The role of genomics in understanding COVID-19 outbreaks in long term care facilities

What has genomics told us about drivers of care home outbreaks during the first wave?

Dinesh Aggarwal[1,2,7,17], Richard Myers[2], William L Hamilton[1,7], Tehmina Bharucha[2,5,6], Niamh M. Tumelty[4], Colin S. Brown[2,5,6], Emma J. Meader[3], Tom Connor[8,9,10], Darren L. Smith[11], Declan T. Bradley[12,13], Samuel Robson[14], Matthew Bashton[11], Laura Shallcross[15], Maria Zambon[2], Ian Goodfellow[16], The COVID-19 Genomics UK (COG-UK) Consortium[19,#], Meera Chand[2,18], Justin O'Grady[10], M. Estée Török[1,7], Sharon J. Peacock[1,17], Andrew J. Page[10,*]

Affiliations

1 University of Cambridge, Department of Medicine, Cambridge, UK.
2 Public Health England, 61 Colindale Ave, London, NW9 5EQ, UK.
3 Norfolk and Norwich University Hospital, Colney Lane, Norwich, NR4 7UY, UK.
4 University of Cambridge, Cambridge University Libraries, Cambridge, UK
5 Oxford Glycobiology Institute, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3RQ, United Kingdom.
6 Lao-Oxford-Mahosot Hospital-Wellcome Trust Research Unit, Microbiology Laboratory, Mahosot Hospital, Vientiane, Lao PDR.
7 Cambridge University Hospital NHS Foundation Trust, Cambridge, UK.
8 Organisms and Environment Division, School of Biosciences, Cardiff University, Cardiff, Wales, UK.
9 Public Health Wales, University Hospital of Wales, Cardiff, UK.
10 Quadram Institute Bioscience, Norwich Research Park, Norwich, NR4 7UQ, UK
11 Hub for Biotechnology in the Built Environment, Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle upon Tyne, NE1 8ST, UK.
12 Public Health Agency, Belfast, BT2 8BS, Northern Ireland, UK.
13 Centre for Public Health, Queen's University Belfast, Belfast, BT12 6BA, Northern Ireland, UK.
14 University of Portsmouth, Centre for Enzyme Innovation, Portsmouth, PO1 2DT, UK.
15 Institute of Health Informatics, University College London, UK.
16 University of Cambridge, Department of Pathology, Cambridge, UK.
17 Wellcome Sanger Institute, Hinxton, Cambridge, UK.
18 Guy's and St Thomas' NHS Foundation Trust, London, UK.
19 https://www.cogconsortium.uk
[#] Full list of consortium names and affiliations are in the appendix
*corresponding author: andrew.page@quadram.ac.uk

Abstract

A review was undertaken of all genomic epidemiology studies on COVID-19 in long term care facilities (LTCF) that have been published to date. It was found that staff and residents were usually infected with identical, or near identical, SARS-CoV-2 genomes. Outbreaks usually involved one predominant lineage, and the same lineages persisted in LTCFs despite infection control measures. Outbreaks were most commonly due to single or few introductions followed by spread rather than a series of seeding events from the community into LTCFs. Sequencing of samples taken consecutively from the same cases showed persistence of the same genome sequence indicating that the sequencing technique was robust over time. When combined with local epidemiology, genomics facilitated likely transmission sources to be better characterised. Transmission between LTCFs was detected in multiple studies. The mortality rate amongst residents was high in all cases, regardless of the lineage. Bioinformatics methods were inadequate in one third of the studies reviewed, and reproducing the analyses was difficult as sequencing data were not available in many cases.

Introduction

A number of studies of COVID-19 in long term care facilities (LTCFs) have reported high mortality (Abrams et al. 2020; Fisman et al. 2020; Burton et al. 2020). Possible explanations for this include recognised risk factors such as increased age and co-morbidities (Jordan, Adab, and Cheng 2020; Fisman et al. 2020). In England and Wales, it has been estimated that nearly 30% (15,819, week ending 16 October 2020) of all deaths due to COVID-19 occurred in LTCFs (ONS 2020) with outbreaks reported in 45% of all LTCFs (Public Health England 2020). Northern Ireland reported even higher rates; 37% (363, week ending 23 October 2020) of deaths in LTCFs were due to COVID-19. Globally, 24.9% of super-spreading events (Al-Tawfiq and Rodriguez-Morales 2020) were linked to LTCFs (Swinkels 2020). The drivers for introduction and transmission of SARS-CoV-2 in the care sector are under investigation and remain incompletely understood.

Genome sequencing of the SARS-CoV-2 virus allows for the underlying genome to be reconstructed, either through *de novo* assembly, or as a reference guided consensus sequence (Tom Connor 2020). These genome sequences can be clustered together based on their similarity. A nomenclature was developed (Rambaut et al. 2020) to label clusters into distinct lineages. This can provide information on the genetic relatedness and enable investigation of the dynamics of outbreaks within and between LTCFs and the wider community. Here, we review the genomic epidemiology studies on COVID-19 in LTCFs that have been conducted to date and provide a summary and interpretation of the key findings (Table 1).

Table 1: Summary of the key findings of this review.	**Evidence**
Community or Hospital acquisition of SARS-CoV-2 in LTCF residents	
1. The majority of LTCF infections were community-acquired ***(moderate)***	Hamilton et al. 2020; Page et al. 2020
2. Approximately 6% of infected LTCF residents had suspected or confirmed hospital-acquired infections in one UK region ***(moderate)***.	Hamilton et al. 2020
3. Staff and residents at the same LTCF were usually infected with the same lineage, including staff with no direct contact with residents ***(moderate)***.	Dautzenberg et al. 2020; Page et al. 2020; Lemieux et al. 2020; Taylor et al. 2020
4. Limited genomic diversity amongst cases from the same LTCFs, indicated a small number of introductions rather than a series of seeding events from the community ***(weak)***.	Dautzenberg et al. 2020; Page et al. 2020; Lemieux et al. 2020; Ladhani et al. 2020
5. Shared clusters between separate LTCFs could be identified ***(moderate)***.	Hamilton et al. 2020; Page et al. 2020; Ladhani et al, 2020
Transmission and outcomes within LTCF	
6. Use of genomic data allowed independent clusters of infections to be identified within LTCFs **(strong)**	Hamilton et al. 2020; Besselaar et al. 2020; Graham et al. 2020; Ladhani et al. 2020; Lemieux et al. 2020; Taylor et al. 2020; (Arons et al. 2020
7. In LTCF outbreaks initial sequencing was useful to identify whether genomes were similar, but subsequent	Lemieux et al. 2020; Taylor et al. 2020

sequencing of large numbers of samples did not add much value ***(moderate)***.	
8. Once two symptomatic individuals were identified in a LTCF, the outbreak was already widespread ***(moderate)***.	Arons et al. 2020
9. Sequencing of samples taken consecutively from the same LTCF residents showed that viral lineages persisted over an extended period of time despite IPC measures. It also showed that the sequencing technique was reproducible ***(moderate)***.	Taylor et al. 2020
10. One lineage usually predominated with the majority of samples being identical or near identical (0 to 1 SNP differences) ***(moderate)***.	Lemieux et al. 2020; Page et al. 2020
11. Within an LTCF residents were more likely to be infected with identical genome sequences if their bedrooms were in close proximity ***(moderate)***.	Arons et al. 2020
12. Mortality rate amongst LTCF residents was high in all cases, with no link to particular lineages ***(moderate)***.	Graham et al. 2020; Ladhani et al. 2020; Taylor et al. 2020
13. Temporal analysis of genomic data allows for the estimation of when an introduction was likely to have occurred ***(moderate)***.	Lemieux et al. 2020
Application of Genomics	
14. The genomic studies reviewed commonly misapplied bioinformatics methods ***(strong).***	Ladhani et al. 2020; Quicke et al. 2020; Graham et al. 2020
15. Minimum quality thresholds set by public archives on SARS-CoV-2 data limit data availability and reproducibility ***(moderate).***	Page et al. 2020; Hamilton et al. 2020; Ladhani et al. 2020;
16. Most studies did not provide adequate epidemiological or metadata to allow analysis to be reproduced ***(strong).***	Dautzenberg et al. 2020; Besselaar et al. 2020; Graham et al. 2020; Ladhani et al. 2020; Quicke et al. 2020; Taylor et al. 2020; Hamilton et al. 2020

* Strength of findings: **strong**, multiple sources of evidence, supported by indepth analysis or experiments; **moderate**, 1 or more sources of evidence which are supported by analysis or experiments; **weak**, 1 or more sources of evidence which are potentially contradictory.

