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

*Dinesh Aggarwal, Richard Myers, William L Hamilton, Tehmina Bharucha, Niamh M Tumelty, Colin S Brown, Emma J Meader, Tom Connor, Darren L Smith, Declan T Bradley, Samuel Robson, Matthew Bashton, Laura Shallcross, Maria Zambon, Ian Goodfellow, Meera Chand, Justin O'Grady, M Estée Török, Sharon J Peacock, Andrew J Page, The COVID-19 Genomics UK (COG-UK) Consortium**

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

Lancet Microbe 2022; 3: e151–58

Published Online September 29, 2021 https://doi.org/10.1016/S2666-5247(21)00208-1

*Full list of Consortium names and affiliations are in the appendix

Department of Medicine (D Aggarwal MRCP, W L Hamilton PhD, M E Török FRCP, Prof S J Peacock PhD), **Cambridge University Libraries** (N M Tumelty MScEcon), **and Department of Pathology** (Prof I Goodfellow PhD), **University of Cambridge, Cambridge, UK; Public Health England, London, UK** (D Aggarwal, R Myers PhD, T Bharucha MRCP, C S Brown FRCPath, Prof M Zambon FRCPath, M Chand FRCPath)**; Cambridge University Hospital NHS Foundation Trust, Cambridge, UK** (D Aggarwal, W L Hamilton, M E Török)**; Oxford Glycobiology Institute, Department of Biochemistry, University of Oxford, Oxford, UK** (T Bharucha, C S Brown)**; Lao-Oxford-Mahosot Hospital, Wellcome Trust Research Unit, Microbiology Laboratory, Mahosot Hospital, Vientiane, Laos** (T Bharucha, C S Brown)**; Norfolk and Norwich University Hospital, Norwich, UK** (E J Meader FRCPath)**; Organisms and Environment Division, School of Biosciences, Cardiff University, Cardiff, Wales, UK** (Prof T Connor PhD)**; Public Health Wales, University Hospital of Wales, Cardiff, UK** (Prof T Connor)**; Quadram Institute Bioscience, Norwich Research Park, Norwich, UK** (Prof T Connor, Prof J O'Grady PhD, A J Page PhD)**; Hub for Biotechnology in the Built Environment, Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle upon Tyne, UK** (Prof D L Smith PhD,

Introduction

Many studies of COVID-19 in long-term care facilities (LTCFs) have reported high mortality.[1–3] Possible explanations for this finding include recognised risk factors such as increased age and comorbidities.[2,4] In England and Wales, it has been estimated that nearly 30% (15 819 of 54 325 total in the week ending Oct 16, 2020) of all deaths due to COVID-19 occurred in LTCFs[5] with outbreaks reported in 45% of all LTCFs.[6] Northern Ireland reported even higher rates; 37% (363 of 988 total in the week ending Oct 23, 2020) of deaths in LTCFs were due to COVID-19. Globally, 24·9% of superspreading events[7] were linked to LTCFs.[8] The drivers for the introduction and transmission of SARS-CoV-2 in the care sector are under investigation and are incompletely understood.

There are multiple tools available to investigate and manage SARS-CoV-2 outbreaks. These include: surveillance-based testing, where PCR testing is preferred to the serology testing of staff and residents; the testing of individuals who are symptomatic with PCR testing; the identification and self-isolation of close contacts; environmental measures such as disinfection; personal protective equipment use; and the self-isolation of individuals who test positive.[9] The genome sequencing of SARS-CoV-2 has been established as a powerful supplementary tool to characterise the transmission dynamics in health-care settings.[10–12] This method entails the investigation of the genetic relatedness of appropriately assembled SARS-CoV-2 sequences, with multiple tools available to identify clusters of infection. The genomic epidemiological investigations of outbreaks are effective in ruling out links between clusters suggested through contact tracing.[13] However, SARS-CoV-2 sequencing alone can encounter limitations, such as difficulty in proving the directionality of the spread or little genomic diversity falsely showing possible transmission events; these limitations can be overcome by integrating this information with epidemiological data.[14,15] This integration can enable the investigation of the dynamics of outbreaks within and between LTCFs and the wider community.

Several studies have used genomic epidemiology to advance the understanding of the transmission of SARS-CoV-2 within LTCFs. These studies vary in size, methods used, and quality. Here, we review the available genomic epidemiology studies on COVID-19 in LTCFs that have been done to date and provide a summary and interpretation of the key findings (table 1; appendix).

Study screening

The database searches identified 110 studies. After the removal of duplicates and the addition of papers identified through the hand-searching of preprint servers, there were 55 studies remaining. An independent review by authors of this study (AJP and NMT) of titles and abstracts identified 27 studies for full text review. After a full text review, 11 genomic epidemiology studies in LTCFs were identified for inclusion in this analysis.

