From 292 care homes
• Sequenced on site: 197 (28.1%); Sequenced at WSI: 503 (71.9%)

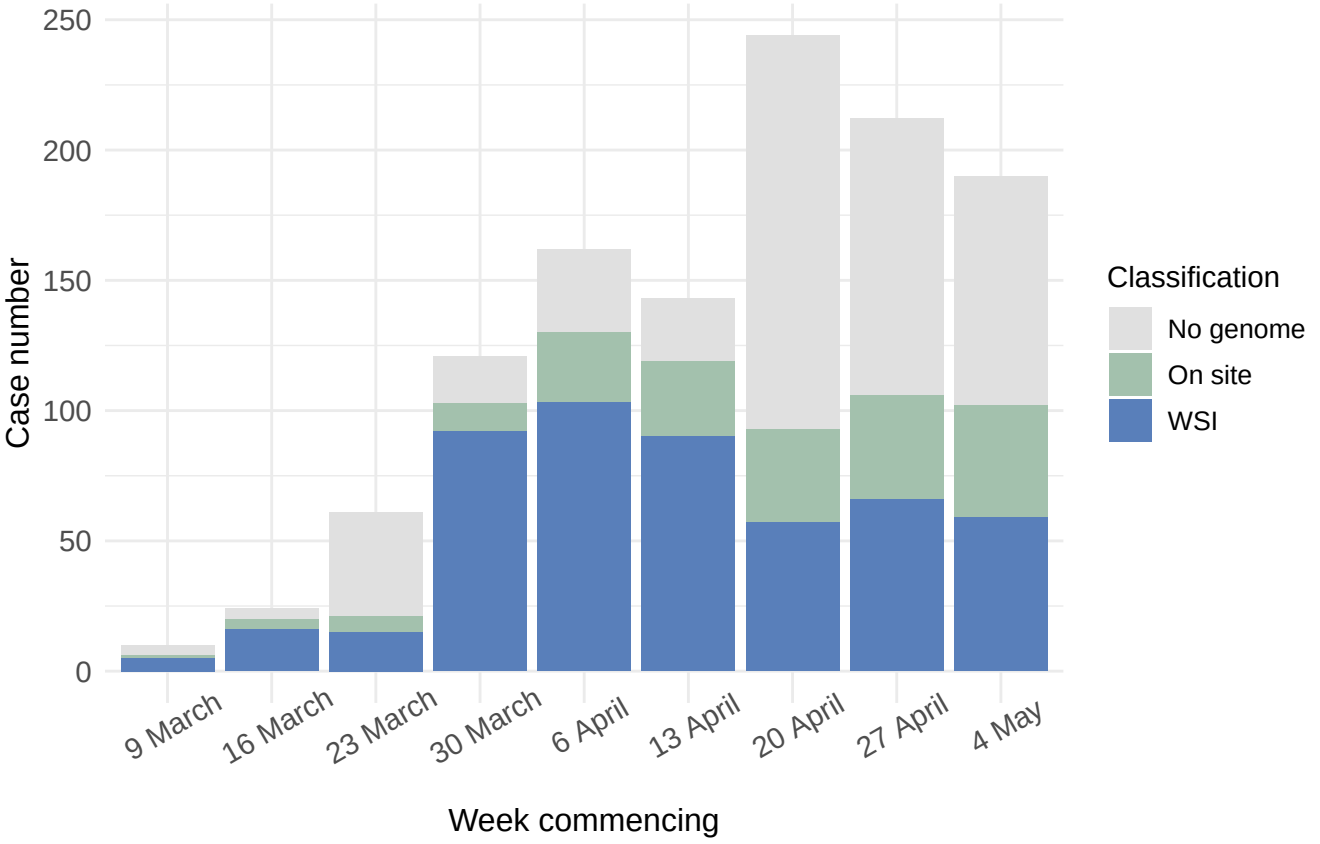

6,600 patients PCR positive
260 patients no address
6,340 patients with address recorded
73 patients with no address are recorded as HCW; counted as not being care home residents
187 patients not HCW and no address; excluded as indeterminable
Step 1: Address phrase search = 765 care home residents
Step 2: CQC care home name search = additional 382 residents identified; 1,147 total care home residents
Step 3: Manual inspection = 89 residents added; 14 subtracted; 1,222 total care home residents
Step 4: Link to CQC registered care homes & manual review = 55 subtracted; 1,167 total care home residents
5,173 patients with address & not care home residents
5,246 patients defined as not care home residents