Search strategy and selection criteria

The studies included in this review were identified by searches of PubMed and MedRxiv (last accessed 3 November 2020). The search terms ‘COVID-19 outbreak’ **or** ‘SARS-CoV-2 outbreak’, **and** 'long term care facility', 'care home', 'skilled nursing facility, 'nursing home' **or** 'residential home', **and** 'genomics', 'genome' **or** 'WGS' were used to identify relevant English-language publications and preprints since January 2020. A focus was placed particularly on studies where genomic epidemiology was used to enhance interpretation of outbreaks. As this is a rapidly emerging field of research, preprints were included but it must be noted that these are not peer reviewed. Study characteristics are summarised in Table 2.

Table 2: Overview of studies using SARS-CoV-2 genome sequencing of samples taken for routine surveillance or during investigation of outbreaks in LTCFs.

Study	**Location**	**Period (2020)**	**Type**	**No. of LTCFs**	**Total number of residents and staff tested**	**No of residents testing positive**	**Number of staff testing positive**	**Cases sequenced**	**No. lineages**
(Dautzenberg et al. 2020)	South East Netherlands	March - April	Surveillance	2	621	NA	133	22	3
(Besselaar et al. 2020)	South Holland	May - June	Outbreak	1	425	113	56	60[d]	1
(Hamilton et al. 2020) #	East of England, UK	February – May	Surveillance	292	6600	1167	NA	700	409
(Page et al. 2020) #	East of England, UK	March - August	Surveillance	6	1035	76	9,3[b]	89	2
(Graham et al. 2020)	London, UK	April	Outbreak	4	383	126	3	19[c]	?
(Ladhani, Chow, Janarthanan, Fok, Crawley-Boevey, Vusirikala, Fernandez, Perez, Tang, Dun-Campbell, Evans, et al. 2020; Ladhani, Chow, Janarthanan, Fok, Crawley-Boevey, Vusirikala, Fernandez, Perez, Tang, Dun-Campbell, Wynne-Evans, et al. 2020)	London, UK	April	Outbreak	6	518	105	53	99	2
(Lemieux et al. 2020) #	Boston, USA	January - May	Surveillance	1	194	82	36	83	3
(Zhang et al. 2020)	California, USA	March - April	Surveillance	2	10	6,1[a]	3	192	1
(Quicke et al. 2020) #	Colorado, USA	N/A	Surveillance	5	454	NA	70	38	1
(Taylor et al. 2020)	Minnesota, USA	April - June	Outbreak	2	600	165	114	105	4
(Arons et al. 2020)	Washington, USA	March	Outbreak	1	89	57	26	34	1

[a] family member of resident, [b] family members of a single staff member, [c] paper states both 17 and 19 samples sequenced so it is not clear which is correct, [d] six of these samples were from an epidemiologically linked hospital outbreak, ? not reported, # preprint prior to peer review.

Discussion

To date, genomic epidemiology studies of SARS-CoV-2 in LTCFs provide a number of insights into transmission in this vulnerable population. The diversity of studies ranged from outbreak investigations with detailed epidemiological data in single LTCFs to prospective surveillance of hundreds of LTCFs, as summarised above. Serial sampling of residents and healthcare workers provided information about the duration of infection in individuals, the duration of outbreaks in LTCFs, and the reproducibility of genome sequencing and lineage identification.

Large outbreaks, such as (Lemieux et al. 2020), in LTCFs generally shared the same characteristics: a single lineage with rapid expansion, resulting in the majority of samples being identical or near identical (1 SNP difference). Residents and staff, including staff with no contact with residents, were usually infected with the same (identical) genome sequence. The direction of transmission cannot be determined from genomic data alone, but the addition of traditional epidemiological data (such as sample dates and co-location of individuals) may allow inferences to be drawn. In many outbreaks more than one lineage was observed (Hamilton et al. 2020) but these sporadic introductions usually represented a small minority of cases. Temporal analysis of genomic data allows estimation of when an introduction into a LTCF is likely to have occurred, making genomics useful to identify the index case (first positive person in a setting), which may have been asymptomatic (Lemieux et al. 2020). The Boston study of (Lemieux et al. 2020) estimated that, after an introduction, 85% of residents were infected within 2-3 weeks, despite extensive infection prevention and control measures being in place. Further, follow up of the London-six study (Ladhani, Chow, Janarthanan, Fok, Crawley-Boevey, Vusirikala, Fernandez, Perez, Tang, Dun-Campbell, Evans, et al. 2020) demonstrated by five weeks the majority of individuals had sero-converted, including 66.4% of staff and 67% of residents who were asymptomatic and tested negative by RT-PCR (Ladhani, Jeffery-Smith, Patel, Janarthanan, Fok, Crawley-Boevey, Vusirikala, Olano, Perez, Tang, Dun-Campbell, et al. 2020). By the time two symptomatic individuals are identified in a LTCF, the outbreak is likely to be widespread (Arons et al. 2020; Ladhani, Chow, Janarthanan, Fok, Crawley-Boevey, Vusirikala, Fernandez, Perez, Tang, Dun-Campbell, Evans, et al. 2020).

Analysis of lineages circulating in a region compared with lineages found within LTCFs (Page et al. 2020; Lemieux et al. 2020) show that there is limited diversity within LTCFs, pointing to a very small number of introductions rather than repeated introduction from the community. Lineages within LTCFs are usually (but not always) different from other LTCFs (Hamilton et al. 2020), with genomics identifying a small number shared lineages in different LTCFs (Hamilton et al. 2020; Page et al. 2020; Ladhani, Chow, Janarthanan, Fok, Crawley-Boevey, Vusirikala, Fernandez, Perez, Tang, Dun-Campbell, Evans, et al. 2020). Taking the sequence diversity found in 292 LTCFs in a region (Hamilton et al. 2020) as a whole and comparing it to a similar number of non-LTCFs residents in the same region, similar numbers of SNP differences were identified in the genomes (LTCF median 8 SNP differences, non-LTCF residents median 9 SNP differences). When looking at a single LTCF (Page et al. 2020), knowing the diversity of circulating lineages within the locality helped rule out local inward transmission.

Looking more closely at the dynamics of an outbreak, the Washington study of (Arons et al. 2020) overlayed unique sequences to a map of the residents' bedrooms and showed a clear spatial signal, with residents more likely to be infected with identical genome sequences if their bedrooms were in close proximity, even with strict infection controls. Genome sequencing also identified examples of links between outbreaks at LTCFs located in the same geographical areas (Page et al. 2020; Ladhani, Chow, Janarthanan, Fok, Crawley-

Boevey, Vusirikala, Fernandez, Perez, Tang, Dun-Campbell, Evans, et al. 2020). In one study (Hamilton et al. 2020) two LTCFs located within 1km of each other had residents infected with identical genomes; a paramedic who visited both also tested positive. In another study (Page et al. 2020) a genetically distinct sub-lineage was found in six different LTCFs within 1 small region. Genomics shows that inter-LTCF transmission of SARS-CoV-2 is a real risk, and is potentially enabled by the use of shared staff or temporary ‘agency’ workers.

When there is an outbreak at a LTCF, the genomes identified in residents and staff, including non-healthcare workers, are usually the same. A high percentage of asymptomatic cases is common, with staff usually accounting for a higher percentage of asymptomatic cases. Therefore, the same SARS-CoV-2 genome can result in both symptomatic and asymptomatic infections. It is important to include staff in testing, although it has been noted that participation rates are often low (Taylor et al. 2020). Even with intensive consecutive weekly testing, enhanced infection control, and transfer of positive residents to dedicated isolation units, the outbreak continued in the Minnesota study, with the same lineages found over an extended period of time (Taylor et al. 2020).

Intensive sequencing of all residents and staff in an outbreak does not provide additional genomic information after the first few sequences. Strategic sub-sampling of staff and residents should be adequate to understand the number of clusters and their relative proportions. However, inadequate sampling does have a large impact on the utility of genomics. Genomics has reduced utility once there is a large outbreak, however, it does provide useful information about how SARS-CoV-2 enters a home, such as via staff or patient movements, and continued ongoing monitoring using genomics can identify new sources of infection (new seeding events), which can help inform policy. As visitors were restricted from visiting LTCFs early in the pandemic, no data is available on their role as a source of introduction. Should visiting be resumed, this will be an important cohort to target for testing and sequencing, to fill this knowledge gap. Further modelling of this is recommended.

Genomes sequenced through prospective surveillance have proven useful for identifying linked outbreaks which may have been missed otherwise (Hamilton et al. 2020; Page et al. 2020; Zhang et al. 2020; Meredith et al. 2020). If the time between taking the swab sample to sequencing and cluster linkage is low, then this can help outbreak investigation and contact tracing efforts. The limitation is that it may take time for an outbreak to be recognised through surveillance activities, where even if the intention is to sequence every positive sample, a large percentage of genome sequences are not available (Hamilton et al. 2020; Page et al. 2020).

