RESEARCH ARTICLE

Figures and figure supplements

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

William L Hamilton *et al*

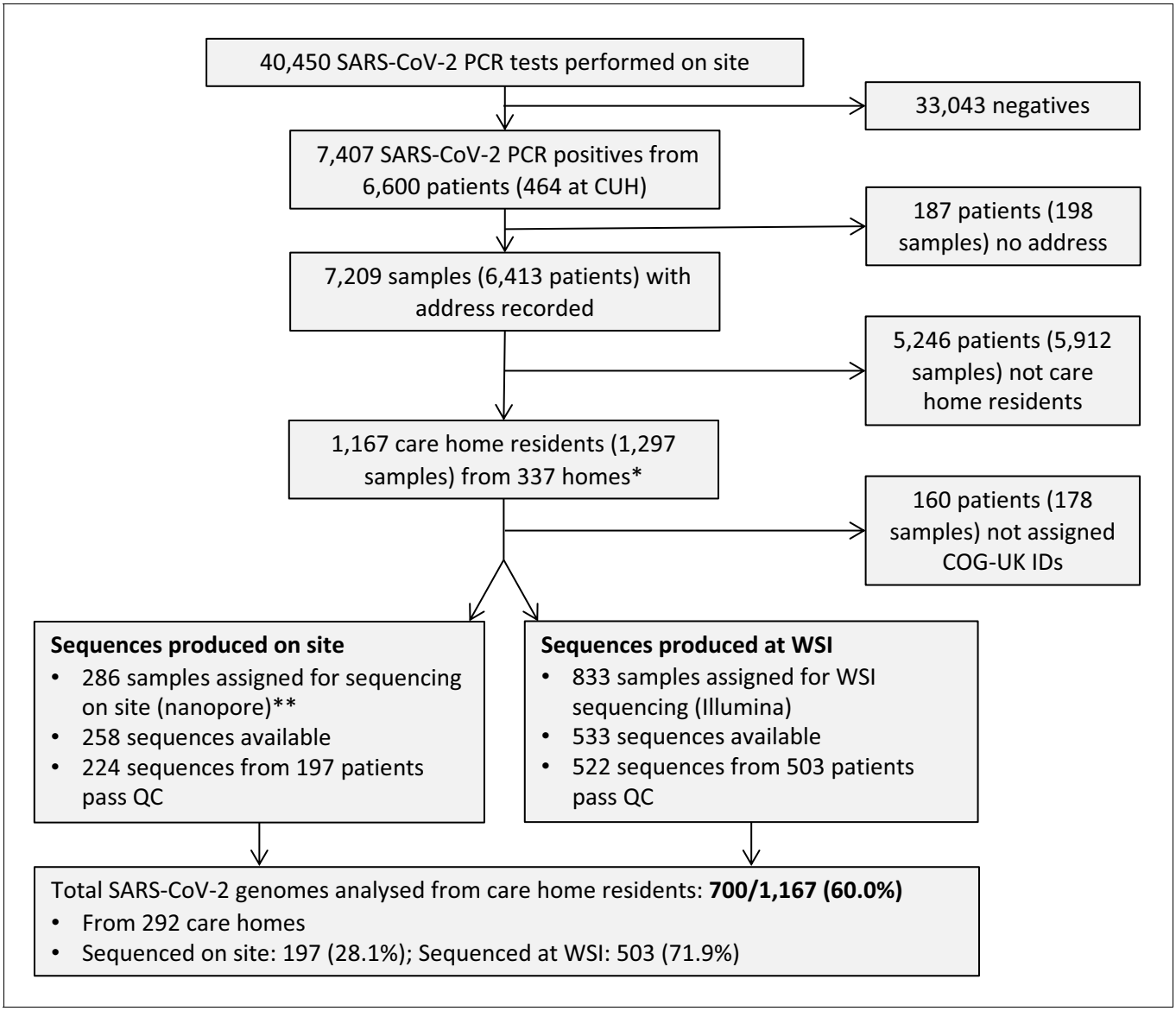


Figure 1. Study flow diagram Out of 6600 patients testing positive in the Cambridge Microbiology Public Health Laboratory (CMPHL) during the study period, 1167 were identified as being care home residents from 337 care homes. (The methodology for assigning care home status is described in main text and ***Figure 1—figure supplement 1***). Out of 1297 samples from 1167 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 Materials and 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.

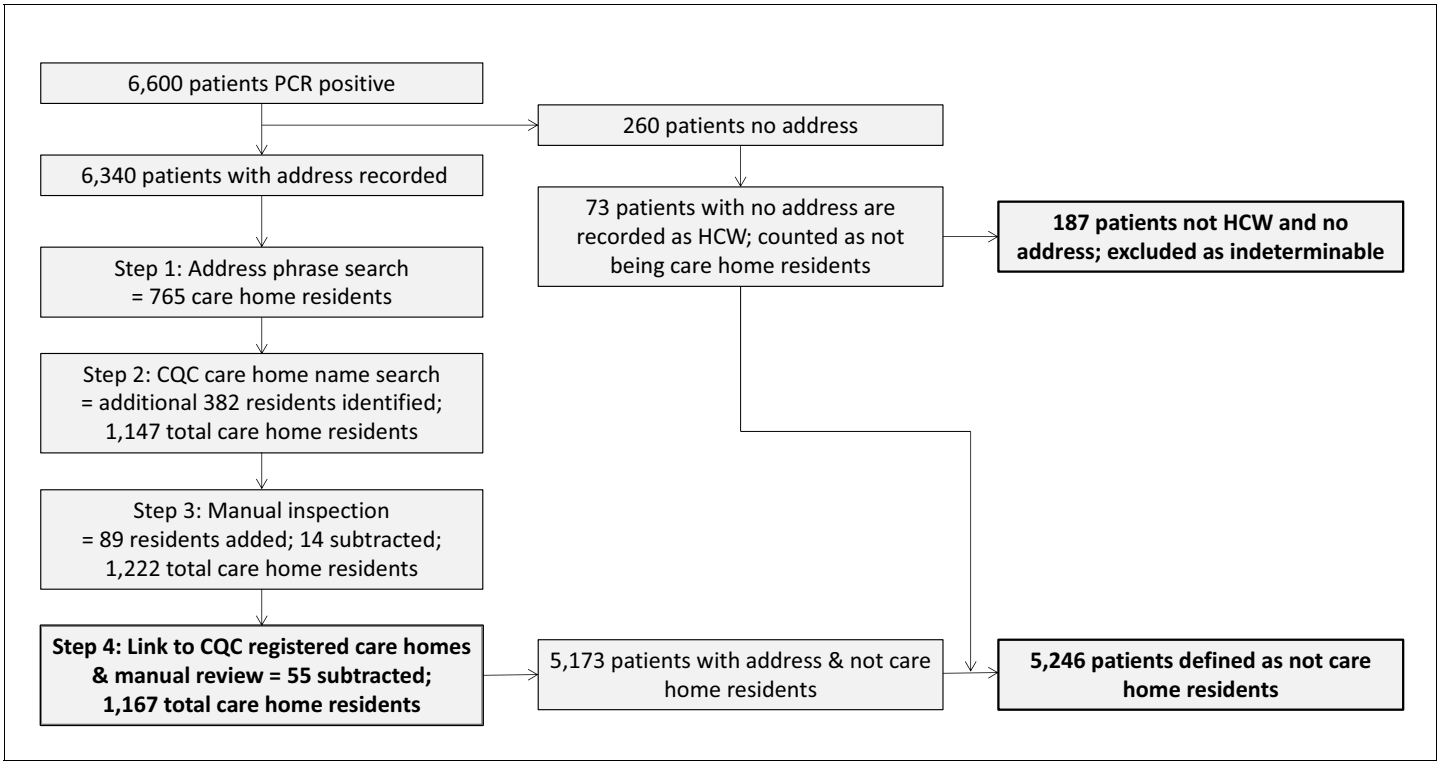


Figure 1—figure supplement 1. Flow diagram for identifying care homes from Cambridge-COGUK metadata Steps for identifying care home residents (further details in Materials and methods). First, the address field in the patient electronic healthcare records was searched for matching terms indicating a 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 1167 care home residents from 337 care homes; 5246 individuals that were not care home residents, and 187 individuals that were indeterminable.

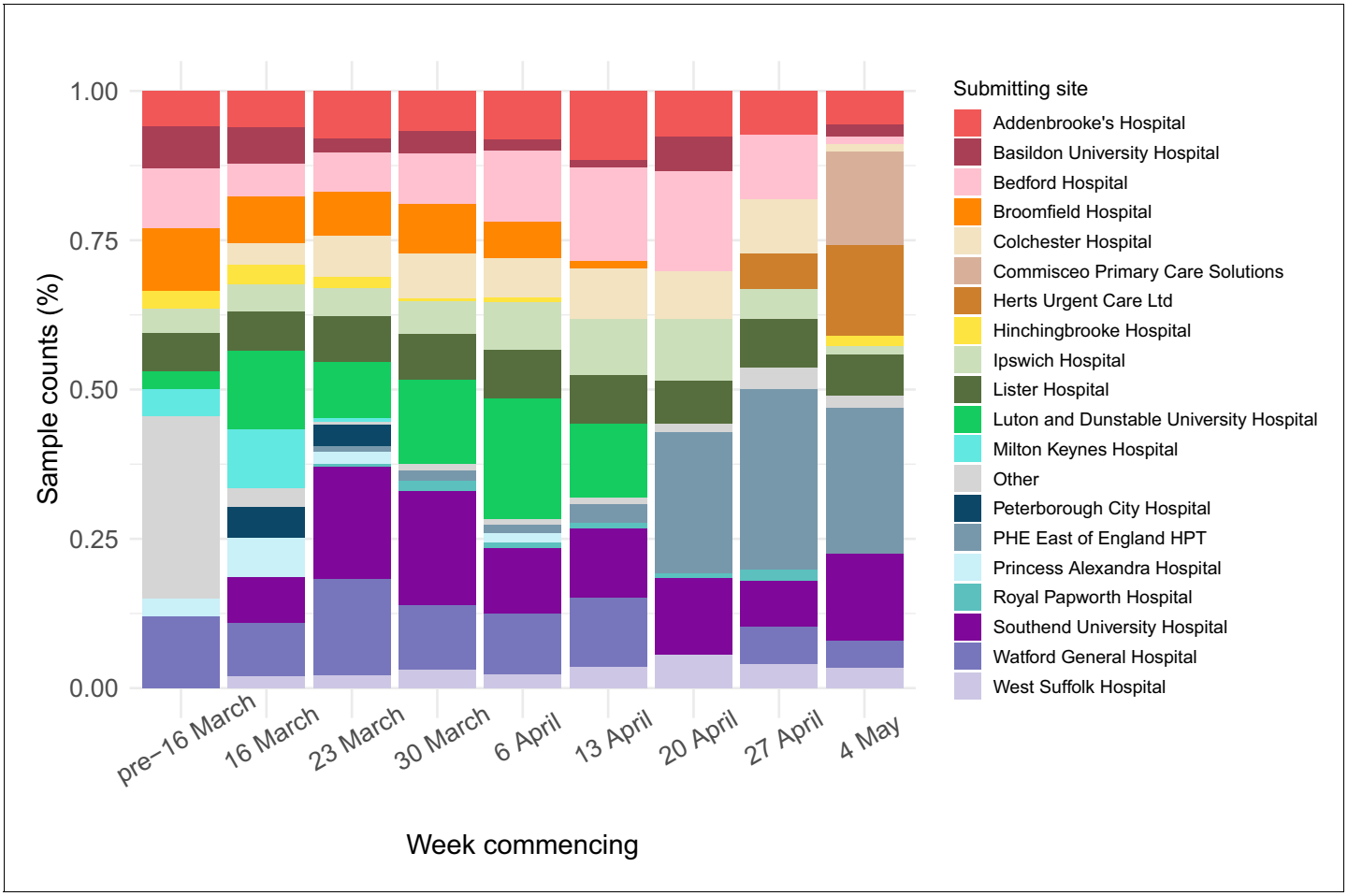


Figure 1—figure supplement 2. Breakdown of main organisations submitting samples to Cambridge PHE Laboratory over study period per week. Only showing sites that submitted samples from >50 people with positive test results over study period, otherwise counted as 'Other'. To maintain patient anonymity, 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.