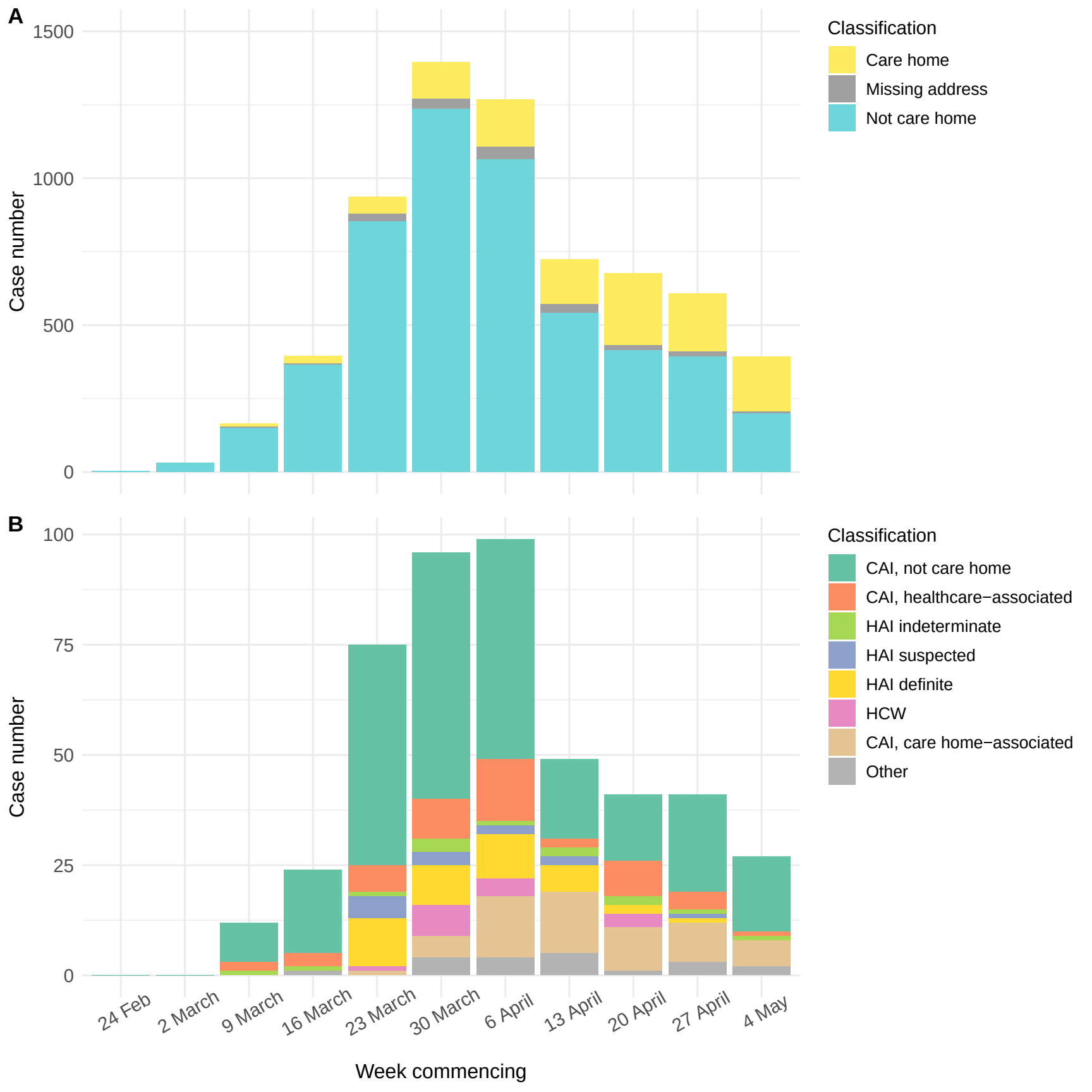

Sample counts (%)
1.00
0.75
0.50
0.25
0.00
pre−16 March
16 March
23 March
30 March
6 April
13 April
20 April
27 April
4 May
Week commencing
Submitting site
Addenbrooke's Hospital
Basildon University Hospital
Bedford Hospital
Broomfield Hospital
Colchester Hospital
Commisceo Primary Care Solutions
Herts Urgent Care Ltd
Hinchingbrooke Hospital
Ipswich Hospital
Lister Hospital
Luton and Dunstable University Hospital
Milton Keynes Hospital
Other
Peterborough City Hospital
PHE East of England HPT
Princess Alexandra Hospital
Royal Papworth Hospital
Southend University Hospital
Watford General Hospital
West Suffolk Hospital

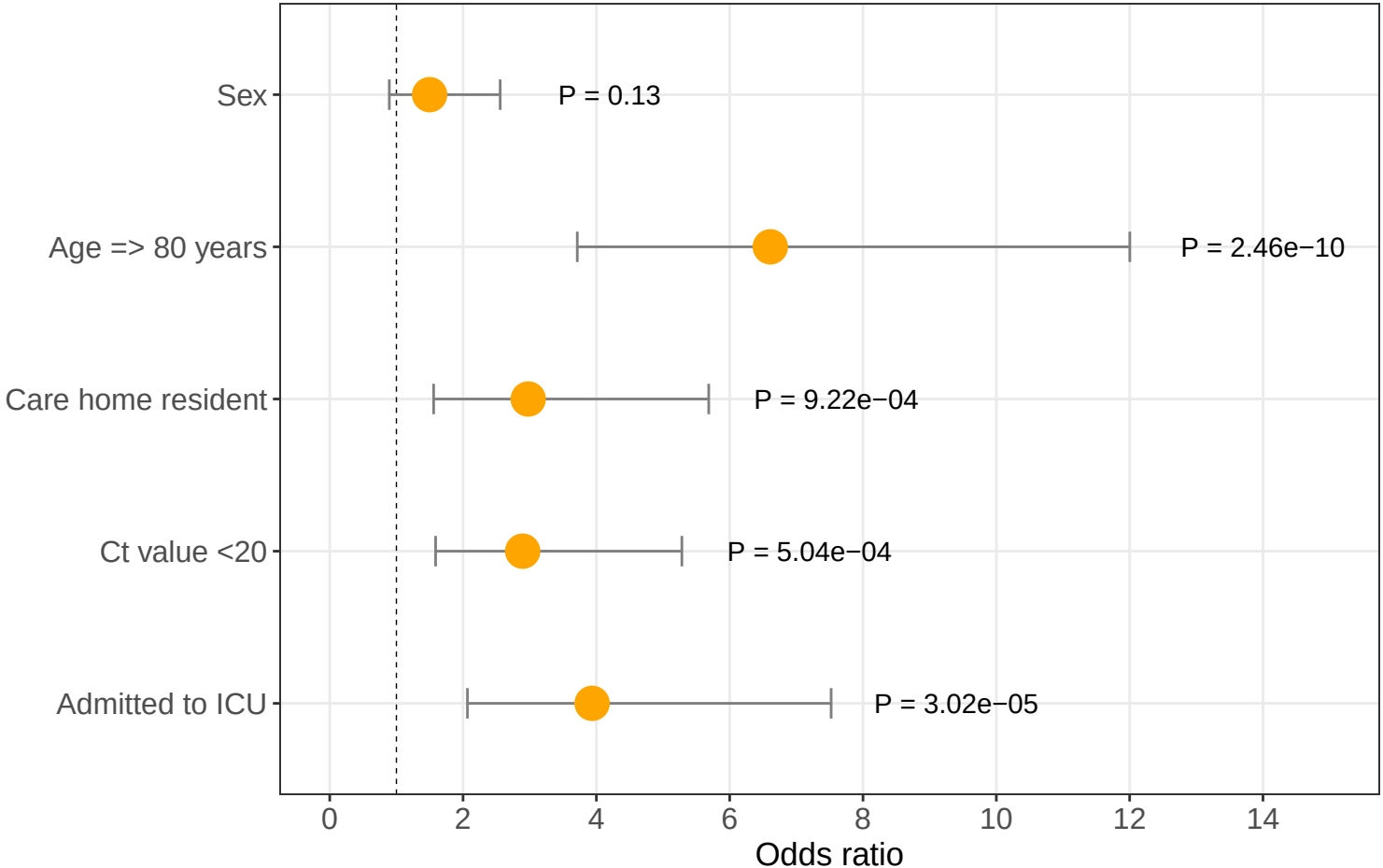

1. First recorded COVID-19 case in UK
2. First EoE positive & start of study
3. Care home testing only in suspected outbreaks
4. UK lockdown
5. Test all symptomatic care home residents
6. COVID-19 testing required for discharge from hospital to care homes
7. Staff & asymptomatic resident testing may be required in outbreaks
8. End of study
9. National care home testing portal opens (homes with residents aged >65 or dementia)
10. National care home testing portal includes homes with residents aged <65
11. Regular care home asymptomatic resident and staff testing announced
Feb 2020
Mar 2020
Apr 2020
May 2020
Jun 2020
Jul 2020


125
100
75
50
25
0
Number of care homes
Essex
Hertfordshire
Bedfordshire
Suffolk
Cambridgeshire
Other
Care home county
Care home type
Nursing home
Residential home