So far, the vast majority of genomic epidemiology studies of SARS-CoV-2 in LTCF have been conducted in the UK, The Netherlands and the USA. Furthermore, two thirds of the global SARS-CoV-2 genomes sequenced to date have been generated by the COG UK consortium. This has enabled detailed analyses on a large scale but also introduced a risk of bias. The dynamics of SARS-CoV-2 transmission in LTCFs in other countries may be different.

LTCF residents who develop severe COVID-19 are often admitted to hospital, which may be the first indication of an outbreak. Samples that are sequenced as part of surveillance studies may therefore represent the tip of the iceberg (Zhang et al. 2020; Lemieux et al. 2020; Hamilton et al. 2020; Page et al. 2020). Hamilton et al. (2020) found that 5.8% of COVID-19 infections in LTCF residents were suspected to be hospital-acquired. Furthermore, 33.1% of patients were discharged within 7 days of their first positive test and could therefore have been infectious at the time of hospital discharge. This has important implications for infection control in LTCFs and public health policy (Hamilton et al. 2020).

Routine genomic surveillance of hospital patients and staff, LTCF residents and staff, and community cases will provide greater insights into transmission dynamics, although supporting studies with additional epidemiological information, such as hospital discharges, patient movements, and discharge locations, would provide a much more informative approach. The mortality rate amongst LTCF residents was high in all cases, regardless of the lineage or lineages circulating within the LTCFs.

Future work

The Vivaldi study (Krutikov et al. 2020) (undergoing open peer review) is a large prospective study in England, UK, covering 105 LTCFs with over 5000 residents and 6500 staff. Whilst it does not provide analysis of genome sequencing data yet, it is an example of an ongoing study that will help fill in many knowledge gaps and can help inform UK policy directly. In a follow-up survey (Vivaldi-1) managers of LTCFs were surveyed to estimate the proportion of staff and residents who had been infected with SARS-CoV-2 since the start of the pandemic and to identify factors associated with infection. In a third study (Vivaldi-2) seroprevalence was explored, starting in Four Seasons LTCFs (a commercial care provider) with the intention of expanding beyond this provider (from October 2020) to identify general attributes that apply across the industry. This study will also assess declining immunity using consecutive antibody testing (three rounds in staff and residents, a further two rounds in residents over a 12-month period).

Table 3: Sequencing and bioinformatics methods used in the LTCF genomic epidemiology studies

Study	Sample preparation	Sequencing	Method	Phylogeny	Data availability#
South East Netherlands	Amplicon	Nanopore	Consensus	Not reported	Not available
South Holland	Amplicon	Nanopore	Consensus	IQ-TREE	Not available
East of England	ARTIC	Nanopore/ Illumina	Consensus	IQ-TREE	Available but not linked
Norfolk	ARTIC	Illumina	Consensus	IQ-TREE	Available
London four	ARTIC	Illumina	Reference guided assembly	IQ-TREE	Not available
London six	WGS	Illumina	Consensus	IQ-TREE	Available but not linked
Boston	Metagenomic	Illumina	Reference guided assembly	IQ-TREE	Available
California	Metagenomic	Illumina	Consensus	IQ-TREE	Available
Colorado	ARTIC	Illumina	Consensus gap filled with reference	Geneious	Not available
Minnesota	ARTIC	Not reported	Not reported	IQ-TREE	Available but not linked
Washington	Not reported	Nanopore	Consensus	Geneious	Available

If data is present in GISAID or a ENA/NCBI/DDJB database it is labelled as ‘Available’, and when there is no linkage information between the samples used in the manuscript and the data in the public archives, it is labelled as ‘not linked’.

Informatics analysis strategies

Bioinformatics methodologies differed between studies, tailored to the sequencing technologies (Illumina or Oxford Nanopore Technologies) and protocols (ARTIC amplicon (Benjamin Farr et al. 2020), metagenomic). Three studies had clear deficiencies relating to the bioinformatics methods used or the results presented. These included assembling amplicons, using poor quality sequencing data in phylogenetic analysis, and imputing reference bases to replace missing bases; the impact of this on downstream analysis is unknown (unpublished Page and MacCannell, 2020).

Most studies performed phylogenetic analysis of their datasets as a final step and presented the results as a dendrogram. The open source bioinformatics software IQ-TREE (Minh et al. 2020) was used in eight out of 11 studies, with commercial software Geneious (Biomatters) used in two. The way in which the analysis was done differed greatly, which had an impact on the granularity presented and largely prevented direct comparisons. Assuming a mutation rate of approximately 2.5 SNPs per month (Page et al. 2020) allows you to estimate the amount of variation expected in a phylogeny at any particular time point in a series. A known source of increased variation is C->U mutations from RNA degradation (De Maio et al. 2020) which are caused by suboptimal sample storage conditions; these should be accounted for. Additionally, most studies do not mention negative controls and no studies have released them publicly. Many sample preparation protocols employ amplification techniques that can also amplify contamination and give false results, particularly when the viral load in the source material is low. Without having all of the underlying raw data (including controls) alongside the sample preparation and sequencing metadata (Griffiths et al. 2020), reanalysis and comparison between studies is difficult and error prone.

As the pandemic progressed, sequencing and bioinformatics methods were refined (detailed in Table 3) making direct comparisons between studies subject to the batch effect. Many studies publicly release their raw sequencing data and/ or consensus/ assembled genomes (Page et al. 2020; Hamilton et al. 2020; Lemieux et al. 2020; Zhang et al. 2020) through the INSDC (Cochrane et al. 2016) and through GISAID (Shu and McCauley 2017). This allows for independent reanalysis, overcoming the effect of variation amongst methods. However, in genomics it is a common poor practice not to release data publicly, or to provide insufficient metadata, such as accession numbers, making reanalysis unfeasible. The use of internationally-agreed open metadata standards for SAR-CoV-2 genomics enables genomic epidemiology on a global scale (Griffiths et al. 2020). In some cases, there are legitimate reasons to withhold data, such as to maintain patient confidentiality where identification maybe possible (e.g. as a result of small sample numbers or the location of LTCFs. In other cases, even if authors wish to deposit all clinically important samples, some of their samples may not meet the minimum quality control thresholds (>90% completeness for GISAID) enforced by the public databases (to aid high quality phylogenetic analysis). For example, samples with low viral load often sequence poorly, leading to incomplete datasets, making reanalysis impossible. Researchers may have a well-meaning desire to make data publicly available which does not meet the stringent quality control thresholds by imputing missing data from a reference genome (Quicke et al. 2020), a common technique in human genetics, but this leads to erroneous results in phylogenetic analysis of SARS-CoV-2. These high quality thresholds on data inhibit reanalysis and reduce available data. The COG-UK consortium has overcome this by making these data available on their website (https://www.cogconsortium.uk/data/) which include lower quality control thresholds (>50% genome completeness).

Recent convergence on a small number of open-source bioinformatics workflows employing best practices should mitigate future issues in this regard (e.g. https://github.com/connor-lab/ncov2019-artic-nf, https://covid19.galaxyproject.org, and https://app.terra.bio/#workspaces/pathogen-genomic-surveillance/COVID-19).

Recommendations

Table 4: Recommendations for measures derived from the use of SARS-CoV-2 genomics in LTCFs.

Transmission of SARS-CoV-2	**Findings**	**Impact**
Limiting the spread of SARS-CoV-2 between hospitals, healthcare workers and LTCF residents is an urgent infection control and public health priority.	2-6, 9, 11-12	Transmission
All staff, not just individuals with direct contact with residents, should be treated as one cohort and subject to the same Infection Prevention Control measures.	3	Transmission
Genomics identifies transmission between staff, between staff and residents, and between care facilities. This should direct future control measures.	2-5	Transmission
Clustering based on physical proximity to the bedroom of an infected resident supports its use as an additional factor to identify at risk individuals and prioritise testing.	11	Transmission, Resource allocation
LTCF Sequencing strategy		
A targeted approach weighted towards sequencing early positive samples in an outbreak coupled with potential epidemiological links can help highlight the source of introduction; widespread sequencing within a care home is unlikely to yield significantly more information	3-4, 7, 9-10	Transmission, Resource allocation
Genomic surveillance in a proportion of samples from LTCFs should be done, to include both patients and staff, allowing the genomic epidemiology of a LTCF to be put into context.	3-5, 7-10, 13	Transmission, Resource allocation
Residents with a recent hospital admission who subsequently test positive should have their genome sequenced to identify hospital seeding of LTCF outbreaks.	2, 6	Transmission, Resource allocation
Ongoing community surveillance with sequencing allows LTCF outbreaks to be better characterised	1-2, 4-5	Transmission, Resource allocation
Recommendations for future research		
Modelling of subsampling strategies within LTCFs is needed to optimally utilise genomic surveillance.	7	Transmission, Research need
Epidemiological and genomic data should be released to public archives with sufficient metadata to enable genomic epidemiology.	15-16	Transmission, Research need
Appropriate, validated, bioinformatics methods should be applied to genomic analysis with domain experts reviewing results to avoid erroneous results	14	Transmission, Research need
A focus on rapid integrated epidemiological and genomic analysis will have the most clinical benefits.	14-16	Transmission, Resource allocation

Having reviewed the available literature, we have drafted some recommendations for the use of genomics to evaluate SARS-CoV-2 in LTCFs, which are summarised in Table 4. We recommend that all staff working in a LTCF (regardless of their role) are treated as a single cohort and subject to uniform infection prevention control measures including appropriate use of personal protective equipment, regular screening for SARS-CoV-2 and genome sequencing of any positive samples. Genome sequencing shows that staff who do not have direct contact with residents have the same lineages in an outbreak as residents and staff with direct contact with residents. Early identification and exclusion of asymptomatic staff may reduce the risk of transmission to residents and other staff. It should be noted that regular screening of staff for asymptomatic infections may sometimes fail to identify an infectious individual that could lead to a superspreading event (Bedford et al. 2020).