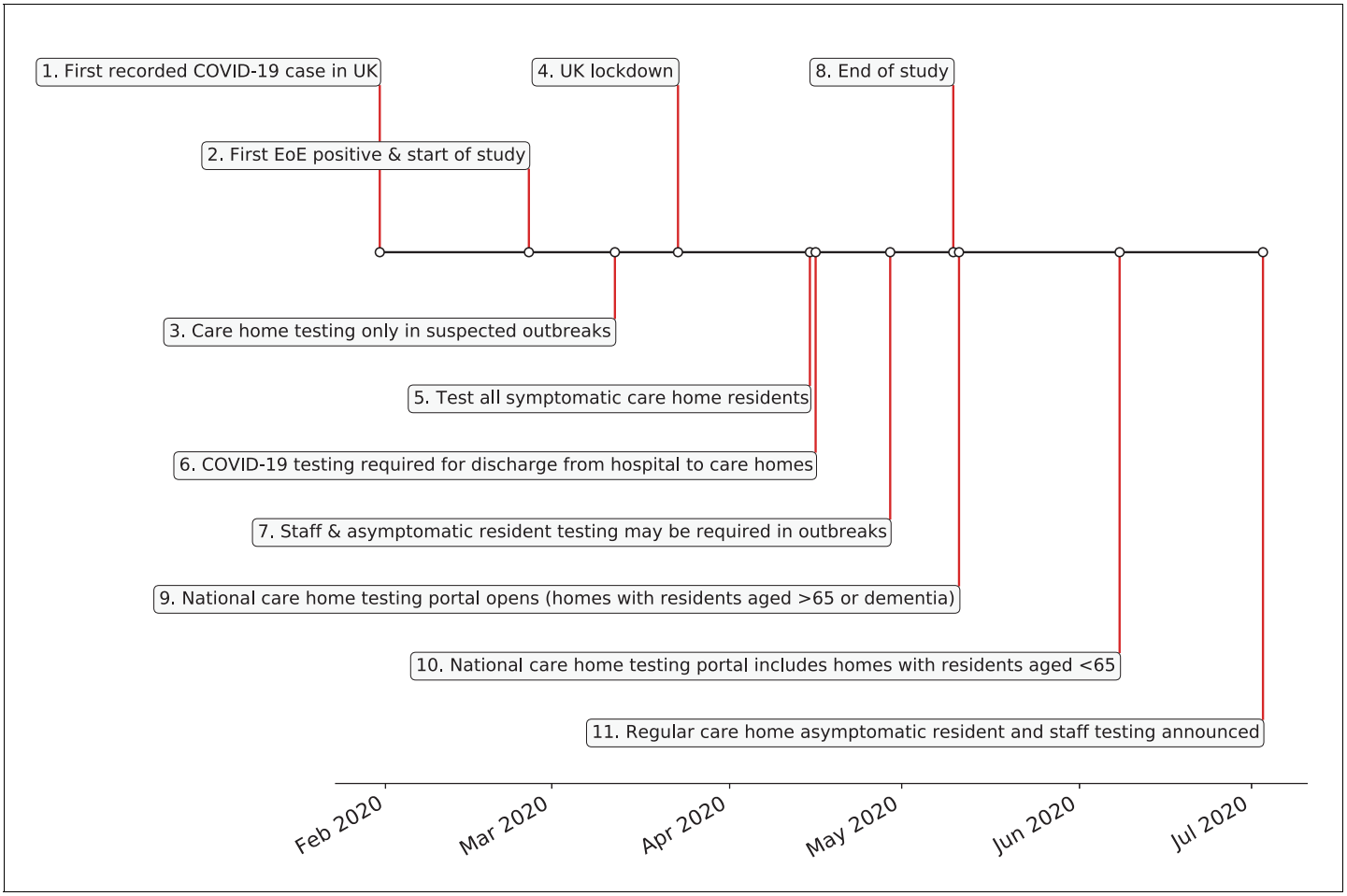


Figure 1—figure supplement 3. UK care home testing policy timeline. (1) 31st January – first recorded case of covid-19 in the UK. (2) 26th February - first case of COVID-19 in the East of England; start date of this study. (3) 12th March – individuals in the community advised to self-isolate for 7 days, 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: ***Public Health England, 2020b***; ***The Health Foundation, 2020***.

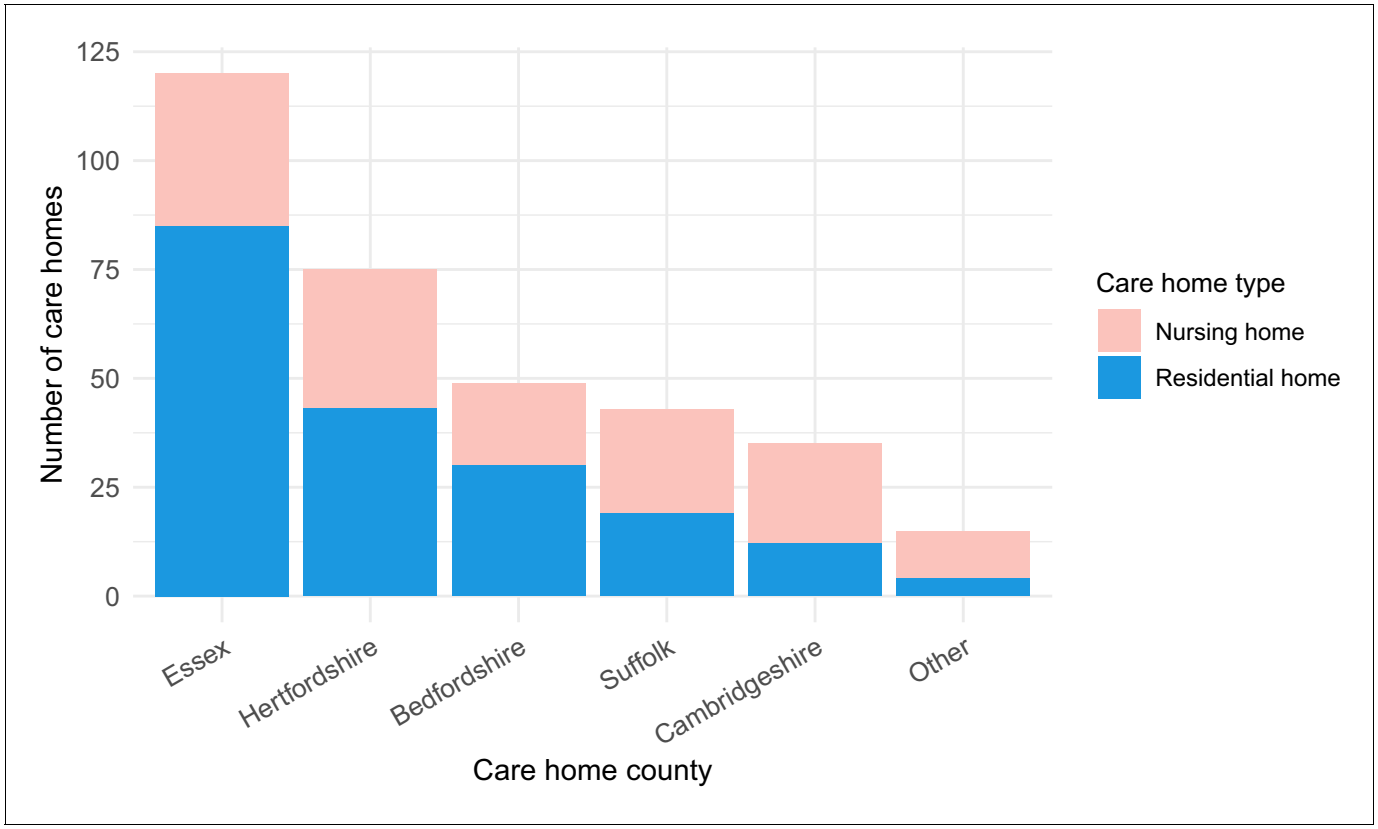


Figure 2. Care home locations by county, showing nursing, and residential homes. Only showing the five counties with the largest number of cases (all >25) to preserve patient anonymity. Definitions of 'nursing home' and 'residential home' are based on Care Quality Commission (CQC) information on whether 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.