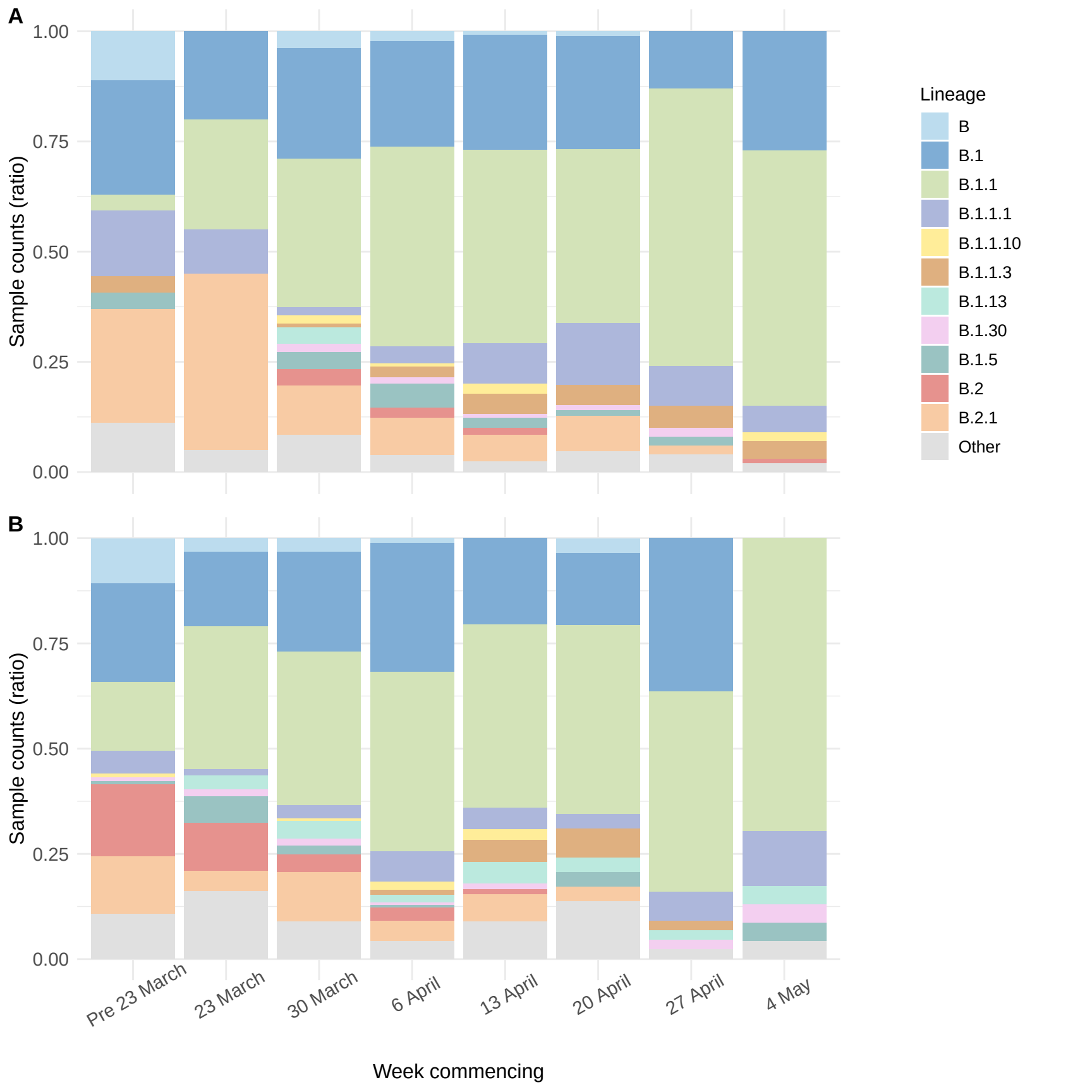

100
75
50
25
0
Count
5
10
15
20
Care home resident cases

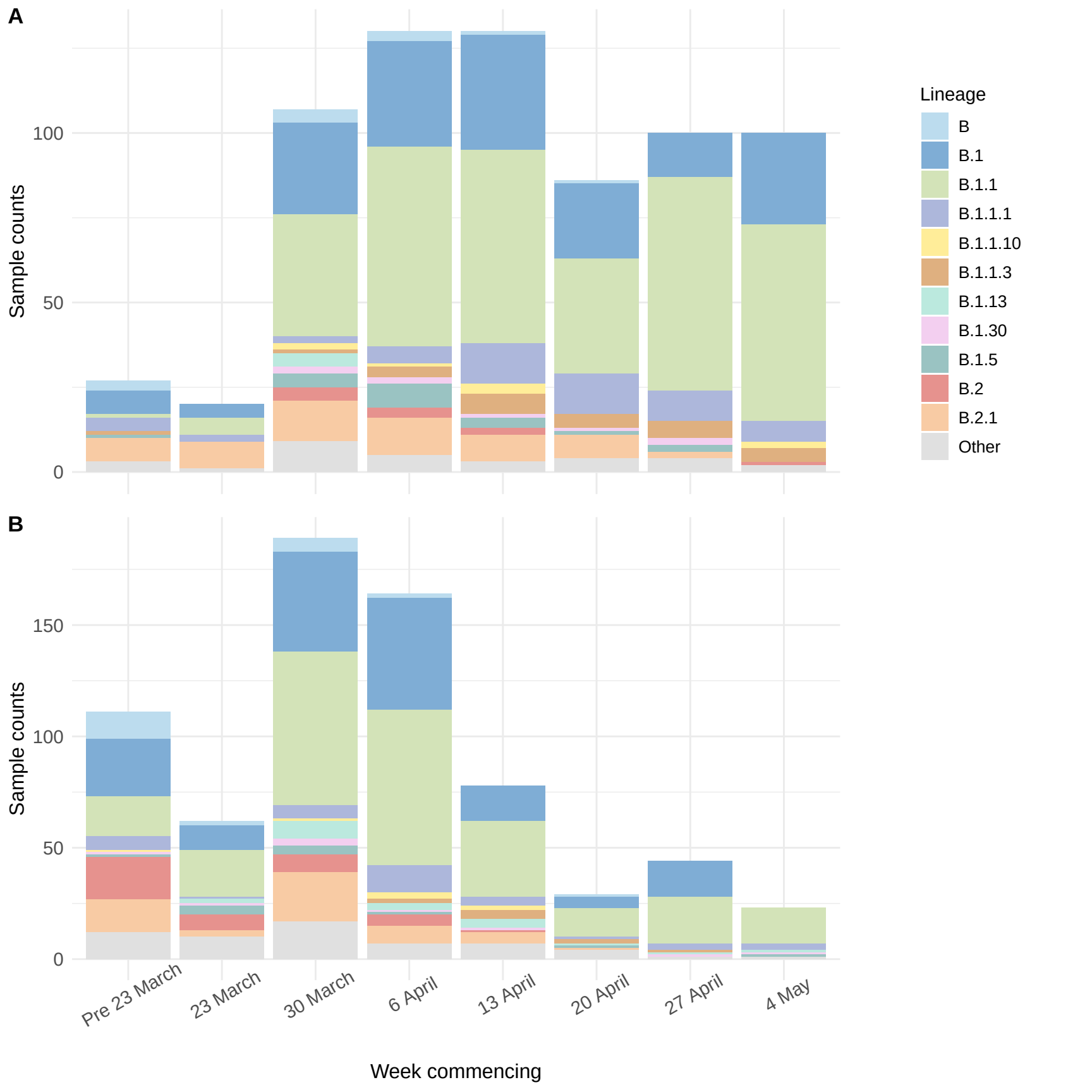

250
200
150
100
50
0
Case number
9 March
16 March
23 March
30 March
6 April
13 April
20 April
27 April
4 May
Week commencing
Classification
No genome
On site
WSI