Sequencing every genome in an outbreak is not recommended as it provides rapidly diminishing returns. Instead strategic sequencing of a subset of samples should be undertaken. The strategy for sequencing positive samples should be weighted towards staff rather than residents as they are at risk of community acquisition and subsequent transmission, whereas residents are less likely to have external contact. Modelling of subsampling strategies within LTCFs is needed.

Once a LTCF resident tests positive, other residents with bedrooms in close proximity should be considered to be at extremely high risk, regardless of contact patterns and other infection control measures, as genomics shows identical genomes are more likely to be found in those in close proximity.

Residents who have had a recent hospital admission and who subsequently test positive (within 14 days of hospital discharge) should have their viral genomes sequenced to distinguish hospital-acquired acquisition from care-home acquisition thus informing outbreak investigation and management. Limiting the spread of COVID-19 between LTCF residents, health care workers and hospitals should be a key target for infection control and prevention.

Raw sequencing data and consensus/assembled genomes should be made available in the public archives in a timely manner with the internationally-recommended minimal set of metadata to enable local, national and international genomic epidemiological analysis (Griffiths et al. 2020). This data sharing is essential to provide context for transmission analysis and outbreak investigations. Bioinformatics analysis of viral data requires additional considerations compared with other organisms. To increase the quality of the analysis, and minimise the probability of missing one of these domain-specific considerations, we recommend the use of validated and tested SARS-CoV-2 pipelines, with domain-specific experts analysing and reviewing the results.

Genomics provides the most clinical benefits and insights if it is integrated with detailed epidemiological data in a timely fashion. We recommend a focus not only on rapidly generating and analysing sequencing data, but also on rapidly collecting and integrating epidemiological data, which is often held in many different databases in different organisations. The ability to combine genomic and epidemiological analysis in a clinically actionable time frame (days rather than months) is crucial to leveraging the clinical benefits of sequencing.

Conclusions

There are many open questions around transmission of SARS-CoV-2 in LTCFs. The Vivaldi study in the UK is structured in such a way that it will help answer some of these over the next phase of the pandemic. Genomics can help understand the initial seeding of outbreaks in LTCFs. For example, they can link existing outbreaks to other LTCFs, identify the likelihood of inter-LTCF transmission, and link outbreaks to hospital cases confirming nosocomial infection. Placing these outbreaks in the context of the wider circulating lineages in the locality can inform about routes of transmission. For example, it can separate local community transmission from other routes of transmission which informs policy and helps limit future outbreaks. Genome sequencing of SARS-CoV-2 has been proven to provide useful insights into the transmission and dynamics of outbreaks. Prospective surveillance provides a backbone of information, helping to inform outbreak analysis. Hidden transmission links are uncovered using genomics that help with interpretation of epidemiology and with contract tracing efforts. Consecutive sampling provides yet more insights into virus longevity and transmission within LTCFs, and the reproducibility of genome sequencing for lineage identification when the same patient is sampled and genome sequencing done repeatedly. The ability to integrate epidemiological and genomic analysis in a clinically actionable timeframe remains a major challenge to realising the clinical benefits of genomics.

Appendix

Summary of major results from each study

A summary of each major study is presented below, grouped alphabetically by country.

South East Netherlands (Dautzenberg et al. 2020)

A study (Dautzenberg et al. 2020) of staff at LTCFs was done in March-April 2020 in south east Netherlands to understand their potential role in the spread of COVID-19 amongst residents in LTCFs. Staff members and a number of residents with mild respiratory symptoms were targeted for testing using rRT-PCR. The policy at the time was for staff with mild respiratory symptoms to wear a mask and still attend for work. At this point in time symptoms of COVID-19 and asymptomatic spread of SARS-CoV-2 were poorly understood.

A high prevalence of positive tests was found amongst staff at LTCFs, with 133 (21%) of those tested being positive; these were predominantly those providing care directly (nursing 74%, 1% physicians, 20% other health care workers) but also staff without patient contact (3%). Whilst the original intention was to test staff with mild respiratory symptoms, many staff with moderate symptoms were identified. Twenty-two samples from staff and residents at two LTCFs were selected for sequencing using an amplicon method, similar to ARTIC. Phylogenetic analysis indicated that the samples clustered by LTCF, with the exception of a single sample which represented a second introduction in one LTCF. Data is not in public archives and only available on request from the authors.

South Holland, Netherlands (Besselaar et al. 2020)

A study (Besselaar et al. 2020) was done to investigate an outbreak at a single LTCF in South Holland in the Netherlands from April to June 2020. A resident tested positive for SARS-CoV-2 after discharge from hospital, seeding an outbreak at the LTCF. This was confirmed by sequencing 60 samples from staff and residents at the LTCF and from the epidemiologically linked hospital ward where the LTCF resident had been an inpatient. All genomes clustered together, with some forming two sub-clusters, confirming the relatedness of the outbreaks. Sequencing helped to understand transmission patterns which would not have been observed using traditional epidemiological techniques. Of the residents tested for SARS-CoV-2 using rRT-PCR, 113 (62.4%) were positive; only four declined to be tested. Residents who tested positive were more likely to be older and have cognitive impairment compared with those that tested negative.

Of the staff tested, 73 (20.8%) were positive; 34 (9%) declined to be tested. A large percentage of staff (65%) reported working while symptomatic. There was no difference in Ct values between symptomatic, pre-symptomatic and asymptomatic individuals, where pre-symptomatic is defined as individuals who were asymptomatic at the point of testing, but later developed symptoms. The sequence data is not available in the public archives and only available on request from the authors.

East of England study, UK (Hamilton et al. 2020)

A large-scale surveillance study of SARS-CoV-2 positive cases was done in the East of England, UK, between February and May 2020 (Hamilton et al. 2020), with genomic analysis of samples from 292 LTCFs (residential and nursing homes). This is considerably larger than previous studies and linked epidemiological data with genomic data. Included in the study were 7,406 samples that tested positive for SARS-CoV-2 from 6,600 patients. Of these 1,167 / 6,413 (18.2%) of the study population were residents in 337 LTCFs. From these 193 / 337 (57.3%) were residential homes and 144 / 337 (42.7%) were nursing homes, mainly located in five counties in the East of England (Cambridgeshire, Bedfordshire, Essex, Hertfordshire and Suffolk). This represents around half of the care homes in the region that had reported suspected or confirmed COVID-19 outbreaks to PHE at the time. Diagnostic samples were tested at the Public Health England (PHE) Clinical Microbiology and Public Health Laboratory (CMPHL) at Cambridge University Hospitals NHS Foundation Trust. Samples were sequenced in the Division of Virology, Department of Pathology, University of Cambridge, as part of the COVID-19 Genomics Consortium UK (COG-UK).

The investigators found 409 distinct viral clusters in 292 LTCFs, corresponding to approximately half of all the LTCFs in the region that had reported outbreaks. Multiple potential transmissions between residents and staff were identified using genomic data. Multiple clusters per care home suggested that independent introductions were common and that within-care home transmission occurred frequently. The median number of cases per care home was two (range 1 - 22), with ten (3%) LTCFs with the highest numbers of cases accounting for 22% of all cases. There was evidence of large-scale outbreaks of identical or near identical (≤ 1SNP difference) lineages in care homes with the largest numbers of genomes. A median of eight SNPs separated genomes within LTCFs, compared with a median of nine SNPs for a random selection of non- LTCFs samples, indicating that genomic diversity amongst positive samples from LTCFs was similar to that from non-LTCFs samples. There were two LTCFs, located within 1km of each other, that had probable inter- LTCF transmission, with links to the same paramedics and shared carers. The genomes present in each had zero SNP differences, with only 2 days between the times that samples were taken. By combining epidemiological and genomic data it was possible to confirm a high probability of transmission. Admission and patient movement data were highlighted as a priority for investigation in relation to transmission.