Study characteristics and quality of outcome measures

Of the 11 studies included, five were done in the USA, four in the UK, and two in the Netherlands. These studies included a wide range of the number of LTCFs (1–292), participants (10–6600), and positive cases (6–1167). Six studies reported findings from the prospective surveillance of individuals and five studies reported genomic sequencing findings that occurred in relation to an outbreak investigation. The serial sampling of residents and health-care 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.[28]

M Bashton PhD); **Public Health Agency, Belfast, UK** (D T Bradley PhD); **Centre for Public Health, Queen's University Belfast, Belfast, UK** (D T Bradley); **University of Portsmouth, Centre for Enzyme Innovation, Portsmouth, UK** (S Robson PhD); **Institute of Health Informatics, University College London, London, UK** (L Shallcross PhD); **Wellcome Sanger Institute, Hinxton, Cambridge, UK** (D Aggarwal, Prof S J Peacock); **Guy's and St Thomas' NHS Foundation Trust, London, UK** (M Chand)

Correspondence to:
Dr Andrew J Page, Quadram Institute Bioscience, Norwich Research Park, Norwich, NR4 7UQ, UK
andrew.page@quadram.ac.uk

See Online for appendix

	Location	Start and end date of study (in 2020)	Type of study*	Number of LTCFs	Total number of residents and staff tested	Number of residents testing positive	Number of staff testing positive	Cases sequenced	Number of clusters†
Dautzenberg et al (2020)[16]‡	Southeast Netherlands	March–April	Surveillance	2	621	NR	133	22	3
van den Besselaar et al (2021)[17]	South Holland	May–June	Outbreak	1	425	113	56	60§	1
Hamilton et al (2021)[18]	East of England, UK	February–May	Surveillance	292	6600	1167	NR	700	409
Page et al (2021)[19]	Norfolk, UK	March–August	Surveillance	6	1035	76	9 and 3¶	89	2
Graham et al (2020)[20]	London, UK	April	Outbreak	4	383	126	3	19‖	NR
Ladhani et al (2020)[21] and Ladhani et al (2020)[22]	London, UK	April	Outbreak	6	518	105	53	99	2
Lemieux et al (2020)[23]	Boston, MA, USA	January–May	Surveillance	1	194	82	36	83	3
Zhang et al (2020)[24]	CA, USA	March–April	Surveillance	2	10	6 and 1**	3	192	1
Gallichote et al (2020)[25]‡	CO, USA	Unknown	Surveillance	5	454	NR	70	38	1
Taylor et al (2020)[26]	MN, USA	April–June	Outbreak	2	600	165	114	105	4
Arons et al (2020)[27]	WA, USA	March	Outbreak	1	89	57	26	34	2

LTCFs=long-term care facilities. NR=not reported. *Surveillance studies are defined as those which involve serial testing to identify positive cases, and outbreak investigations are those which involve the testing or sequencing, or both, of positivity after a case (or a defined number of cases) of SARS-CoV-2 have been identified. †Clusters are not uniformly defined in all papers. ‡Preprint before peer review. §Six of these samples were from an epidemiologically linked hospital outbreak. ¶Family members of a single staff member. ||Paper states both 17 and 19 samples sequenced, so it is not clear which is correct. **Family member of resident.

Table 1: **Overview of studies using SARS-CoV-2 genome sequencing of samples taken for routine surveillance or during investigation of outbreaks in LTCFs**

Nine studies sequenced both staff and patients to better understand the transmission dynamics within a LTCF. One study assessed transmission in staff alone[25] and one study in patients alone.[27] The studies were done between February and August, 2020. The study characteristics are detailed in table 2. Bioinformatics methods differed between studies, tailored to the sequencing technologies (Illumina, San Diego, CA, USA, or Oxford Nanopore Technologies, Oxford, UK) and sample preparation methods (ARTIC amplicon,[29,30] metagenomic, and whole genome sequencing), meaning that direct comparisons cannot be made because of the differing methods. Three studies were found to display deficiencies relating to the bioinformatics methods used or the results presented. These deficiencies included assembling amplicons,[20] using poor-quality sequencing data in phylogenetic analysis,[21] and imputing reference bases to replace missing bases;[25] the effect of these methods on downstream analysis is unknown.