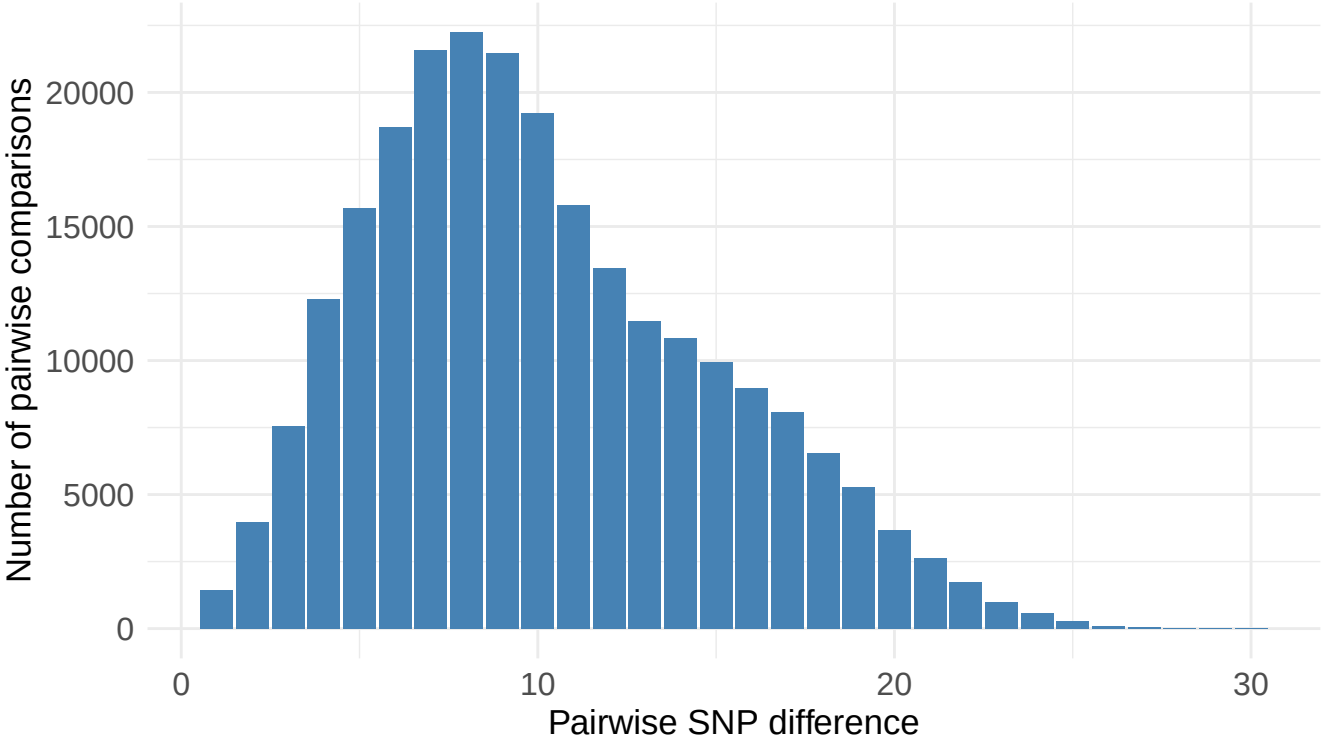

A
1500
1000
500
0
Case number
Classification
Care home
Missing address
Not care home
B
100
75
50
25
0
Case number
Classification
CAI, not care home
CAI, healthcare−associated
HAI indeterminate
HAI suspected
HAI definite
HCW
CAI, care home−associated
Other
24 Feb
2 March
9 March
16 March
23 March
30 March
6 April
13 April
20 April
27 April
4 May
Week commencing

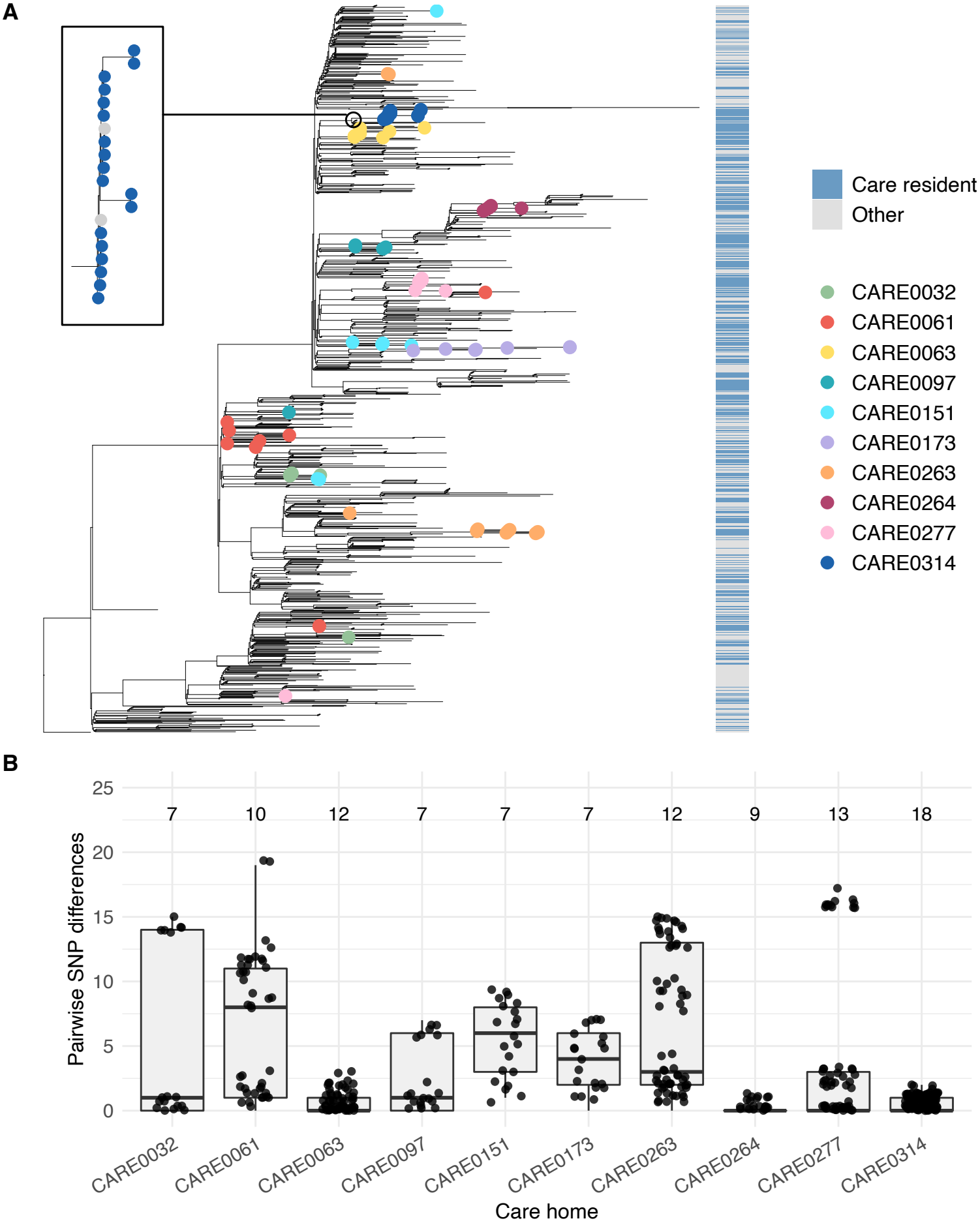

Sex
P = 0.13
Age => 80 years
P = 2.46e−10
Care home resident
P = 9.22e−04
Ct value <20
P = 5.04e−04
Admitted to ICU
P = 3.02e−05
0
2
4
6
8
10
12
14
Odds ratio