The proportion of LTCF residents testing positive increased as general transmission decreased during lockdown, although it should be noted that sample collection strategies changed during the study period. Cases in LTCFs appeared more resistant to non-pharmaceutical intervention (NPI) measures. The study investigators also examined links between LTCFs cases and hospital admissions (Hamilton, personal communication). During the study period 470 / 694 (67.7%) of LTCF residents had at least one hospital admission, and 398 / 694 (57.3%) were admitted to hospital with COVID-19 infection. 40 / 694 (5.8%) cases were categorised as suspected hospital-acquired COVID-19 infections. Furthermore, 230 / 694 (33.1%) of individuals were discharged from hospital within seven days of their first positive test, and could therefore have been infectious at the time of hospital discharge. Limiting the spread of SARS-CoV-2 between hospitals, healthcare workers and LTCF residents should be an urgent infection control and public health priority.

During the study period, no new viral lineages from outside the UK were observed in the entire dataset; this included genomes from LTCF and non-LTCF samples, suggesting travel restrictions had been successful in minimising new importations. The genome sequencing was based on the ARTIC protocol (https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w) utilising Nanopore and Illumina sequencing platforms as part of the COG-UK consortium. Overall, viral sequence data was not available for 40% of samples from LTCF residents testing positive; this was due to a combination of missing samples, mismatches between metadata and sequences, genomes not passing quality control, or sequencing being unavailable at the time of analysis. This highlights the practical difficulties in undertaking genomic surveillance when there are large numbers of samples. Despite availability of: all the consensus genomes in the database; the Global Initiative on Sharing All Influenza Data (GISAID) (Shu and McCauley 2017); and al the raw data in the European Nucleotide Archive (ENA), the links between LTCF samples used for the analysis were not available (for reasons of patient confidentiality), although they may be requested from the authors.

Norfolk, UK (Page et al. 2020)

A large-scale surveillance study (Page et al. 2020) was undertaken in Norfolk, UK from March to August 2020 as part of the COG-UK consortium. A total of 42% (n=1,035) of all samples from SARS-CoV-2 positive cases from the hospital testing system within the region (covering hospitals, LTCF, health care workers) were sequenced. An outbreak in one LTCF was investigated using data from genome sequencing of samples that had been prospectively collected. It was noted that the genomes from this LTCF were identical to each and formed a distinct sub lineage that included genomes from additional cases clustered in small geographical areas around the LTCF within a short period of time (between 8 April and 1 June 2020). As most of these additional cases were in older people (>85 years old), it was suspected that they represented LTCF clusters. It was later confirmed that the sub lineage was circulating in six LTCFs.

There were 90 cases caused by this sub lineage, of which 64 were known to be LTCF residents, nine were healthcare workers, and three were family members of healthcare workers. The majority of the LTCF infections were community-acquired. Twelve of these cases were admitted to a hospital (two were admitted twice to three different hospital trusts). Six had a community-acquired infection, testing positive within seven days of admission, three were inconclusive due to missing data, one had a probable hospital-acquired infection and tested positive within seven days of discharge, and two had a definite hospital-acquired infection (https://www.gov.uk/government/publications/wuhan-novel-coronavirus-infection-prevention-and-control/epidemiological-definitions-of-outbreaks-and-clusters-in-particularsettings). All LTCF residents with hospital-acquired infections were tested prior to discharge, suggesting that the adult social care IPC Department of Health and Social Care measures announced in May (which required this) were being followed. In the time period covered by the study, while patients required a test before they could be discharged to a LTCF, a positive test did not preclude them from returning to the LTCF just that adequate IPC measures needed to be taken. Given that some of this cohort of patients tested positive for community-acquired infections in May, a number of weeks after the adult social care IPC Department of Health and Social Care measures were announced, suggests that these measures may not have been sufficient.

This multi- LTCF outbreak was only identified using genomic epidemiology and the links between LTCFs were unknown prior to this investigation. The sub lineage involved was not detected in cases identified as part of community testing. This study indicated inter- LTCF transmission was likely. The initial seeding events were unknown. A limitation of the study is that the collection date of a sequenced sample may not be the first positive test for that case. All data analysed as part of this study is detailed in the preprint, with accession numbers for each sample linked to accession numbers for the public archives, which enables reanalysis.

London four study, UK (Graham et al. 2020)

An investigation of an outbreak in four LTCFs in London was done in April 2020 (Graham et al. 2020). Residents were tested for SARS-CoV-2 by rRT-PCR at two time points one week apart; residents who were initially positive were not re-tested. Of all those residents tested 126 (40%) were positive; of these 54 (42.9%) were asymptomatic. There was a high COVID-19 mortality rate (26%) amongst residents during the study period. In terms of ethnicity, 18.5% of the residents were Black, Asian and minority ethnic (BAME) and the mortality rate for these residents was similar to that of white residents.

A subset (n=70, 11%) of asymptomatic staff at three of the four LTCFs were offered SARS-COV-2 testing. Only three (4%) out of 70 tested positive, but it was noted that staff absence rates due to sickness or self-isolation were more than twice the normal rate during the study period. Furthermore, three different diagnostic platforms were used for testing, including one with a lower sensitivity than the others. These changes in methodology may have affected the reliability of the study results.

In terms of sequencing, samples from one staff member and 17 (19 reported in the manuscript) residents were sequenced. However, samples that were sequenced were not representative of all the LTCFs; there was only one sample sequenced from one LTCF and only two samples sequenced from a second LTCF. The bioinformatics methods used (assembling amplicons) is generally regarded as poor practice and likely to result in errors in the sequences generated. Furthermore, no sequencing data were publicly available, precluding reanalysis of the primary data.

The genomic data identified a cluster with one staff member and two residents at a single care home. Most of the samples clustered by care home, and visual inspection of the phylogenetic tree presented indicates there were two clusters in each of two care homes, although this was not entirely clear from the data presented. The data were compared to a random selection of UK samples to provide background context and help to show separation between clusters.

London six study, UK (Ladhani, Chow, Janarthanan, Fok, Crawley-Boevey, Vusirikala, Fernandez, Perez, Tang, Dun-Campbell, Evans, et al. 2020)

A study of six London (UK) LTCFs experiencing COVID-19 outbreaks was done over the Easter weekend (10-12 April, 2020) (Ladhani, Chow, Janarthanan, Fok, Crawley-Boevey, Vusirikala, Fernandez, Perez, Tang, Dun-Campbell, Evans, et al. 2020; Ladhani, Chow, Janarthanan, Fok, Crawley-Boevey, Vusirikala, Fernandez, Perez, Tang, Dun-Campbell, Wynne-Evans, et al. 2020). The LTCSs in this region had outbreaks of COVID-19 early on in the UK epidemic, before there was full recognition of the extent of community transmission and the frequency of asymptomatic transmission. The study was of 518 individuals and included both staff and residents; 105 residents and 53 staff tested positive for SARS-CoV-2 using rRT-PCR. Virus was cultured by Public Health England Colindale to ascertain infectivity and patients were tested serologically for evidence of previous infection (Ladhani, Jeffery-Smith, Patel, Janarthanan, Fok, Crawley-Boevey, Vusirikala, Olano, Perez, Tang, Dun-Campbell, et al. 2020).

Of those who tested positive for SARS-CoV-2 there was a high percentage of asymptomatic cases amongst staff (49%, n=26) and residents (44%, n=46), indicating that symptom screening has low sensitivity. Detection of outbreaks was often delayed if based on symptoms as, by then there were already high rates of asymptomatic infection in both staff and residents. The rRT-PCR cycle threshold values (Ct), which indicate SARS-CoV-2 viral load, were similar across different age groups, and between symptomatic and asymptomatic cases. Infectivity in culture was also similar across different age groups, and between symptomatic and asymptomatic cases. A high percentage of cases with symptoms tested negative by rRT-PCR (15%, n=24 residents; 9%, n=19 staff). This may indicate that: sampling was inadequate; viral loads were too low to be detected (early/late infection); or that diagnostic screening by rRT-PCR with a single target gene may underestimate infection. Residents who were symptomatic and tested positive by rRT-PCR had a higher mortality rate than those who were symptomatic and tested negative by rRT-PCR (36% versus 4%).