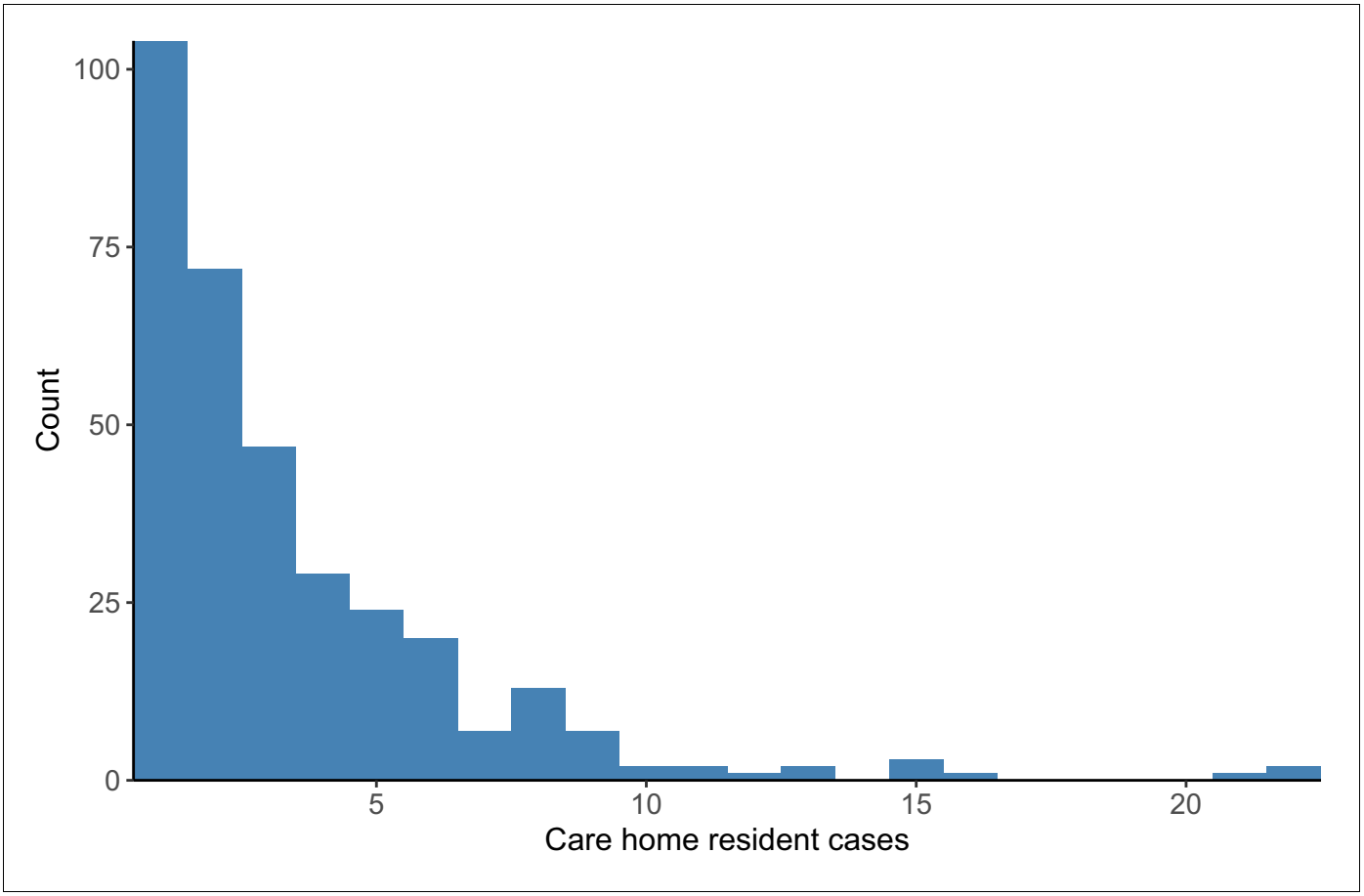


Figure 2—figure supplement 1. Distribution of cases per care home. The number of positive cases per care home was highly skewed, such that a relatively small number of care homes contributed a large proportion of cases (right-hand side of the plot). Plot produced with R package *ggplot2* using geom_histogram with binwidth = 1.

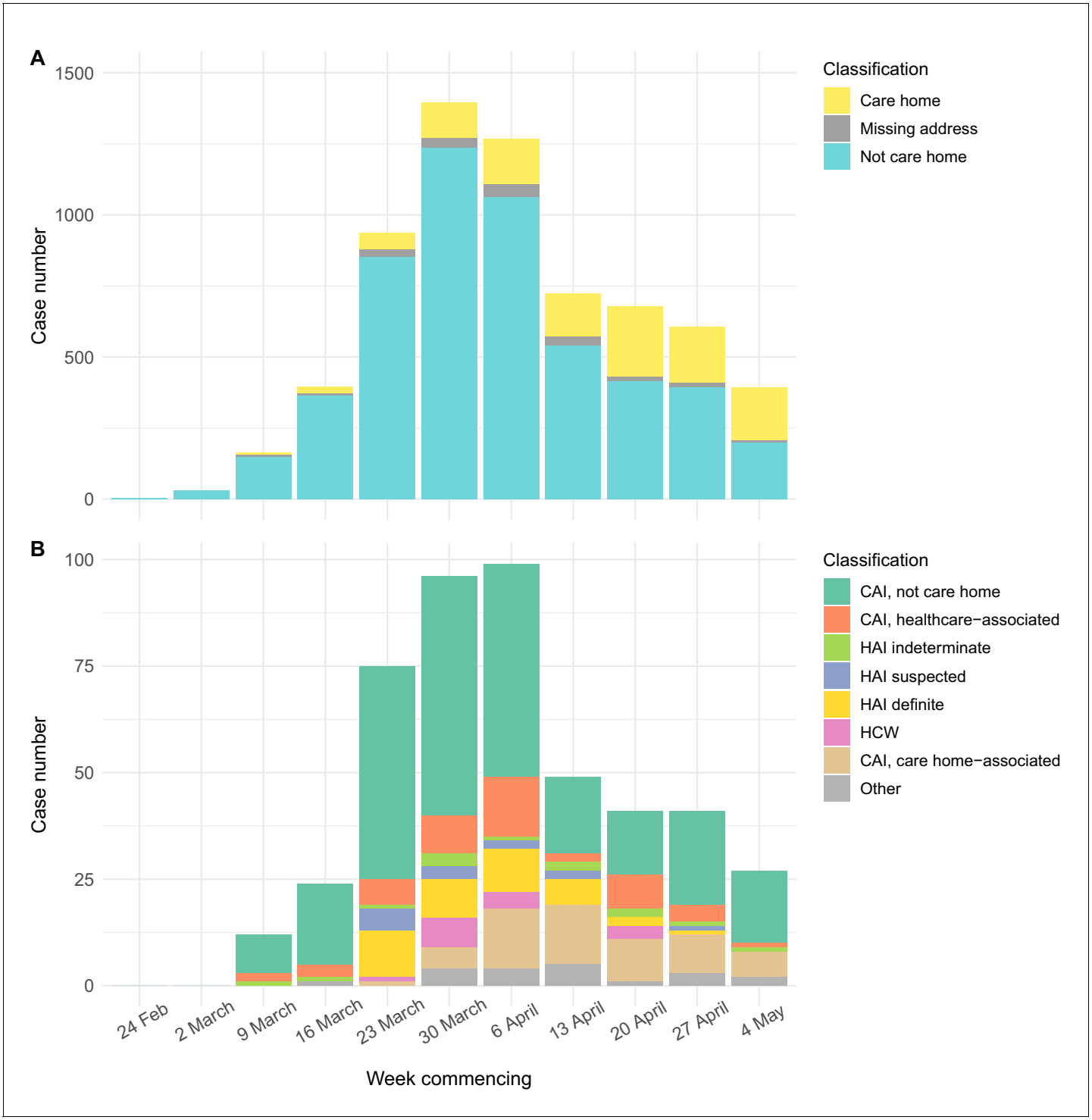


Figure 3. Epidemic curves for EoE and CUH showing care home residents. Number of positive cases per week over the study period for different infection sources, for all samples tested from EoE at the Cambridge PHE laboratory (**A**), or those tested at CUH acute medical services (**B**). Peak of the epidemic 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

Figure 3 continued on next page

Figure 3 continued

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***).

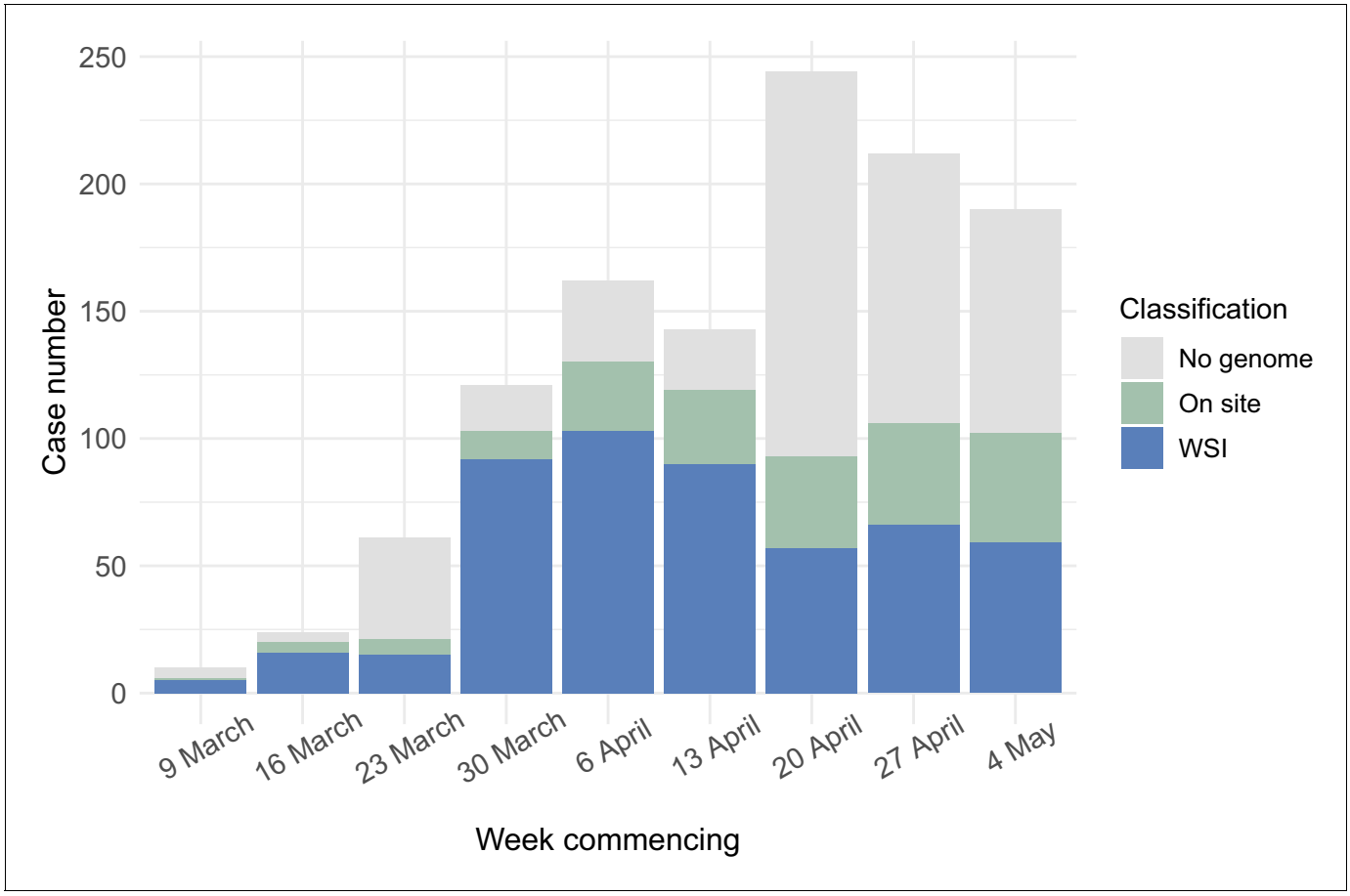


Figure 3—figure supplement 1. Care home residents per week showing genome sequencing site. Plot shows total care home residents testing positive per week over the study period, showing number of care home residents with genomes included in the study broken down by sequencing location (on site in the Department of Pathology, Division of Virology or at the Wellcome Sanger Institute).