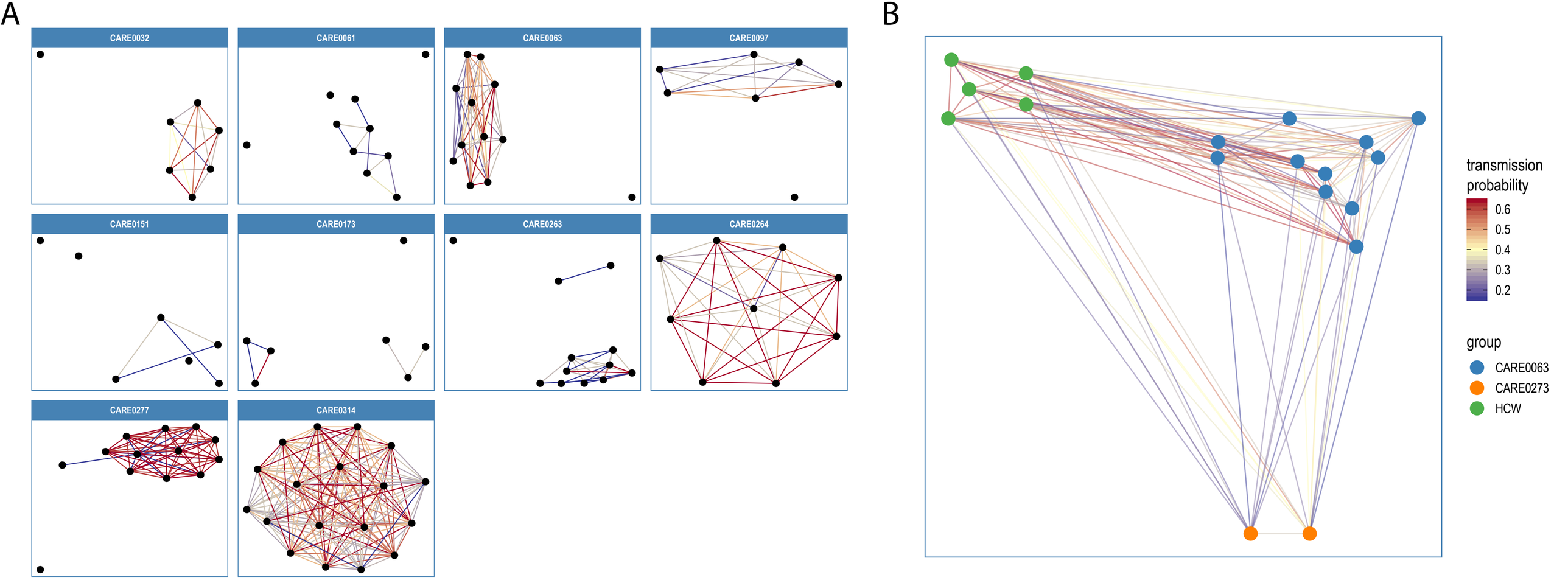
Mortality
Age
Care resident
Mortality
Age
Care resident
300
200
100
0
100
75
50
25
0
Died
Survived
0
25
50
75
100
Care resident
Independent

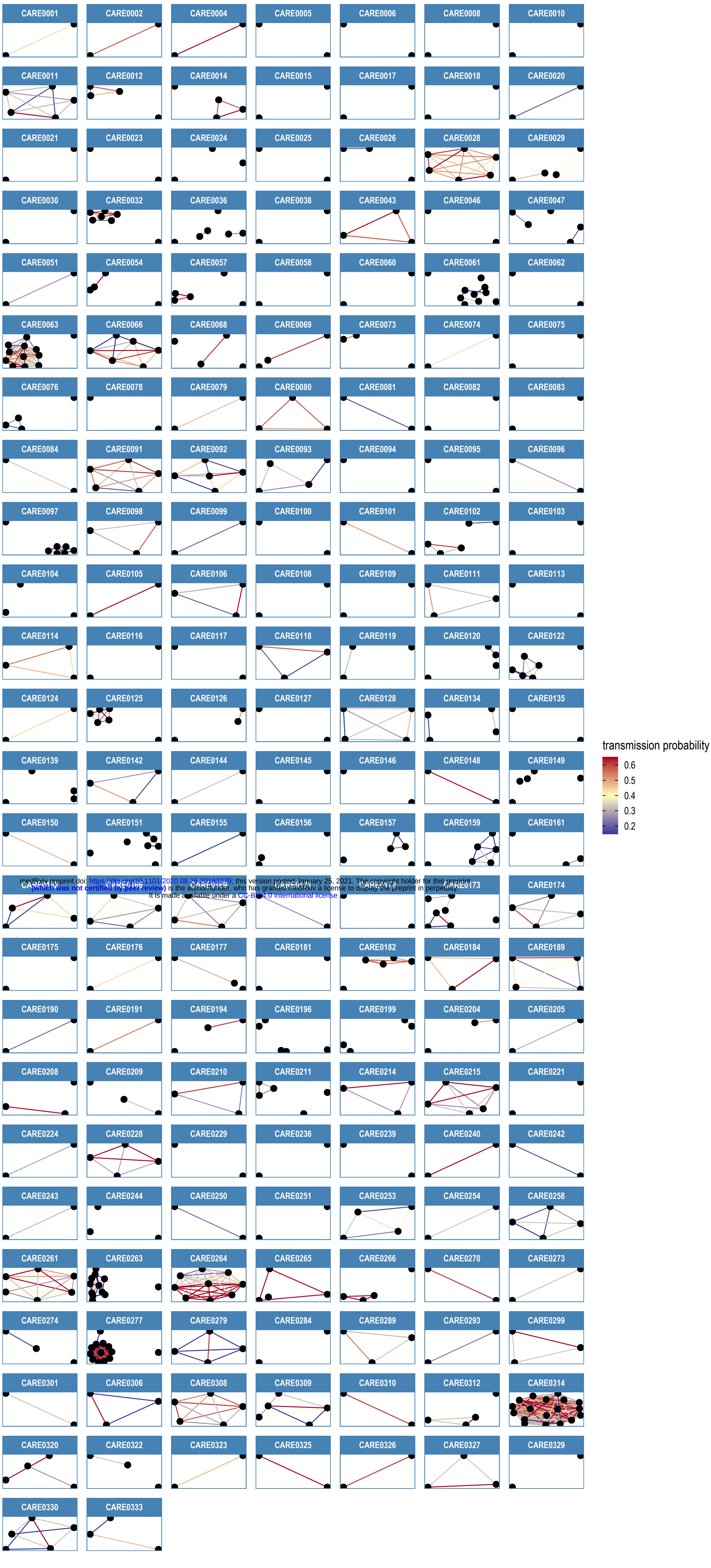

A
B
1.00
0.75
0.50
0.25
0.00
Sample counts (ratio)
Pre 23 March
23 March
30 March
6 April
13 April
20 April
27 April
4 May
Week commencing
Lineage
B
B.1
B.1.1
B.1.1.1
B.1.1.10
B.1.1.3
B.1.13
B.1.30
B.1.5
B.2
B.2.1
Other