Genome sequencing of 99 out of 158 cases that tested positive (62%) revealed two distinct lineages (Rambaut et al. 2020) predominantly B.1 and B.2.1, two of the most common UK lineages. All six LTCFs had both lineages, and genomes from both staff and residents were interspersed throughout the phylogenetic tree, likely due to the low genetic diversity of SARS-CoV-2 genomes in April, 2020. To provide genomic context we examined the publicly available virus genomes from the COVID-19 Genomics Consortium UK that had been collected in the week before and the week after the Easter weekend (5 - 19 April 2020) in the Greater London region. At this time, diagnostic testing was directed towards people with symptoms, so may not have been representative of community spread, but did provide an indication of the virus diversity within this small geographic region. A total of 44 lineages were observed from 617 genomes, including the London Six genomes. However, the London Six publications did not provide sample accession numbers so it was not possible to identify them within the public archives. The most common lineages in Greater London at the time were B.1.1.1 (n=298, 48%) followed by B.1 (n=90, 14%), B.2.1 (n=78, 13%) and B.1.5 (n=21, 3%).

The study found that there were up to nine separate introductions into a single LTCF. Reanalysis for this review indicates that there was more sequence diversity than expected for samples from the Easter time period, and the number of introductions into a single LTCF was likely to be six rather than nine. This over-estimation of introductions and sequencing diversity was caused by some poor-quality sequence data, including missing data, leading to bioinformatics artefacts. Had there been a high level of introductions into a single LTCF we would expect to observe more lineages, particularly the most common lineage for the region. The raw sequencing data and genomes are available in the public archives, however the specific samples used for this study were not detailed in the papers to maintain patient confidentiality. This limits public reanalysis.

Boston, USA (Lemieux et al. 2020)

A large community surveillance study was conducted between January and May 2020 in Boston, USA (Lemieux et al. 2020). In this study 850 SARS-CoV-2 positive samples were sequenced metagenomically using Illumina, with reference-guided assembly. Over 80 introductions were estimated to have occurred in the region over the study period.

A sub-study conducted in April 2020 analysed an outbreak in a single LTCF, where a planned relocation of residents led to universal screening of residents and staff for SARS-CoV-2. Out of those tested 82 (85%) residents and 36 (37%) staff tested positive. A total of 83 (67%) genomes were sequenced. From these 75 (90%) genomes formed a single closely-related cluster; 59 were identical (no SNP differences) and shared a distinct mutation (G3892T) with unknown significance. Genome sequencing indicated a recent introduction from a single source. Estimates for the most recent common ancestor allowed the authors to estimate that the time from introduction to widespread positive testing in residents was 2 - 3 weeks. Two additional introductions (three genomes each) were also observed but did not disseminate widely. The three introductions highlight the risk of introduction into a high-risk setting, despite strict infection control measures which had been in place from two weeks prior to the estimated introduction date. By tracking the mutation distinct to this outbreak as part of continued regional surveillance and sequencing indicated that there was little onward spread from this initial superspreading event. The raw read sequence data and the assembled genomes are deposited in the public archives (NCBI), allowing for reanalysis.

California, USA (Zhang et al. 2020)

A four-week prospective surveillance study was done on 192 patients with COVID-19 in a hospital in Los Angeles, California, USA between March and April 2020 (Zhang et al. 2020). Genome sequencing found that 85% of genomes were European lineages and 15% were Asian, indicating multiple sources of introduction. Out of all the samples, 113 (69%) yielded genomes of sufficient quality for use in phylogenetic analysis (>50% reconstructed consensus genome). The percentage of the genome that could be reconstructed was closely correlated with the number of viral copies in the primary sample.

From phylogenetic analysis of the sequenced SARS-CoV-2 genomes, a cluster of ten patients was identified: five of these were residents from a single LTCF, while the other five were associated with a LTCF one block away (three staff members, a family member of a resident and one resident). Another related case was identified in a person living near one of the LTCFs. Genome sequencing was used to establish connections between these cases; the genomes were identical (or near identical) to each other and belonged to lineage B.1. In total the study identified three large clusters, only one of which included genomes from a LTCF. This study demonstrates the effectiveness of prospective surveillance in detecting and linking outbreaks in LTCFs, and thus enhancing contact tracing efforts. The data are deposited in GISAID, with samples clearly described allowing for reanalysis.

Colorado, USA (Quicke et al. 2020)

A prospective surveillance study of staff at five LTCFs was done over a six-week period in Colorado, USA (Quicke et al. 2020). This involved consecutive testing for SARS-CoV-2 in staff at five LTCFs to investigate the prevalence of asymptomatic and pre-symptomatic positive tests. Staff voluntarily enrolled and were swabbed weekly throughout the study period, including if they developed symptoms. Staff with and without direct contact with residents were included in the study. A total of 70 staff members tested positive and rates of infection varied between LTCFs. The median number of consecutive weekly positive tests was 2 (range 1 to 5), indicating a detection window in the nasopharynx of most people of at least 8 days. Some individuals tested positive for five consecutive weeks, and some tested positive intermittently. The levels of viral RNA tended to decline over the duration of infection and corresponded to low levels of infectious virus in culture.

A total of 48 genomes from positive samples were sequenced, ten of which came from five staff members collected over two consecutive weeks. The ARTIC amplicon protocol was used, with Illumina sequencing; gaps in the consensus sequences were filled with bases from the reference genome, so the results should be treated with caution. Of those sequenced, 36 genomes from one LTCF clustered together, and a further five clustered with another LTCF. Transmission within the workplace was likely, but community transmission could not be ruled out. Of the five staff members with two sequenced genomes each, three had genomes that differed in SNPs between the two consecutive samples; this high rate of within-host mutation is likely to be due to a bioinformatics error associated with filling gaps with bases from the reference genome. The sequence data are not publicly available and could not be reanalysed.

Minnesota, USA (Taylor et al. 2020)

A prospective surveillance study was done in two LTCFs experiencing COVID-19 outbreaks between April and June 2020 in Minnesota, USA (Taylor et al. 2020). Residents (n=261) and staff (n=480) were offered SARS-CoV-2 testing up to six times and, once they tested positive, were not re-tested. Participation rates in testing varied by LTCF, with 17% of residents in one refusing to be tested initially. Of those that were tested, 165 (64%) residents tested positive, 33 of these (20%) were hospitalised and 52 (31%) died. Residents testing positive were isolated in a COVID-19 specific unit, but this had no impact on overall transmission as indicated by the continued identification of positive cases throughout the study. This study demonstrated the utility of serial (repeated) sampling of the same individuals for detecting new cases as they occurred, with the detection of new cases rapidly diminishing throughout the study.

Severe challenges were encountered with staff testing. Staff were reluctant to participate (71%), and when they did, did so only once. Overall, 114 (33%) staff tested were positive, of which 58 (51%) were symptomatic and working on the day of testing. Of those staff testing positive, 41 (12%) were not involved directly in care provision. There were delays of up to 12 days in obtaining test results, with staff incurring financial losses if they self-isolated without a positive test. Four staff members were hospitalised and two staff members died.

In this study genomes from 105 samples were sequenced using the ARTIC protocol. Genomes were clustered into two groups separated by LTCF (i.e all genomes in one cluster were from residents and staff of one LTCF while all genomes in the other cluster were from residents and staff of the other LTCF); this indicates within-home transmission and no evidence for transmission between LTCFs. Only 37% of positive samples were available for sequencing; samples from early in the outbreaks were missing. However, this was sufficient to be reasonably confident of the underlying clusters and dynamics observed. In one LTCF there appeared to be a second potential introduction event from the community, rooted earlier in the tree. However, as there were few specimens from early stages of the outbreak for sequencing, the full genetic evolutionary history cannot be elucidated further. Sequence data from this study are available on GISAID, but there are no sample identifiers or accession numbers in the manuscript to enable linkage or re-analysis of the data.

Washington, USA (Arons et al. 2020)

This study in Washington state, USA, (Arons et al. 2020) was the first to report the use of genome sequencing to investigate a large COVID-19 outbreak in a LTCF. Following positive results from SARS-CoV-2 testing of one staff member and one resident, samples from residents were tested by Public Health–Seattle and King County (PHSKC) and the Centers for Disease Control and Prevention (CDC) on two occasions one week apart. Not all residents were tested. Real-time reverse transcription polymerase chain reaction (rRT-PCR) was used for identification and a subsample of positive samples were selected for culture and sequencing. No new residents were admitted to the LTCF after the first resident tested positive. Enhanced infection prevention and control (IPC) measures focused on symptomatic residents and staff were implemented after the first resident tested positive. However, testing 3 days after implementation showed widespread transmission had already occurred. There was a high percentage of positive cases amongst residents (64%, n=57), most of whom were asymptomatic or pre-symptomatic at the time of testing (82%, n=48). Fifteen residents died (26%) and a further 11 were hospitalised. No serological testing was done. The viral load (based on the rRT-PCR cycle threshold values [Ct]) and the percentage of cultures testing positive for virus were similar between symptomatic, pre-symptomatic and asymptomatic cases. Symptomatic staff (40%; n = 138) were advised to seek testing externally by their health care provider; of these 19% tested positive (n=26). Asymptomatic staff were not advised to be tested; this may have underestimated the infection rate. The role of staff in the introduction or transmission of SARS-CoV-2 was not fully explored or analysed in the study.