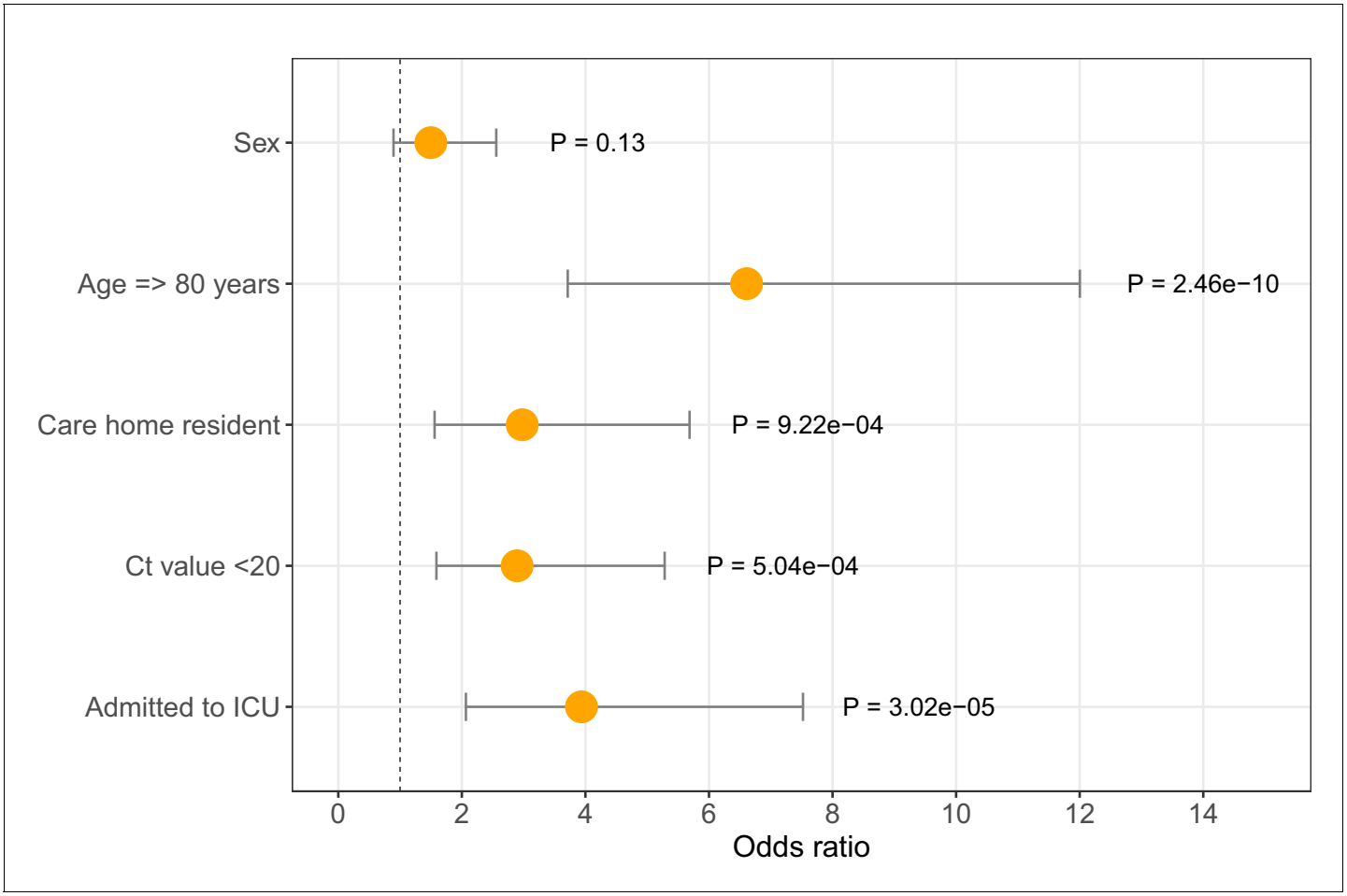


Figure 4. Odds ratios for mortality at 30 days. Logistic regression analysis showing odds of death at 30 days (with 95% confidence intervals) for five available metadata variables: patient sex, age (here categorised as ≥80 years), whether they were a care home resident, the diagnostic Ct value (here 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.

Figure 4—figure 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()*.

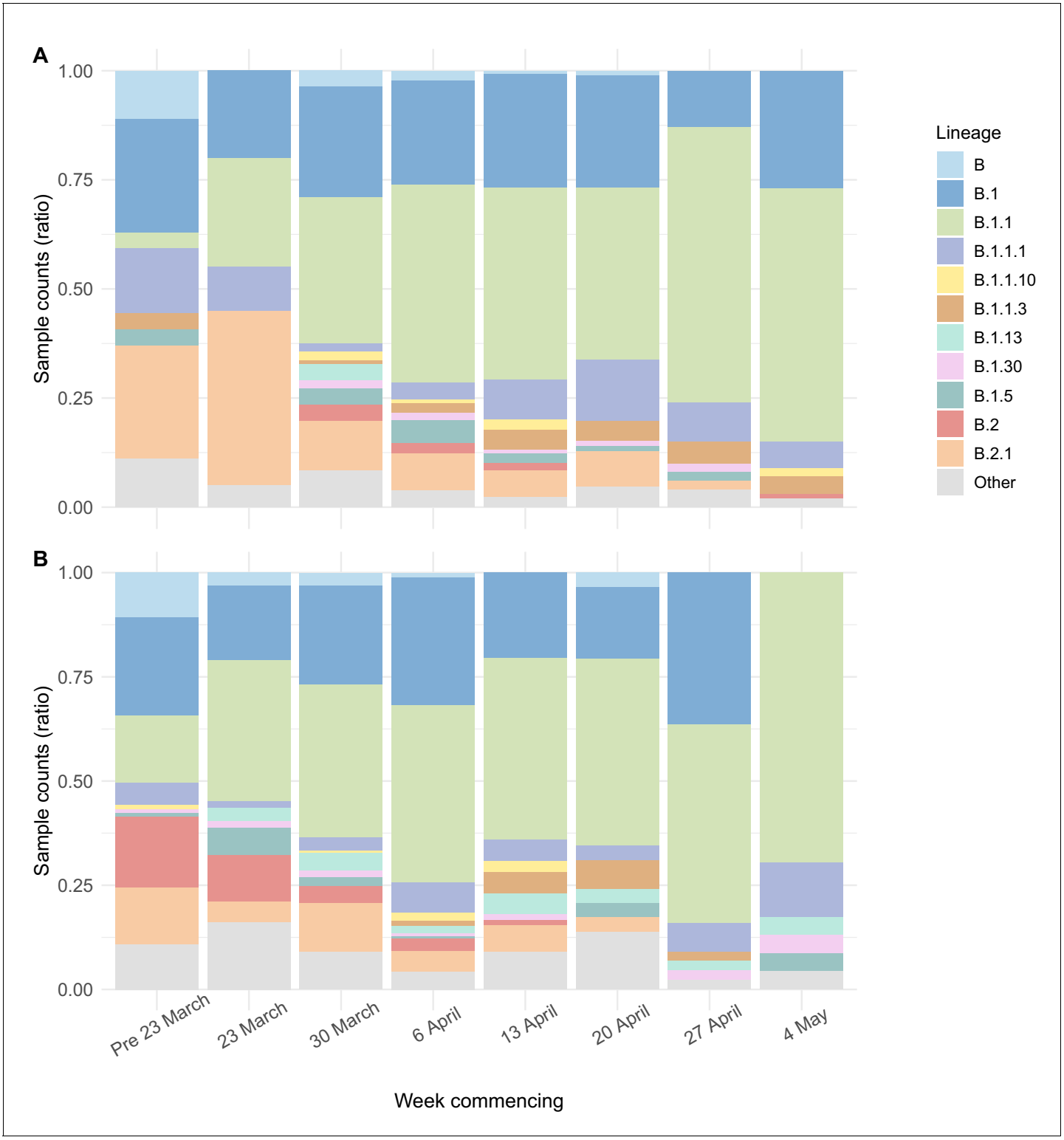


Figure 5. Viral lineage compositions in care home and non-care home samples. Plots showing the ratios of SARS-CoV-2 viral lineages for 700 care home resident genomes (**A**) and a randomly selected subset of 700 non-care home residents (**B**). The proportion of lineage B.1.1 increased over the study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

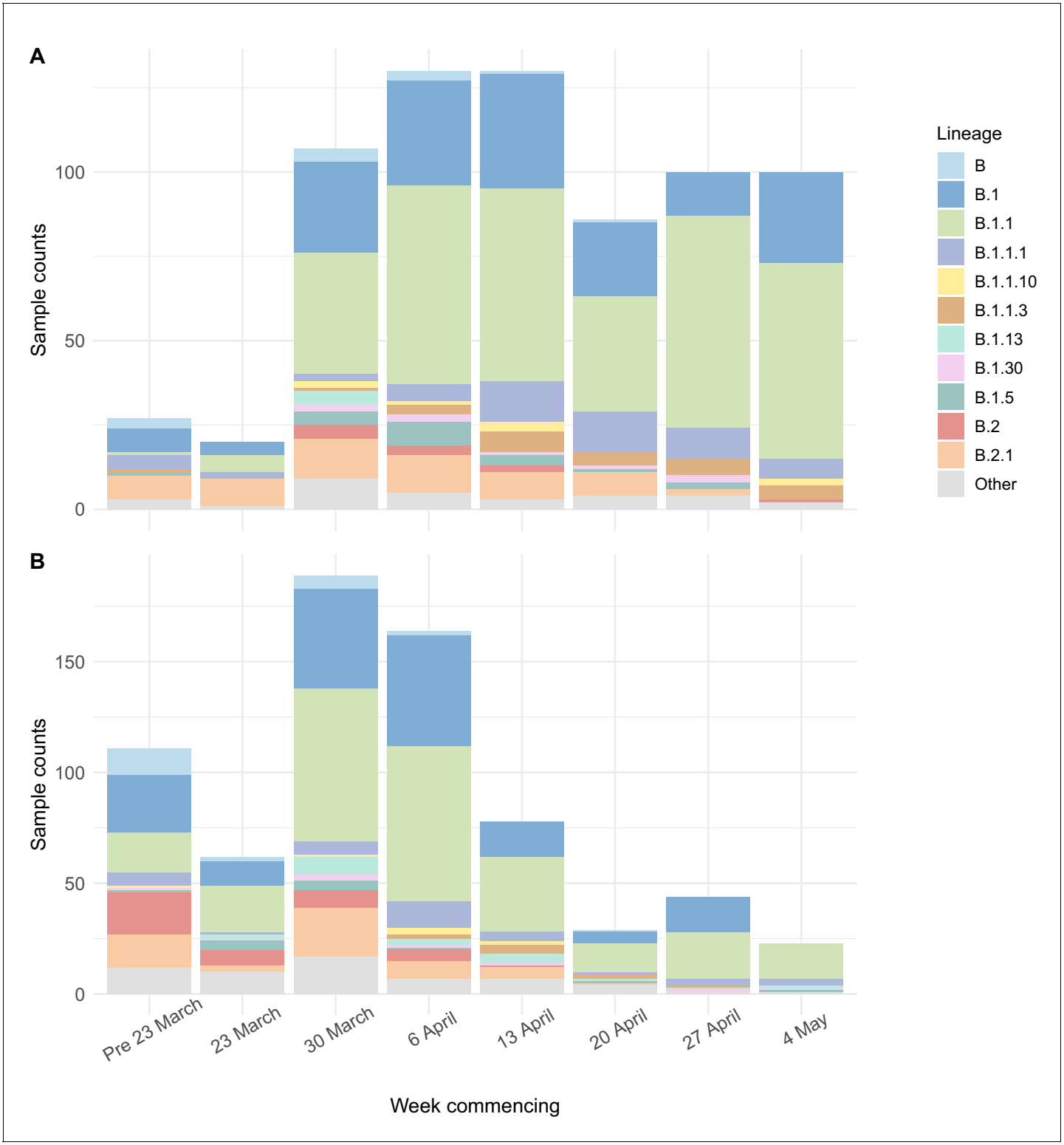


Figure 5—figure supplement 1. Viral lineage compositions in care home and non-care home samples by count. Plots showing the counts of SARS-CoV-2 viral lineages for 700 care home resident genomes (**A**) and a randomly selected subset of 700 non-care home residents (**B**). Lineages defined using *pangolin*. Data also presented in ***Table 5***.