A
B
Sample counts
0
50
100
150
Pre 23 March
23 March
30 March
6 April
13 April
20 April
27 April
4 May
Week commencing
Lineage
B
B.1
B.1.1
B.1.1.1
B.1.1.10
B.1.1.3
B.1.13
B.1.30
B.1.5
B.2
B.2.1
Other

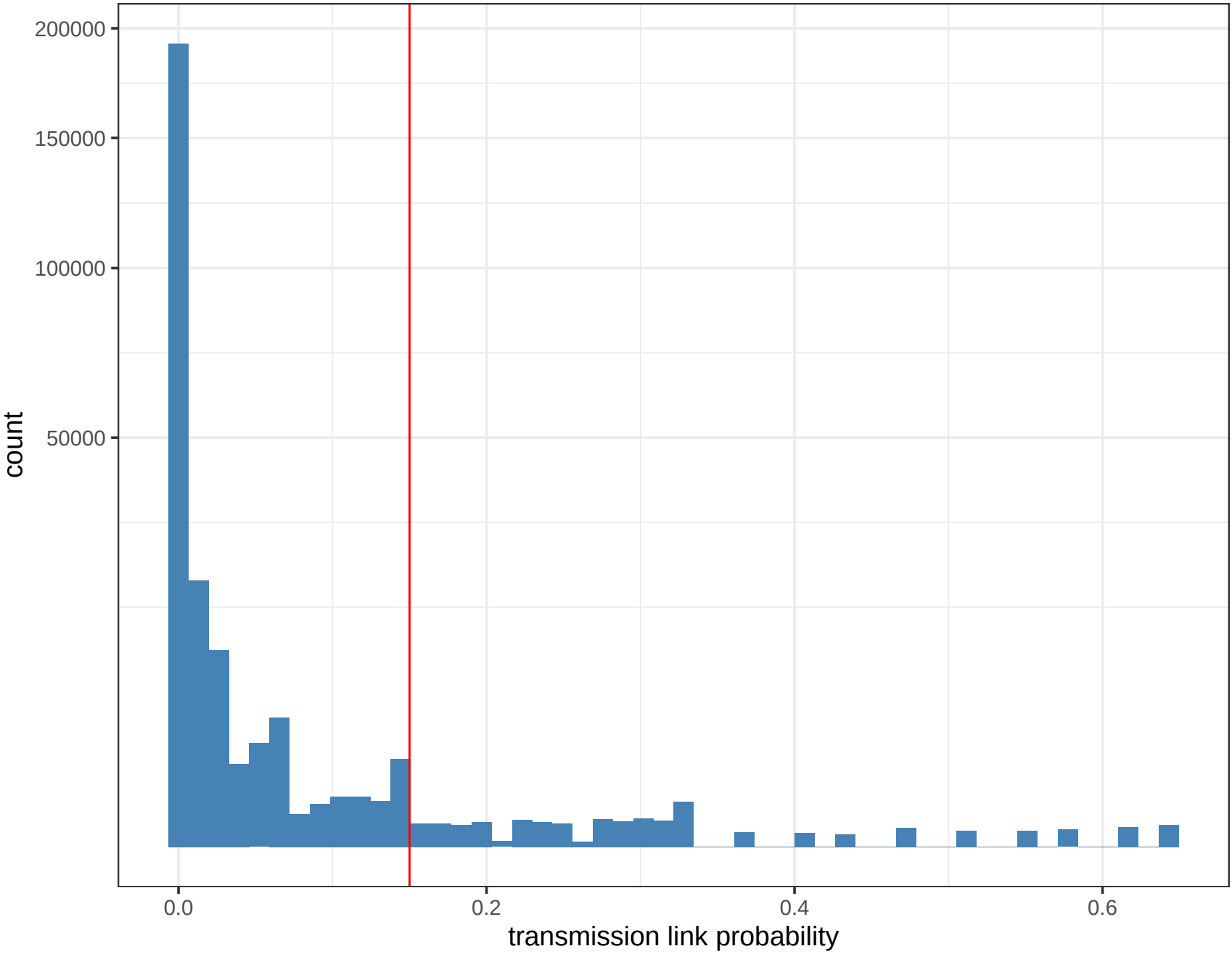

Number of pairwise comparisons
20000
15000
10000
5000
0
0
10
20
30
Pairwise SNP difference