The doubling time was faster than in the surrounding community, but this may have been due to the identification of asymptomatic cases within the LTCS; only symptomatic individuals were tested in the community. The IPC measures focused on symptomatic cases, but the high prevalence of underlying conditions (cognitive impairment, chronic coughs) amongst residents made identification of COVID-19 symptoms difficult, particularly in the early stages of infection.

Nanopore sequencing was done on samples from 34 residents that tested positive; for five of these residents samples were taken twice, one week apart, sequenced, and the viral genomes found to be identical in the second test in each case. This demonstrated the reproducibility of SARS-CoV-2 genome sequencing. Bioinformatics analysis showed that 79% (n=27) of positive samples mainly clustered into two groups, separated by a single SNP difference; a small number of outlier samples containing up to 4 SNP differences. This confirmed the relatedness of genomes found in the residents' samples. Identical sequences were given a unique identifier and, when these were related back to a map of the facility and the location of the residents' bedrooms, there was a very clear spatial signal; residents in adjacent bedrooms were more likely to have 100% identical consensus genomes than not. Phylogenetic analysis of publicly available genomes at the time showed that the genomes from the LTCS samples were very closely related to those found elsewhere in the locality (Washington, USA). Sequencing data used in the paper has been publicly deposited in two archives, with sufficient information in the paper to enable the genomic analysis to be fully reproduced.

Other studies

There is one study from Hungary (Kemenesi et al. 2020) that sequenced a single LTCF resident's positive SARS-CoV-2 sample.

An unpublished study (personal communication Guthrie, Templeton and Holden) sequencing SARS-CoV-2 positive samples from staff and residents of LTCFs in Scotland as part of the COG-UK consortium. There was evidence for a connection between the genomes from staff and residents' samples in some LTCFs. Outbreaks were heterogenous in size, duration and pattern (some explosive, some more drawn out some with long gaps between cases).

Another unpublished study (personal communication Bashton, Young, Nelson, Smith) done as part of the COG-UK consortium sequenced genomes from staff and residents testing positive at 64 LTCFs in the North Yorkshire, South Tees region. Of these LTCFs, 36 had multiple positive samples enabling genome sequencing and cluster analysis. Sequence data analysis using Civet (https://github.com/artic-network/civet) detailed six outbreak clusters. One of these clusters involved three LTCFs and associated staff from their local National Health Service (NHS) trust. Another involved two LTCFs and an associated staff member. This demonstrated not only transmission between residents within LTCFs, but more complex transmission chains between LTCFs and local hospitals.

Definitions

We use the term 'long term care facility' which encompasses terms used to describe similar facilities in different countries such as: 'skilled nursing facility', 'care home', 'nursing home', 'elderly care home', and 'residential home'.

For the purposes of this review we use the ECDC definition of nosocomial infection (https://www.ecdc.europa.eu/en/covid-19/surveillance/surveillance-definitions).

Funding

DA is a Clinical PhD Fellow and gratefully supported by the the Wellcome Trust. AJP, JOG gratefully acknowledge the support of the Biotechnology and Biological Sciences Research Council (BBSRC); their research was funded by the BBSRC Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent project BBS/E/F/000PR10352, also Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/CCG1860/1). SR, DS and MB gratefully acknowledge support from Research England's Expanding Excellence in England (E3) Fund. IG is a Wellcome Senior Fellow and supported by the Wellcome Trust (207498/Z/17/Z). MET was supported by a Clinician Scientist Fellowship (funded by the Academy of Medical Sciences and the Health Foundation) and by the NIHR Cambridge Biomedical Research Centre. The COVID-19 Genomics UK (COG-UK) Consortium is supported by funding from 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 funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

All authors read the manuscript and consented to its publication. AJP lead the review and wrote the first draft of the manuscript. DA undertook enhanced analysis for the UK studies and provided clinical oversight. NMT, AJP, DA contributed to the literature search. WLH, IG, MET provided additional insights into the East of England study. LS provided insights into the Vivaldi study. JOG, AJP, EJM provided insights into the Norfolk study. TC, MC, TB, CB, MH, DTB, MB provided public health insights and guidance. RM undertook reanalysis of most of the UK studies. SJP instigated the review and provided overall leadership.

Conflicts of interest and disclosures

None declared.

Ethics

The UK studies were conducted as part of surveillance for COVID-19 infections under the auspices of Section 251 of the NHS Act 2006 and/or Regulation 3 of The Health Service (Control of Patient Information) Regulations 2002. They 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).

Acknowledgements

We thank members of the COVID-19 Genomics Consortium UK for their contributions to generating the data used in some of these studies. We thank Dr Judith Pell for critically assessing and improving this manuscript.

References

Abrams, Hannah R., Lacey Loomer, Ashvin Gandhi, and David C. Grabowski. 2020. ‘Characteristics of U.S. Nursing Homes with COVID-19 Cases’. *Journal of the American Geriatrics Society* 68 (8): 1653–56. https://doi.org/10.1111/jgs.16661.

Al-Tawfiq, J. A., and A. J. Rodriguez-Morales. 2020. ‘Super-Spreading Events and Contribution to Transmission of MERS, SARS, and SARS-CoV-2 (COVID-19)’. *Journal of Hospital Infection* 105 (2): 111–12. https://doi.org/10.1016/j.jhin.2020.04.002.

Arons, Melissa M., Kelly M. Hatfield, Sujan C. Reddy, Anne Kimball, Allison James, Jesica R. Jacobs, Joanne Taylor, et al. 2020. ‘Presymptomatic SARS-CoV-2 Infections and Transmission in a Skilled Nursing Facility’. *The New England Journal of Medicine* 382 (22): 2081–90. https://doi.org/10.1056/NEJMoa2008457.

Bedford, Trevor, Jennifer K. Logue, Peter D. Han, Caitlin R. Wolf, Chris D. Frazar, Benjamin Pelle, Erica Ryke, et al. 2020. ‘Viral Genome Sequencing Places White House COVID-19 Outbreak into Phylogenetic Context’. *MedRxiv*, November, 2020.10.31.20223925. https://doi.org/10.1101/2020.10.31.20223925.

Benjamin Farr, Diana Rajan, Emma Betteridge, Lesley Shirley, Michael Quail, Naomi Park, Nicholas Redshaw, et al. 2020. ‘COVID-19 ARTIC v3 Illumina Library Construction and Sequencing Protocol’, May. https://doi.org/10.17504/protocols.io.bgq3jvyn.

Besselaar, Judith Henriette van den, Reina S. Sikkema, Fleur M. H. P. A. Koene, Laura W. van Buul, Bas B. Oude Munnink, Ine Frenay, Rene te Witt, Marion P. G. Koopmans, Cees M. P. M. Hertogh, and Bianca M. Buurman. 2020. ‘A COVID-19 Nursing Home Transmission Study: Sequence and Metadata from Weekly Testing in an Extensive Nursing Home Outbreak’. *MedRxiv*, September, 2020.09.15.20195396. https://doi.org/10.1101/2020.09.15.20195396.

Burton, Jennifer K., Gwen Bayne, Christine Evans, Frederike Garbe, Dermot Gorman, Naomi Honhold, Duncan McCormick, et al. 2020. ‘Evolution and Impact of COVID-19 Outbreaks in Care Homes: Population Analysis in 189 Care Homes in One Geographic Region’. *MedRxiv*, July, 2020.07.09.20149583. https://doi.org/10.1101/2020.07.09.20149583.

Cochrane, Guy, Ilene Karsch-Mizrachi, Toshihisa Takagi, and International Nucleotide Sequence Database Collaboration. 2016. ‘The International Nucleotide Sequence Database Collaboration’. *Nucleic Acids Research* 44 (D1): D48–50. https://doi.org/10.1093/nar/gkv1323.

Dautzenberg, Mirjam Jeanne Dorine, Andrea Eikelenboom-Boskamp, Jacqueline Janssen, Miranda Drabbe, Ewoud de Jong, Eefke Weesendorp, Marion Koopmans, and Andreas Voss. 2020. ‘Healthcare Workers in Elderly Care: A Source of Silent SARS-CoV-2 Transmission?’ *MedRxiv*, September, 2020.09.07.20178731. https://doi.org/10.1101/2020.09.07.20178731.

De Maio, Nicola, Conor Walker, Rui Borges, Lukas Weilguny, Greg Slodkowicz, and Nick Goldman. 2020. ‘Issues with SARS-CoV-2 Sequencing Data’. Virological. 5 May 2020. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.