Most studies did phylogenetic analysis of their datasets as a final step and presented the results as a dendrogram. Of the 11 included studies, the open source bioinformatics software IQ-TREE[31] was used in eight studies, PhyML[32] was used in one study, a combination of Molecular Evolutionary Genetics Analysis across computing platforms[33] software and commercial software Geneious (Biomatters, Auckland, New Zealand) was used in two studies, and the software used was not stated in one study. One study also did an analysis with Bayesian Evolutionary Analysis Sampling Trees version 2.6.2.[23] The way in which the analysis was done differed greatly, which had an effect on the granularity presented and largely prevented direct comparisons. Three studies defined a consistent method in their study to identify clusters.[18,19,27] These studies used Phylogenetic Assignment of Named Global Outbreak Lineages software or a more granular clustering algorithm in the COVID-19 Genomics UK (COG-UK) Consortium pipeline,[19] transcluster algorithm,[18] or single nucleotide polymorphism (SNP) distance.[27] One study provides a rationale for how clusters were identified without providing a defined criteria for selecting genomic clusters.[23]

Furthermore, assuming a mutation rate of approximately 2·5 SNPs per month[19] allows for the estimation of the amount of variation expected in a phylogeny at any particular timepoint in a series. Additionally, only two studies mention the use of negative controls[18,19] and no studies have released the sequencing reads found in the negative controls publicly. Many sample preparation protocols use amplification techniques that can also amplify contamination and give false results, particularly

	Location	Sample preparation method	Sequencing	Method of genome construction strategy	Software used to infer phylogenetic trees	Data availability*
Dautzenberg et al (2020)[16]	Southeast Netherlands	Amplicon	Nanopore	Consensus	NR	Not available
van den Besselaar et al (2021)[17]	South Holland	Amplicon	Nanopore	Consensus	IQ-TREE	Not available
Hamilton et al (2021)[18]	East of England, UK	Amplicon†	Nanopore or Illumina	Consensus	IQ-TREE and PhyML	Available but not linked
Page et al (2021)[19]	Norfolk, UK	Amplicon†	Illumina	Consensus	IQ-TREE	Available
Graham et al (2020)[20]	London	Amplicon†	Illumina	Reference-guided assembly	IQ-TREE	Not available
Ladhani et al (2020)[21]	London	Whole genome sequencing	Illumina	Consensus	IQ-TREE	Available but not linked
Lemieux et al (2020)[23]	Boston, MA, USA	Metagenomic	Illumina	Reference-guided assembly	IQ-TREE and Bayesian Evolutionary Analysis Sampling Trees	Available
Zhang et al (2020)[24]	CA, USA	Metagenomic	Illumina	Consensus	IQ-TREE	Available
Gallichote et al (2020)[25]	CO, USA	Amplicon†	Illumina	Consensus gap filled with reference	Geneious	Not available
Taylor et al (2020)[26]	MN, USA	Amplicon†	NR	NR	IQ-TREE	Available but not linked
Arons et al (2020)[27]	WA, USA	NR	Nanopore	Consensus	Geneious	Available

NR=not reported. The companies for the sequencing methods are: Nanopore, Oxford Nanopore Technologies, Oxford, UK, and Illumina, San Diego, CA, USA. *If data are present in the Global Initiative on Sharing Avian Influenza Data or the International Nucleotide Sequence Database Collaboration database they are labelled as available, and when there is no linkage information between the samples used in the article and the data in the public archives, they are labelled as not linked. †Amplicon sequencing uses the ARTIC protocol.[29]

Table 2: **Sequencing and bioinformatics methods used in the long-term care facilities genomic epidemiology studies**

when the viral load in the source material is low.[18,19,20,25] Without having all underlying raw data (including controls) alongside the sample preparation and sequencing metadata,[34] reanalysis and comparison between studies is difficult and prone to error.

Many studies publicly release their raw sequencing data or consensus or assembled genomes, or a combination,[18,19,23,24] through the International Nucleotide Sequence Database Collaboration (INSDC)[35] and the Global Initiative on Sharing Avian Influenza Data (GISAID).[36] This sharing allows for independent reanalysis, overcoming the effect of variation among methods between studies.

However, in genomics it is a common poor practice to not release data publicly or to provide insufficient metadata, such as accession numbers, thus making reanalysis unfeasible. The use of open metadata standards that are internationally agreed upon for SAR-CoV-2 genomics enables genomic epidemiology on a global scale. In some cases, there are legitimate reasons to withhold data, such as to maintain patient confidentiality where identification might be possible (eg, as a result of small sample numbers or the location of LTCFs). In other cases, even if the authors wish to deposit all clinically important samples, some of their samples might 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 a low viral load often sequence poorly, leading to incomplete datasets and making reanalysis impossible. Researchers might have a well meaning desire to make publicly available data that does not meet the stringent quality-control thresholds by imputing missing data from a reference genome,[25] a common technique in human genetics, but this leads to erroneous results in the 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 challenge by making these data available on their website and through the INSDC, which has lower quality-control thresholds than GISAID (>50% genome completeness). Convergence on a small number of open-source bioinformatics workflows using best practices should mitigate future issues in this regard (eg, a Nextflow pipeline with a focus on COVID-19, one global resource on the Galaxy platform for the analysis of SARS-CoV-2 data, and a workspace on the Terra app with COVID-19 genomic data and workflows.