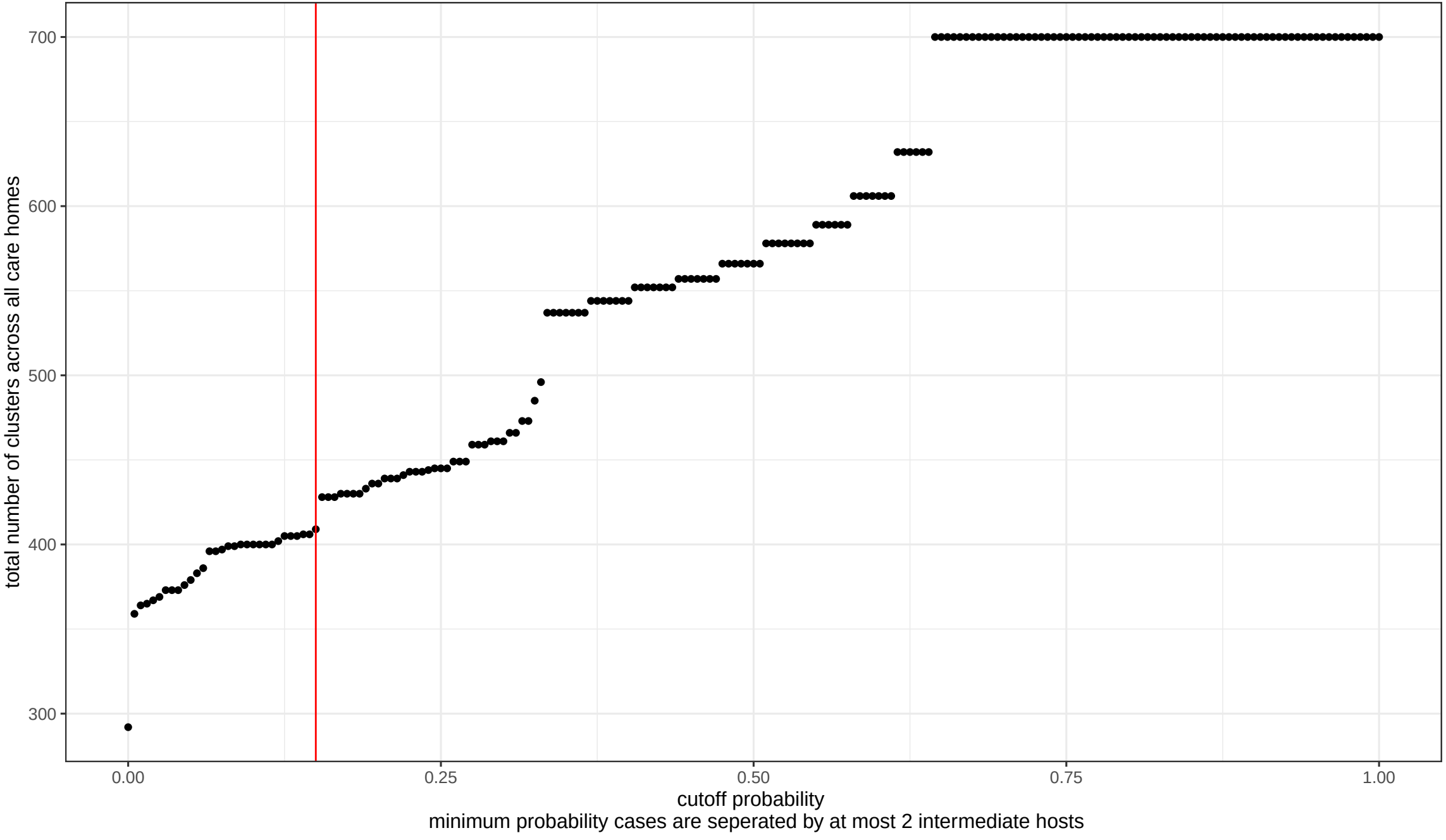

A
Care resident
Other
CARE0032
CARE0061
CARE0063
CARE0097
CARE0151
CARE0173
CARE0263
CARE0264
CARE0277
CARE0314
B
Pairwise SNP differences
25
20
15
10
5
0
7
10
12
7
7
7
12
9
13
18
CARE0032
CARE0061
CARE0063
CARE0097
CARE0151
CARE0173
CARE0263
CARE0264
CARE0277
CARE0314
Care home

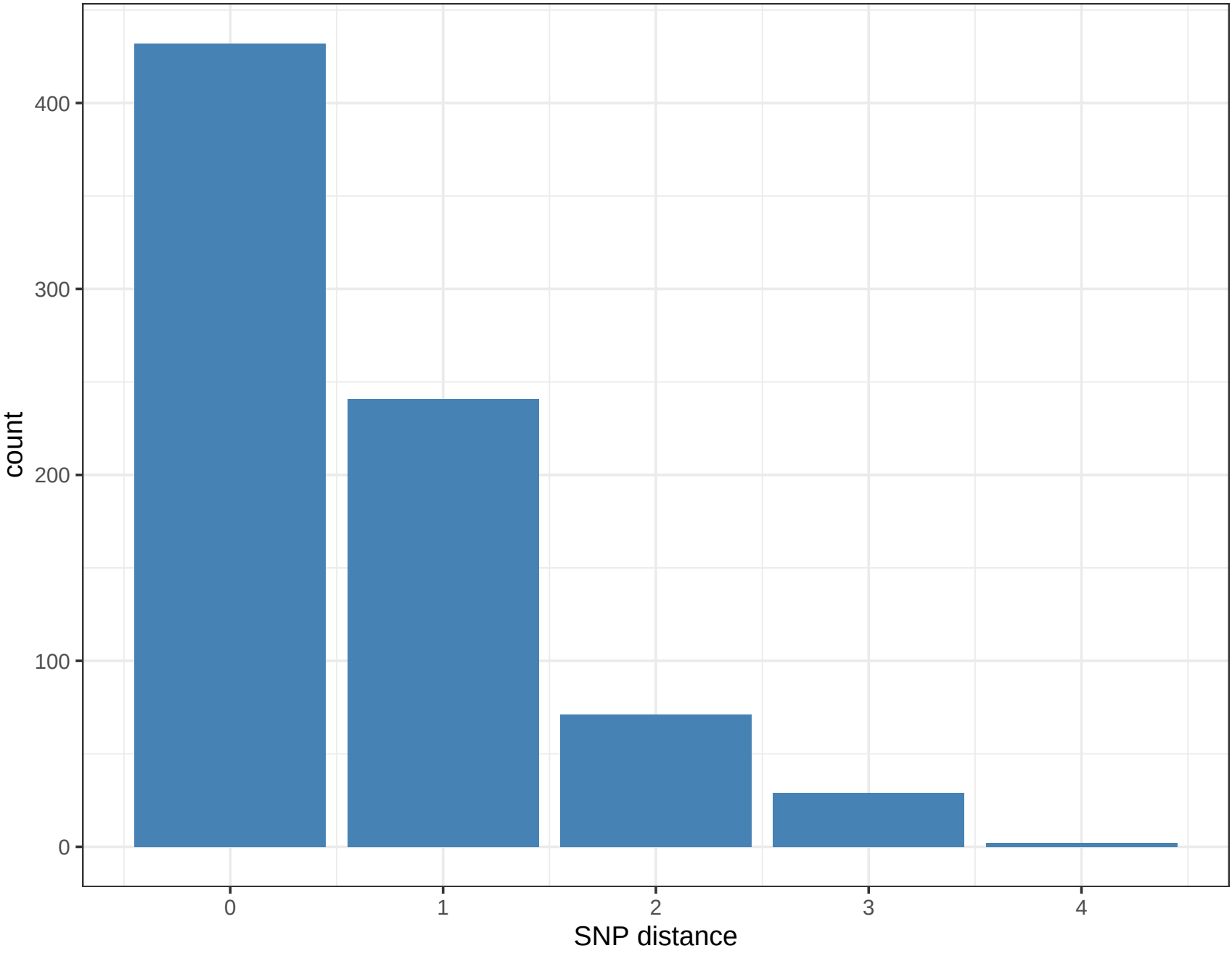

A
CARE0032
CARE0061
CARE0063
CARE0097
CARE0151
CARE0173
CARE0263
CARE0264
CARE0277
CARE0314
B
transmission probability
0.6
0.5
0.4
0.3
0.2
group
CARE0063
CARE0273
HCW