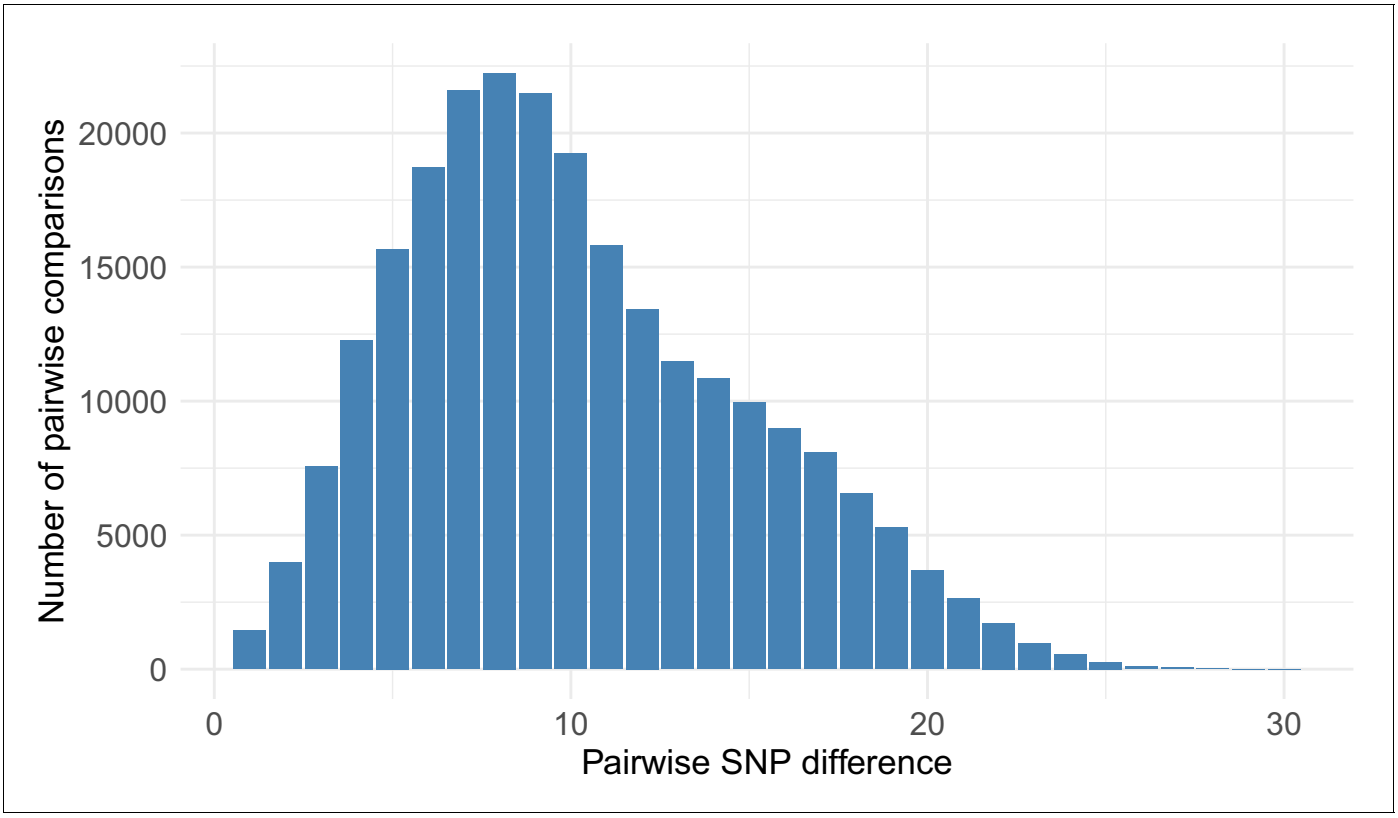


Figure 5—figure supplement 2. Distribution of pairwise SNP differences between care home samples. Pairwise SNP differences between the 700 care home residents (244,650 comparisons). There was a median of eight single nucleotide polymorphisms (SNPs) separating care home genomes (interquartile range, IQR 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).

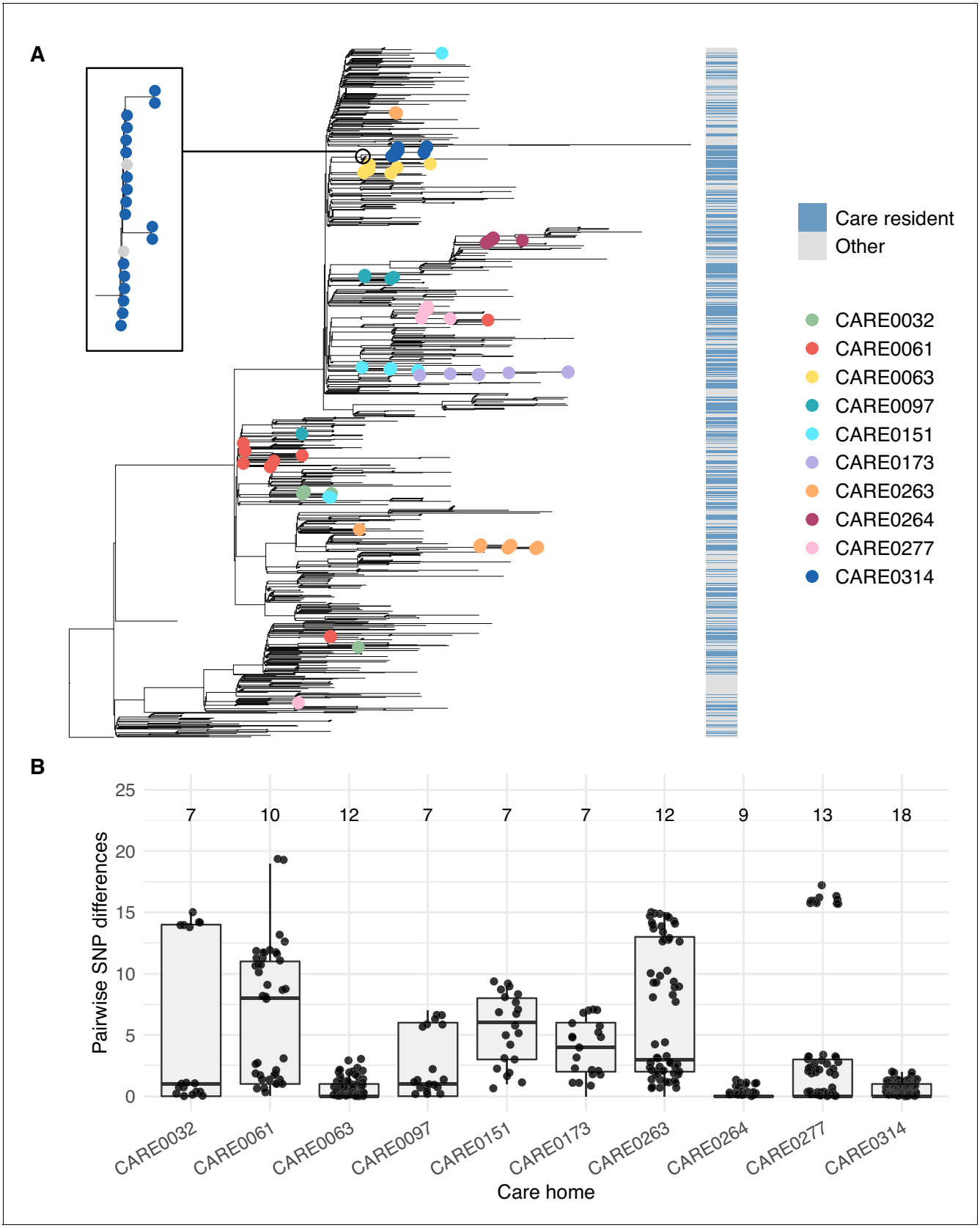


Figure 6. Care home clustering on viral phylogenetic tree and within-care home pairwise SNP differences. (**A**) Phylogenetic tree of 1400 East of England SARS-CoV-2 genomes rooted on a sample from Wuhan, China, collected December 2019, including 700 care home residents and 700

Figure 6 continued on next page

Figure 6 continued

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 10 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 10 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***.