Fisman, David N., Isaac Bogoch, Lauren Lapointe-Shaw, Janine McCready, and Ashleigh R. Tuite. 2020. ‘Risk Factors Associated With Mortality Among Residents With Coronavirus Disease 2019 (COVID-19) in Long-Term Care Facilities in Ontario, Canada’. *JAMA Network Open* 3 (7): e2015957. https://doi.org/10.1001/jamanetworkopen.2020.15957.

Graham, N. S. N., C. Junghans, R. Downes, C. Sendall, H. Lai, A. McKirdy, P. Elliott, et al. 2020. ‘SARS-CoV-2 Infection, Clinical Features and Outcome of COVID-19 in United Kingdom Nursing Homes’. *The Journal of Infection* 81 (3): 411–19. https://doi.org/10.1016/j.jinf.2020.05.073.

Griffiths, Emma J., Ruth E. Timme, Andrew J. Page, Nabil-Fareed Alikhan, Dan Fornika, Finlay Maguire, Catarina Inês Mendes, et al. 2020. ‘The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology’, August. https://doi.org/10.20944/preprints202008.0220.v1.

Hamilton, William L., Gerry Tonkin-Hill, Emily Smith, Charlotte Houldcroft, Ben Warne, Luke Meredith, Myra Hosmillo, et al. 2020. ‘COVID-19 Infection Dynamics in Care Homes in the East of England: A Retrospective Genomic Epidemiology Study’. *MedRxiv*, September, 2020.08.26.20182279. https://doi.org/10.1101/2020.08.26.20182279.

Jordan, Rachel E., Peymane Adab, and K. K. Cheng. 2020. 'Covid-19: Risk Factors for Severe Disease and Death'. *BMJ* 368 (March). https://doi.org/10.1136/bmj.m1198.

Kemenesi, Gábor, László Kornya, Gábor Endre Tóth, Kornélia Kurucz, Safia Zeghbib, Balázs A. Somogyi, Viktor Zöldi, Péter Urbán, Róbert Herczeg, and Ferenc Jakab. 2020. 'Nursing Homes and the Elderly Regarding the COVID-19 Pandemic: Situation Report from Hungary'. *GeroScience*, May, 1–7. https://doi.org/10.1007/s11357-020-00195-z.

Krutikov, Maria, Tom Palmer, Alasdair Donaldson, Fabiana Lorencatto, Gill Forbes, Andrew Copas, James Robson, et al. 2020. 'Study Protocol: Understanding SARS-Cov-2 Infection, Immunity and Its Duration in Care Home Residents and Staff in England (VIVALDI)'. *Wellcome Open Research* 5 (October): 232. https://doi.org/10.12688/wellcomeopenres.16193.1.

Ladhani, Shamez N., J. Yimmy Chow, Roshni Janarthanan, Jonathan Fok, Emma Crawley-Boevey, Amoolya Vusirikala, Elena Fernandez, Marina Sanchez Perez, Suzanne Tang, Kate Dun-Campbell, Edward Wynne- Evans, et al. 2020. 'Investigation of SARS-CoV-2 Outbreaks in Six Care Homes in London, April 2020'. *EClinicalMedicine* 26 (September). https://doi.org/10.1016/j.eclinm.2020.100533.

Ladhani, Shamez N., J. Yimmy Chow, Roshni Janarthanan, Jonathan Fok, Emma Crawley-Boevey, Amoolya Vusirikala, Elena Fernandez, Marina Sanchez Perez, Suzanne Tang, Kate Dun-Campbell, Edward Wynne-Evans, et al. 2020. 'Increased Risk of SARS-CoV-2 Infection in Staff Working across Different Care Homes: Enhanced CoVID-19 Outbreak Investigations in London Care Homes'. *The Journal of Infection* 81 (4): 621–24. https://doi.org/10.1016/j.jinf.2020.07.027.

Ladhani, Shamez N., Anna Jeffery-Smith, Monika Patel, Roshni Janarthanan, Jonathan Fok, Emma Crawley-Boevey, Amoolya Vusirikala, Elena Fernandez Ruiz De Olano, Marina Sanchez Perez, Suzanne Tang, Kate Dun-Campbell, et al. 2020. 'High Prevalence of SARS-CoV-2 Antibodies in Care Homes Affected by COVID-19: Prospective Cohort Study, England'. *EClinicalMedicine* 0 (0). https://doi.org/10.1016/j.eclinm.2020.100597.

Lemieux, Jacob, Katherine J. Siddle, Bennett M. Shaw, Christine Loreth, Stephen Schaffner, Adrianne Gladden-Young, Gordon Adams, et al. 2020. 'Phylogenetic Analysis of SARS-CoV-2 in the Boston Area Highlights the Role of Recurrent Importation and Superspreading Events'. *MedRxiv: The Preprint Server for Health Sciences*, August. https://doi.org/10.1101/2020.08.23.20178236.

Meredith, Luke W., William L. Hamilton, Ben Warne, Charlotte J. Houldcroft, Myra Hosmillo, Aminu S. Jahun, Martin D. Curran, et al. 2020. 'Rapid Implementation of SARS-CoV-2 Sequencing to Investigate Cases of Health-Care Associated COVID-19: A Prospective Genomic Surveillance Study'. *The Lancet Infectious Diseases* 20 (11): 1263–72. https://doi.org/10.1016/S1473-3099(20)30562-4.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era'. *Molecular Biology and Evolution*, February. https://doi.org/10.1093/molbev/msaa015.

ONS. 2020. ‘Deaths Registered Weekly in England and Wales, Provisional - Office for National Statistics’. 16 October 2020. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregisteredweeklyinenglandandwalesprovisional/weekending16october2020.

Page, Andrew J., Alison E. Mather, Thanh Le Viet, Emma J. Meader, Nabil-Fareed J. Alikhan, Gemma L. Kay, Leonardo de Oliveira Martins, et al. 2020. ‘Large Scale Sequencing of SARS-CoV-2 Genomes from One Region Allows Detailed Epidemiology and Enables Local Outbreak Management’. *MedRxiv*, September, 2020.09.28.20201475. https://doi.org/10.1101/2020.09.28.20201475.

Public Health England. 2020. ‘COVID-19: Number of Outbreaks in Care Homes - Management Information’. GOV.UK. 27 August 2020. https://www.gov.uk/government/statistical-data-sets/covid-19-number-of-outbreaks-in-care-homes-management-information.

Quicke, Kendra, Emily Gallichote, Nicole Sexton, Michael Young, Ashley Janich, Gregory Gahm, Elizabeth J. Carlton, Nicole Ehrhart, and Gregory D. Ebel. 2020. ‘Longitudinal Surveillance for SARS-CoV-2 RNA Among Asymptomatic Staff in Five Colorado Skilled Nursing Facilities: Epidemiologic, Virologic and Sequence Analysis’. *MedRxiv: The Preprint Server for Health Sciences*, June. https://doi.org/10.1101/2020.06.08.20125989.

Rambaut, Andrew, Edward C. Holmes, Verity Hill, Áine O’Toole, J. T. McCrone, Chris Ruis, Louis du Plessis, and Oliver G. Pybus. 2020. ‘A Dynamic Nomenclature Proposal for SARS-CoV-2 to Assist Genomic Epidemiology’. *BioRxiv*, April, 2020.04.17.046086. https://doi.org/10.1101/2020.04.17.046086.

Shu, Yuelong, and John McCauley. 2017. ‘GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality’. *Eurosurveillance* 22 (13): 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

Swinkels, Koen. 2020. ‘Covid-19 Superspreading Events Database’. Medium. 1 November 2020. https://kmswinkels.medium.com/covid-19-superspreading-events-database-4c0a7aa2342b.

Taylor, Joanne, Rosalind J. Carter, Nicholas Lehnertz, Lilit Kazazian, Maureen Sullivan, Xiong Wang, Jacob Garfin, et al. 2020. ‘Serial Testing for SARS-CoV-2 and Virus Whole Genome Sequencing Inform Infection Risk at Two Skilled Nursing Facilities with COVID-19 Outbreaks - Minnesota, April-June 2020’. *MMWR. Morbidity and Mortality Weekly Report* 69 (37): 1288–95. https://doi.org/10.15585/mmwr.mm6937a3.

Tom Connor. 2020. *Connor Lab Nextflow Pipeline for Running the ARTIC Network’s Field Bioinformatics Tools*. https://github.com/connor-lab/ncov2019-artic-nf.

Zhang, Wenjuan, John Paul Govindavari, Brian D. Davis, Stephanie S. Chen, Jong Taek Kim, Jianbo Song, Jean Lopategui, Jasmine T. Plummer, and Eric Vail. 2020. ‘Analysis of Genomic Characteristics and Transmission Routes of Patients With Confirmed SARS-CoV-2 in Southern California During the Early Stage of the US COVID-19 Pandemic’. *JAMA Network Open* 3 (10): e2024191. https://doi.org/10.1001/jamanetworkopen.2020.24191.