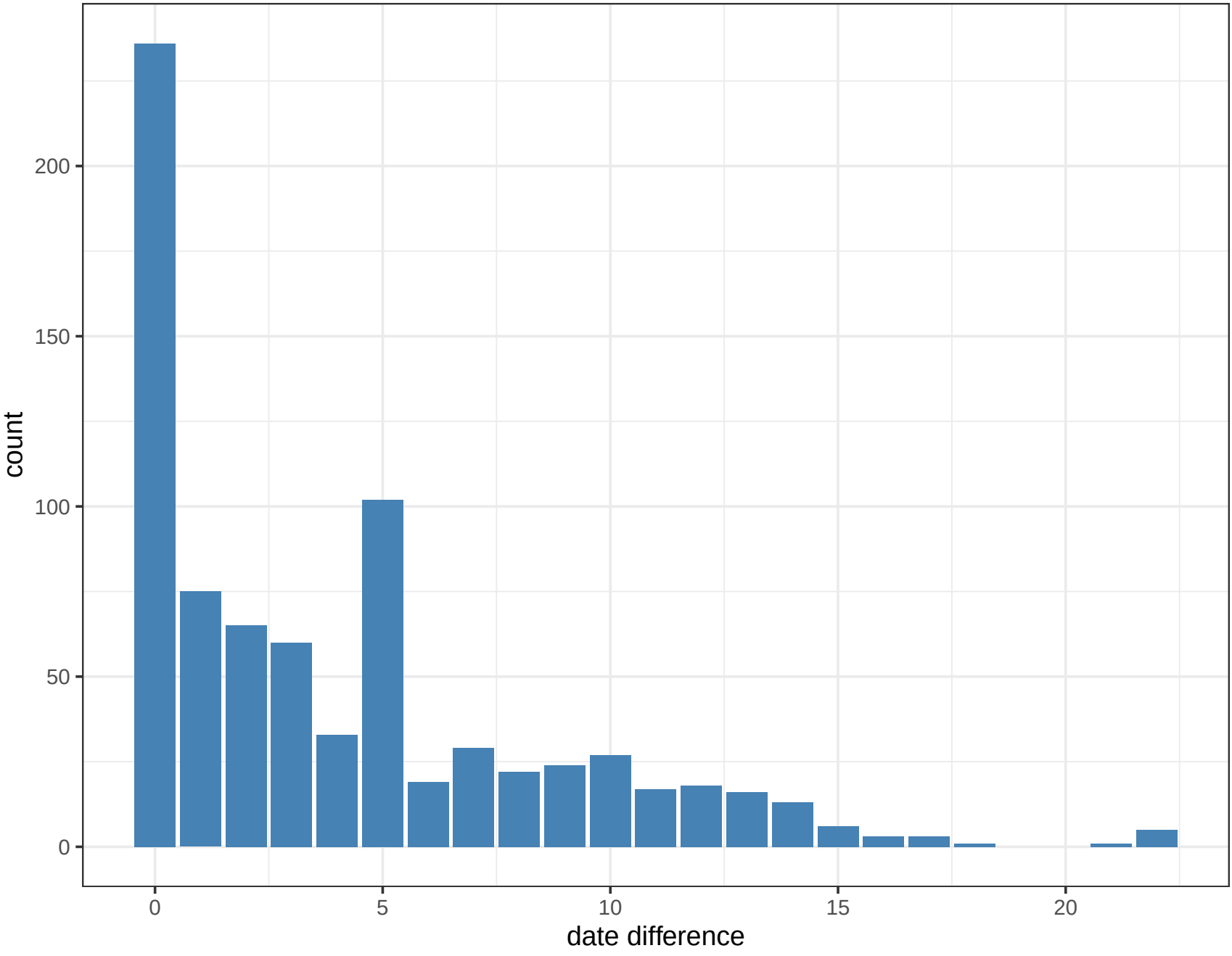


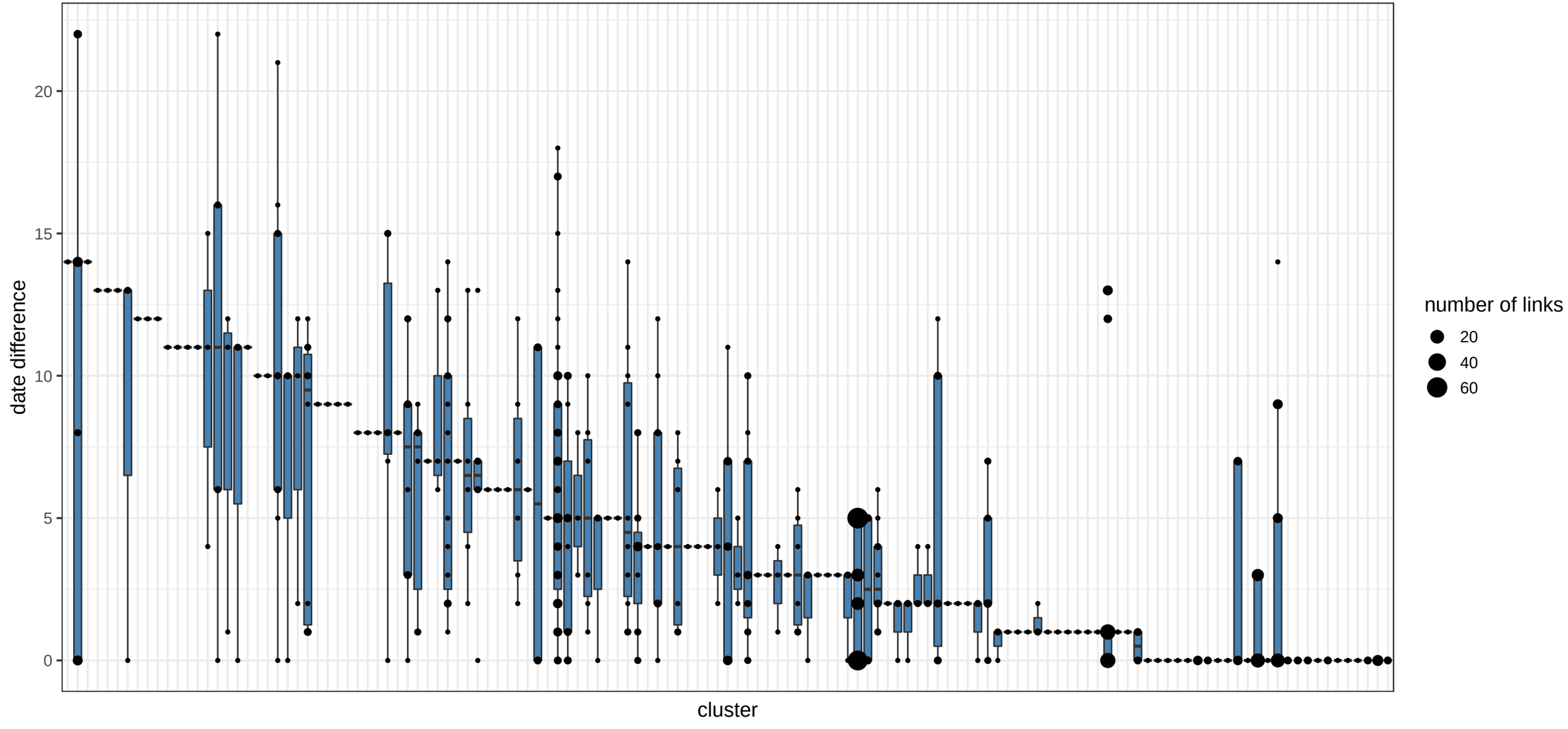
medRxiv preprint doi: https://doi.org/10.1101/2020.08.26.20182279; this version posted January 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-4.0 International license


200000
150000
100000
50000
count
0.0
0.2
0.4
0.6
transmission link probability


700
600
500
400
300
total number of clusters across all care homes
0.00
0.25
0.50
0.75
1.00
cutoff probability
minimum probability cases are seperated by at most 2 intermediate hosts


count
400
300
200
100
0
0
1
2
3
4
SNP distance


count
200
150
100
50
0
0
5
10
15
20
date difference


date difference
20
15
10
5
0
cluster
number of links
20
40
60