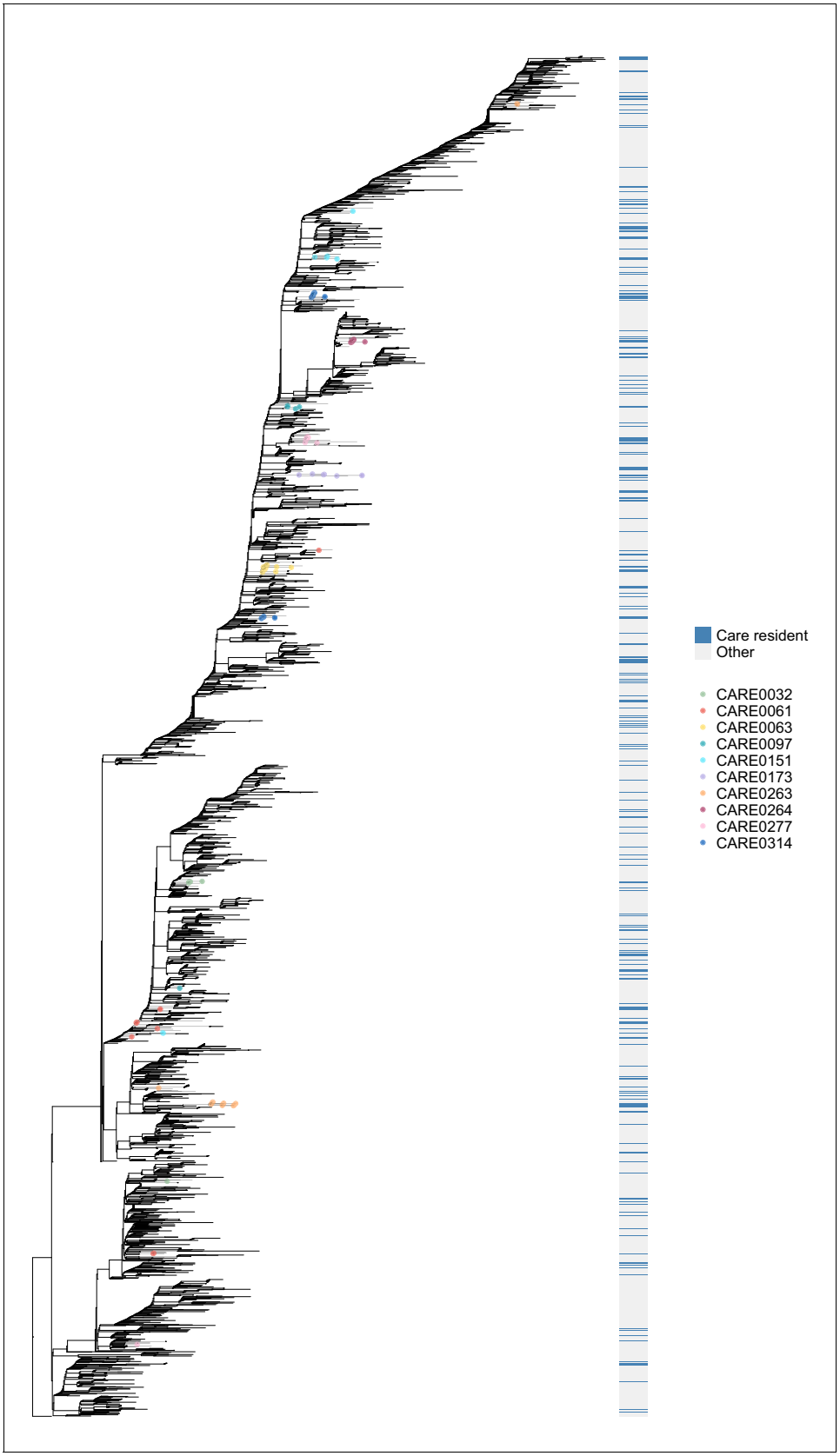


Figure 6—figure supplement 1. Phylogenetic tree of all available genomes highlighting care home and non-care home samples. Of the 6600 individuals in the study, 1167 were identified as care home residents and 5246 were not care home residents (187 were undetermined). 700/1167 (60.0%)

Figure 6—figure supplement 1 continued on next page

Figure 6—figure supplement 1 continued

care home residents had genomes available that passed quality control (QC) filtering at time of analysis. Of 5246, 3745 (71.4%) non-care home residents had genomes available and passing the same QC filtering at time of analysis, accessed from the COG-UK public database (https://www.cogconsortium.uk/data/). This tree comprises all 700 care home and 3745 non-care home genomes from the study (total 4445 samples), rooted on a 2019 genome from Wuhan, China. As with ***Figure 6***, 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. This supports the findings shown in ***Figure 6*** using the randomly selected sub-sample of non-care home samples, (1) that care home genomes were phylogenetically intermixed with non-care home genomes (consistent with transmission between care homes and outside of care homes) and (2) that, using the 10 care homes with the largest number of samples as examples, some care homes were monophyletic (such as CARE0314) while others were polyphyletic (such as CARE0061). Even for polyphyletic care homes (implying multiple independent introductions of the virus among residents), the majority of samples were usually attributable to a single dominant cluster (described further in main text).

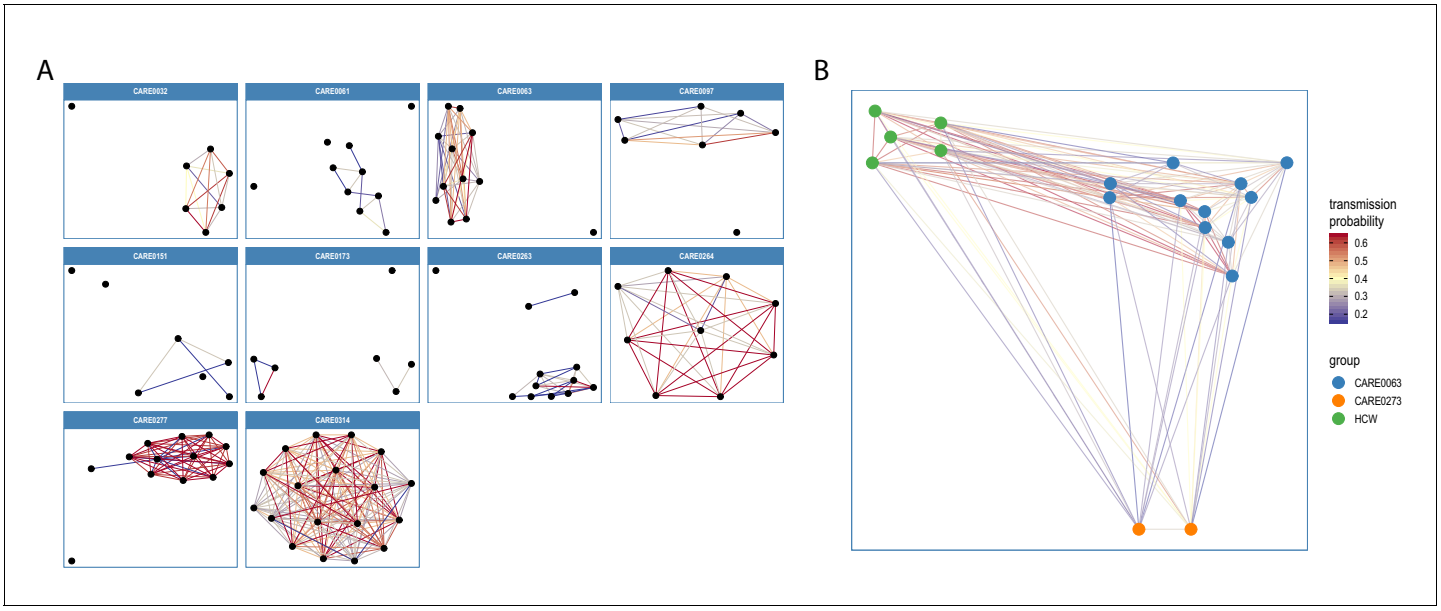


Figure 7. Visualisations of SARS-CoV-2 clusters among care home residents. Transmission networks were produced using a derivative of the *transcluster* algorithm, which incorporates pairwise date and genetic differences to estimate the probability of cases being connected within a defined number of intermediate hosts. Clusters were defined using a probability threshold of ≥15% for cases being linked by ≤2 intermediate hosts (further details in Materials and 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 10 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.

Figure 7—figure supplement 1. Transmission network diagrams for all care homes with two 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

Figure 7—figure supplement 1 continued on next page

Figure 7—figure supplement 1 continued

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 Materials and methods). This figure displays data from all care homes with ≥2 samples with genomic data.

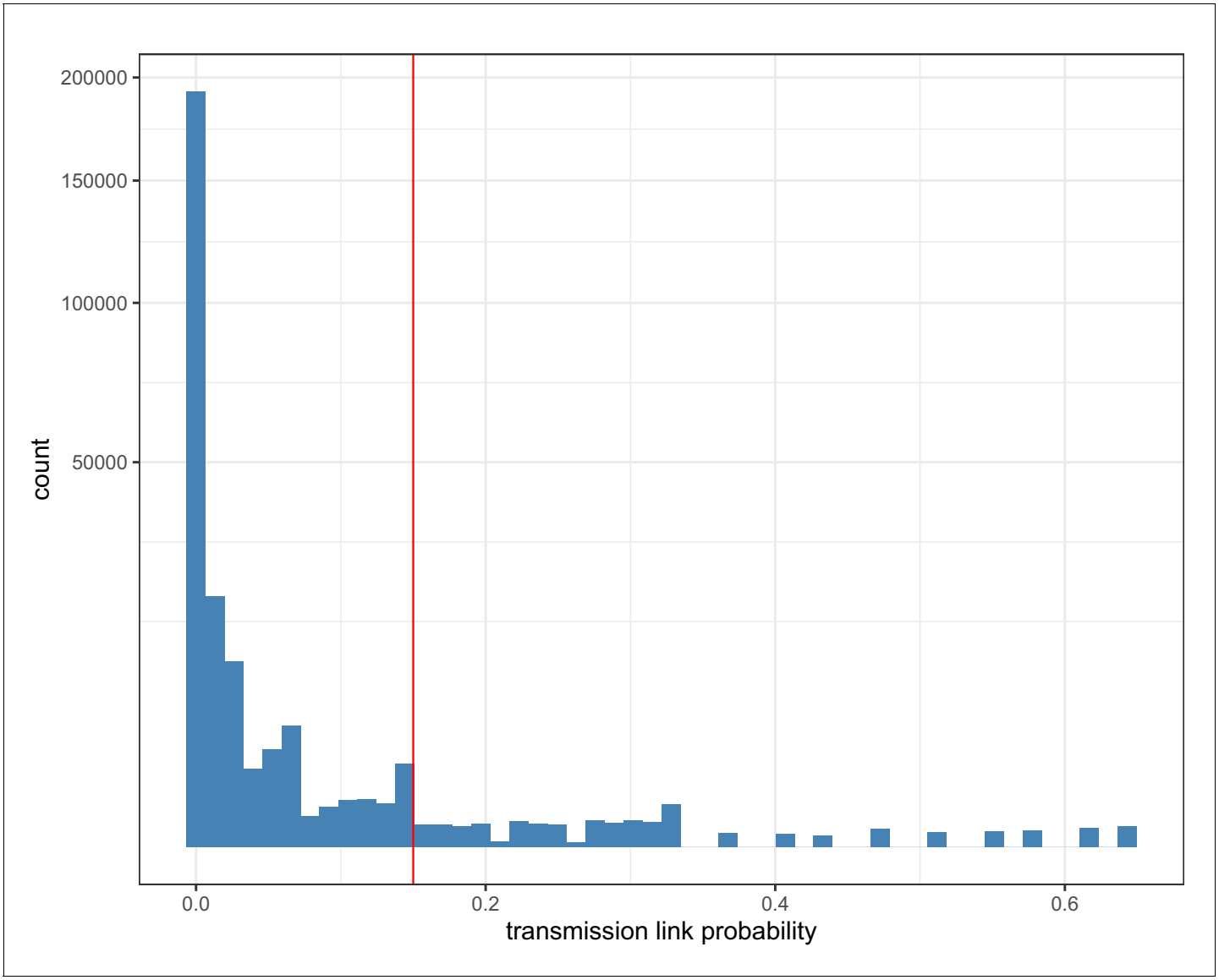


Figure 7—figure supplement 2. Histogram of pairwise transmission probabilities between care home samples. Histogram of the pairwise probabilities for cases being connected by ≤2 intermediate hosts for all 700 care home residents as inferred by the *transcluster* algorithm, with vertical red line at 0.15 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.

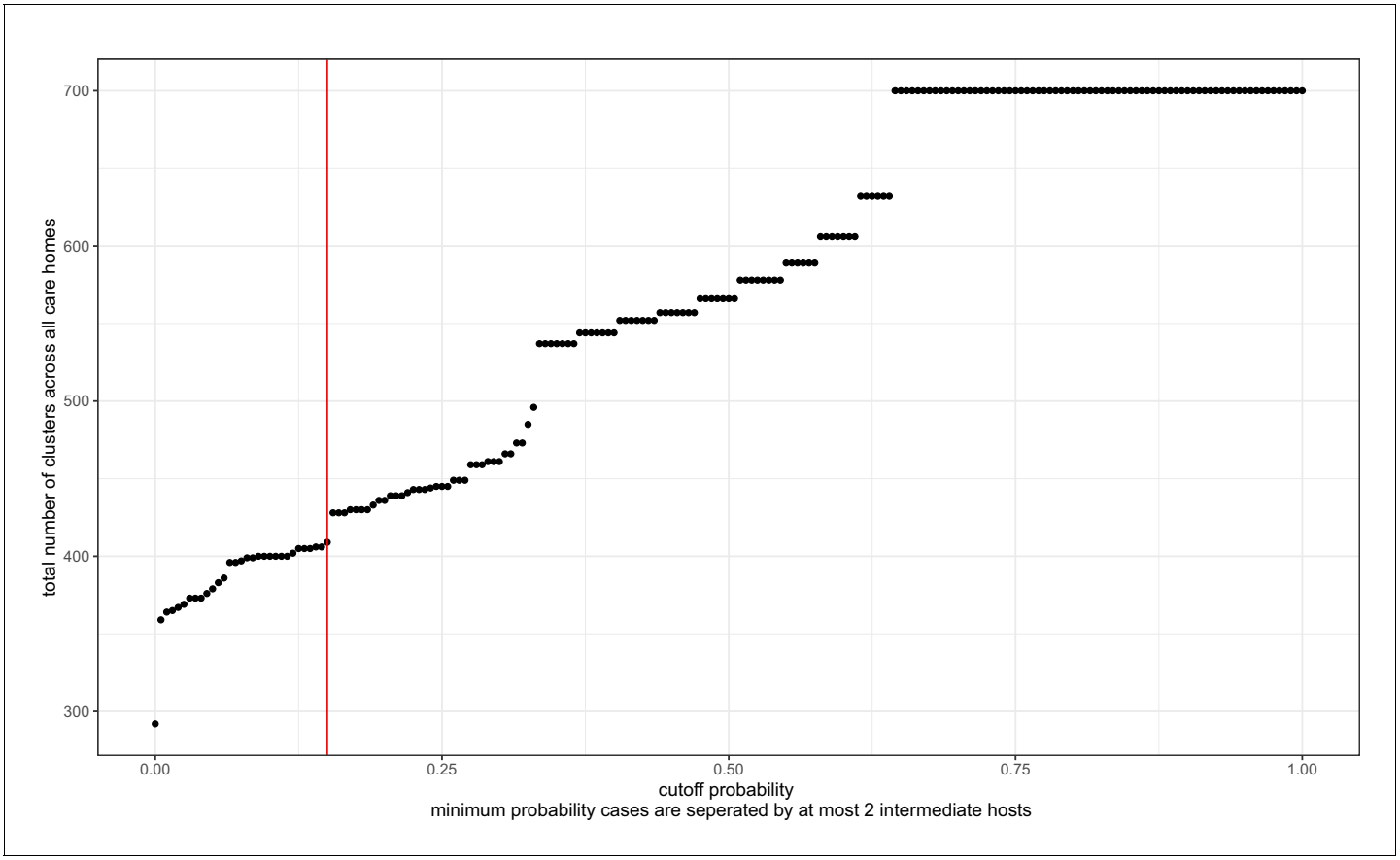


Figure 7—figure supplement 3. Transmission probability threshold vs number of care home clusters. The *transcluster* algorithm computes the likelihood of two samples being linked within a given number of intermediate hosts, based on the date and genetic differences between samples (assuming a given serial interval and mutation rate, further details in Materials and 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—figure supplements 4*** and ***5***) and because of a 'jump' in the number of clusters occurring at that point.

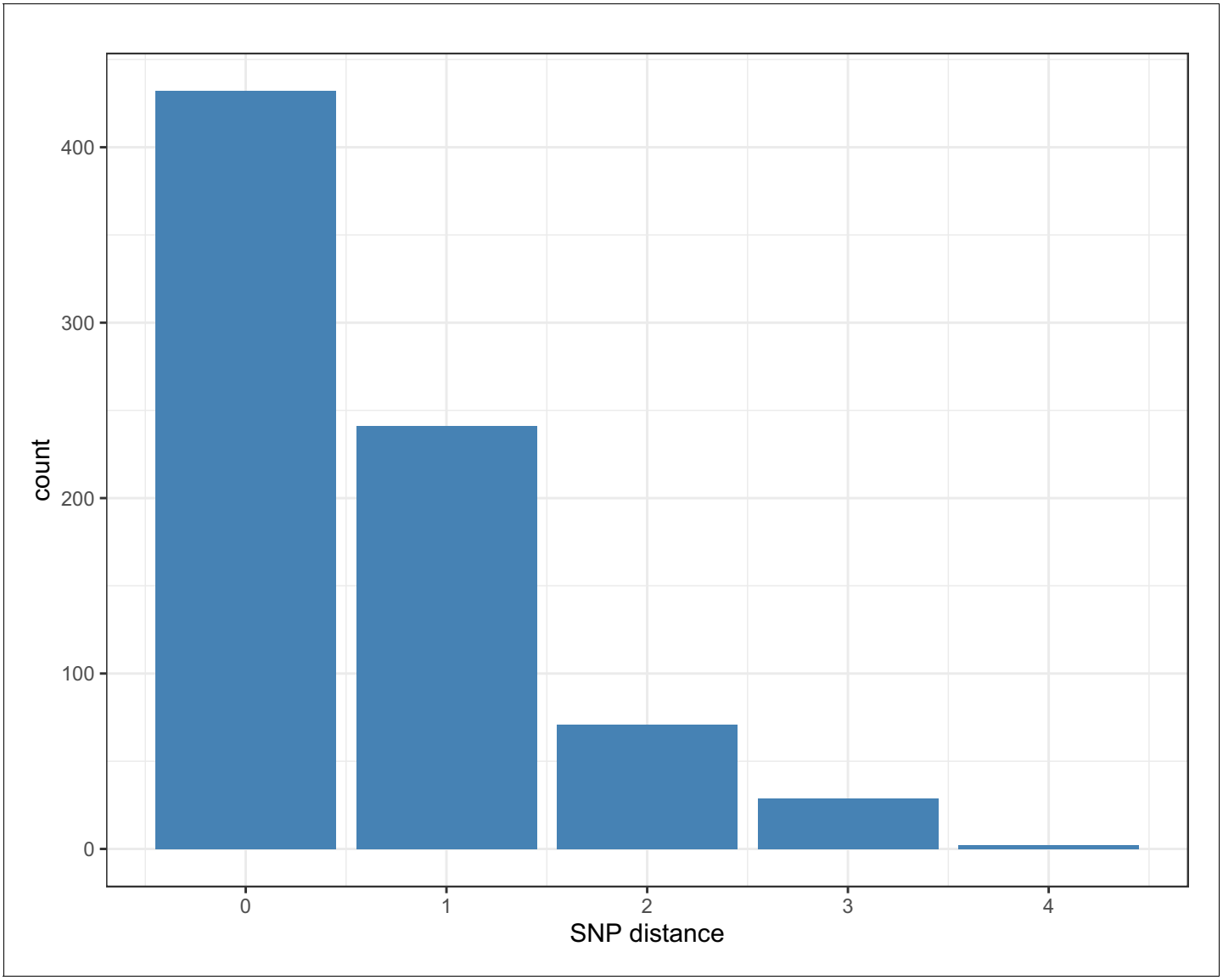


Figure 7—figure supplement 4. Pairwise SNP difference distribution between samples within clusters. Within each cluster, 673/775 (86.8%) of pairwise links that had a ≥15% probability of transmission with ≤2 intermediate hosts had 0 or one pairwise SNP differences (maximum 4).

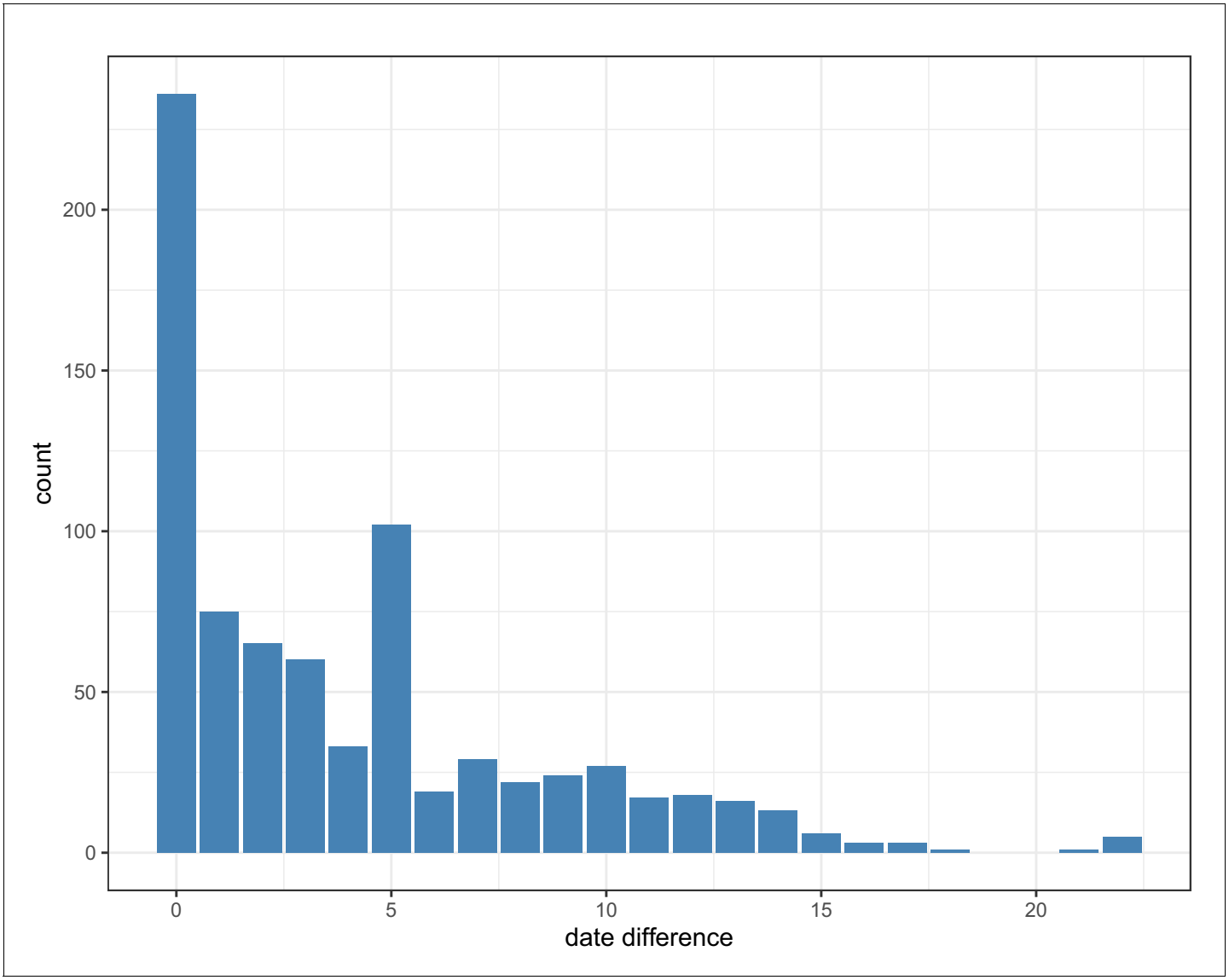


Figure 7—figure supplement 5. Pairwise date difference distribution between samples within clusters, aggregated across dataset. Within each cluster, 756/775 (97.5%) of pairwise links that had a $\geq$15% probability of transmission with $\leq$2 intermediate hosts cases were sampled <14 days apart (maximum 22 days).

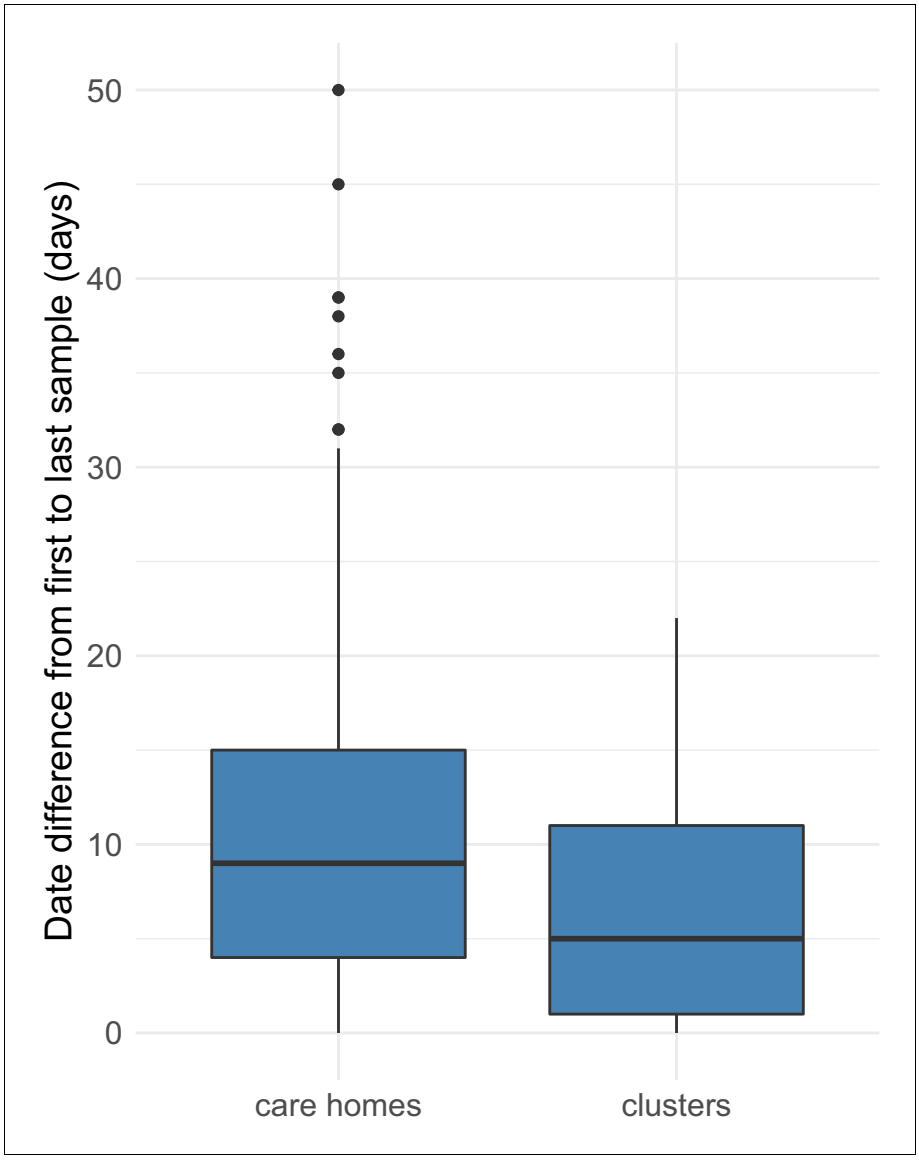


Figure 7—figure supplement 6. Distributions of date ranges (from first to last sampling dates) for care homes vs clusters. Date ranges were calculated by subtracting the date of the first sample from the last sample for each care home (left) or cluster (right). Care homes and clusters were only included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

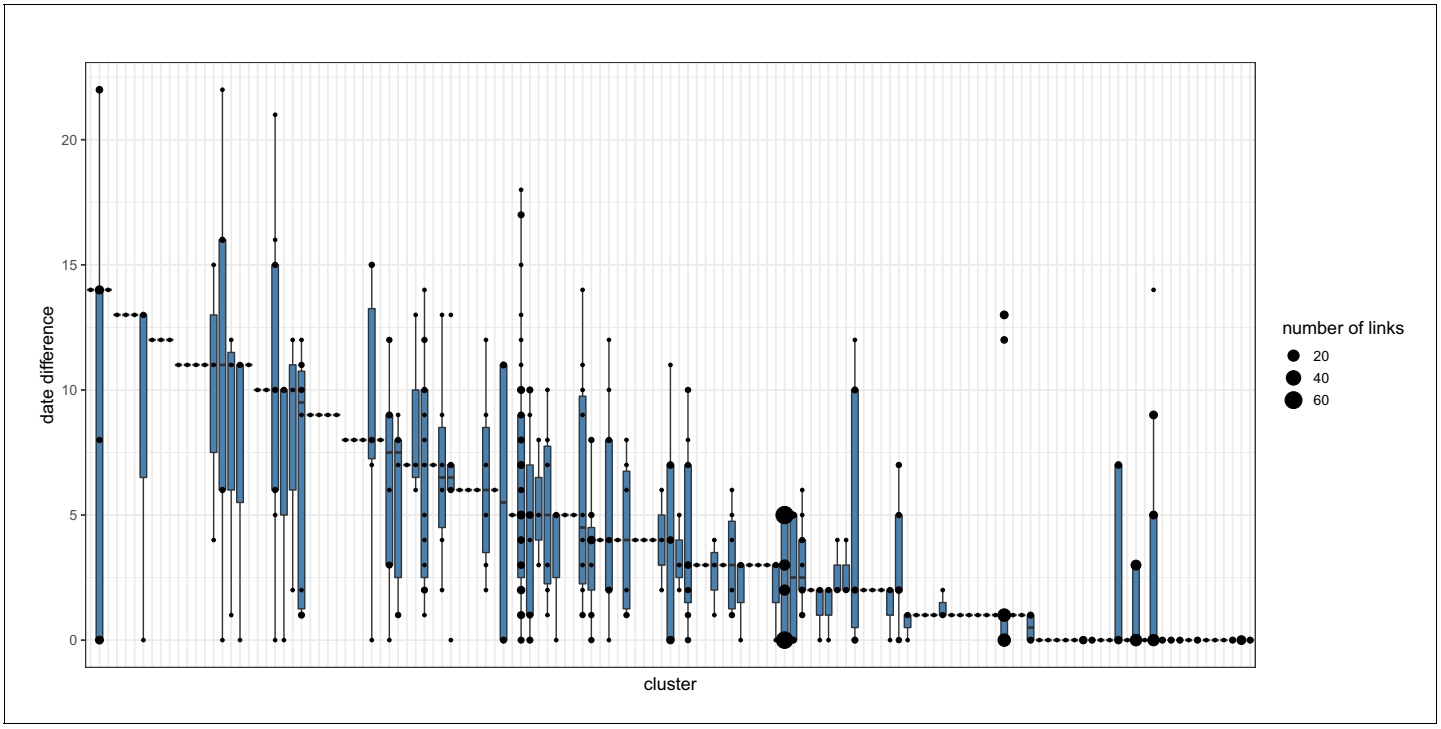
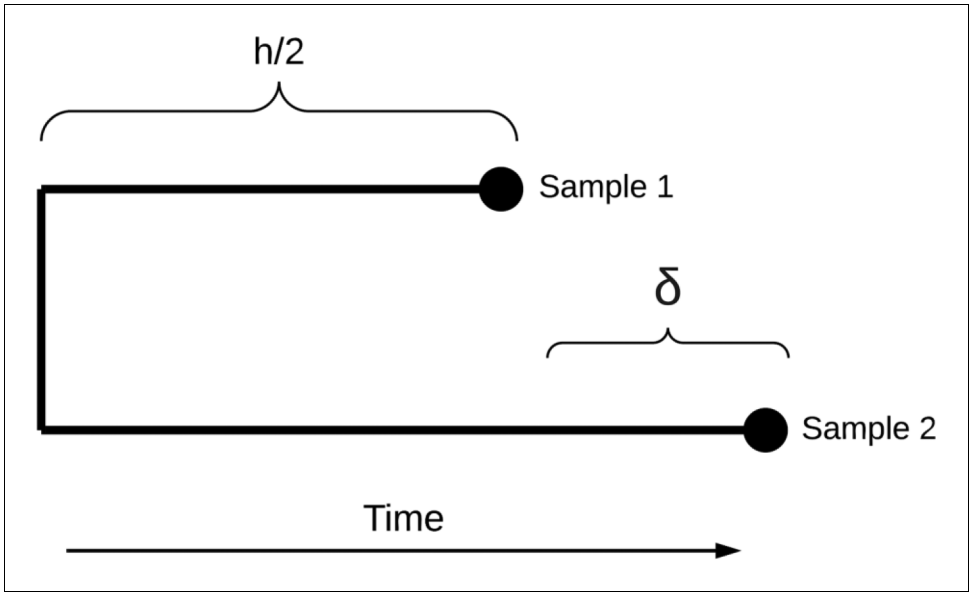


Figure 7—figure supplement 7. Pairwise date difference distribution between samples within each cluster. Boxplots indicate the median and interquartile ranges for the number of days separating samples found to be within the same transmission cluster by the *transcluster* algorithm. The boxplots are overlaid 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.



Scheme 1. Diagram representing transmission dynamics between two samples.