For the **COG-UK Consortium**, see https://www.cogconsortium.uk/

For the **Nextflow pipeline**, see https://github.com/connor-lab/ncov2019-artic-nf

For the **Galaxy resource**, see https://covid19.galaxyproject.org

For the **Terra app resource**, see https://app.terra.bio/#workspaces/pathogen-genomic-surveillance/COVID-19

Summary of findings

To date, genomic epidemiology studies of SARS-CoV-2 in LTCFs provide many insights into transmission in this susceptible population. The diversity of studies ranged from outbreak investigations with detailed epidemiological data in single LTCFs to the prospective surveillance of hundreds of LTCFs, as summarised earlier. Key findings from the included studies are shown in the panel.

Large outbreaks in LTCFs, such as in the study by Lemieux and colleagues,[23] generally shared the same characteristics: a single cluster with rapid expansion, resulting in most samples being identical or near identical (only a one SNP difference). Residents and staff, including staff who had no contact with residents, were usually infected with the same (identical) genome sequence. The direction of transmission cannot be established from genomic data alone, but the addition of traditional epidemiological data (such as sample dates and the co-location of individuals) might allow inferences to be drawn. In many outbreaks more than one cluster was observed,[18] but these sporadic introductions usually represented only a few facilities. The temporal analysis of genomic data allows estimation of when an introduction into a LTCF is likely to have occurred, making genomics useful for providing an estimate for when an outbreak began.[23] The paper by Lemieux and colleagues[23] in Boston, MA, USA, estimated that, after an introduction, 85% of residents were infected within 2–3 weeks, despite extensive infection prevention and control measures being in place. By the time two symptomatic individuals are identified in a LTCF, the outbreak is likely to already be widespread.[21,27]

An analysis of lineages circulating in a region compared with lineages found within LTCFs[19,23] show that there is little diversity within LTCFs, indicating a small number of introductions rather than repeated introductions from the community. Distinct clusters are usually (but not always) seen between LTCFs,[18] with genomics identifying a small number of shared genomic clusters in different LTCFs.[18,19,21] Taking the sequence diversity found in 292 LTCFs in a region[18] as a whole and comparing it to a similar number of residents not in a LTCF in the same region, similar numbers of SNP differences were identified in the genomes (the median number of SNP differences in residents in a LTCF was eight, and in residents not in a LTCF was nine). When looking at a single LTCF,[19] knowing the diversity of the circulating lineages within the locality helped to rule out local inward transmission.

Looking more closely at the dynamics of an outbreak, the study by Arons and colleagues[27] in WA, USA, 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.[19,21] In one study,[18] two LTCFs located within 1 km of each other had residents infected with identical genomes; a paramedic who visited both facilities also tested positive. In another study[19] a genetically distinct sub-lineage was found in six different LTCFs within one small region. Genomics reveals that the 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-health-care workers, are usually the same. A high percentage of asymptomatic individuals is common, with staff usually accounting for a higher percentage of those who are asymptomatic. 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.[26] Even with intensive consecutive testing every week, enhanced infection control, and the transfer of residents who test positive to dedicated isolation units, the outbreak continued in the study in MN, USA, with the genomically similar clusters found over an extended period of time.[26]

***Panel:* Key findings**

Community or hospital acquisition of SARS-CoV-2 in residents of LTCFs

1 Most LTCF infections were community-acquired (moderate).[18,19]
2 Approximately 6% of residents with SARS-CoV-2 infection in LTCFs had suspected or confirmed hospital-acquired infections in one UK region (moderate).[18]
3 Little genomic diversity among the SARS-CoV-2 infections in staff and residents from the same LTCFs. This finding indicated a small number of introductions rather than a series of seeding events from the community (weak).[16,19,21,23,26]
4 Shared clusters between separate LTCFs could be identified (moderate).[18,19,21]

Transmission and outcomes within LTCFs

5 The use of genomic data allowed independent clusters of infections to be identified within LTCFs (strong).[17,18,20,21,23,26,27]
6 In LCTF outbreaks, initial sequencing was useful to identify whether genomes were similar, but the subsequent sequencing of large numbers of samples did not add much value (moderate).[23,26]
7 Once two symptomatic individuals were identified in a LTCF, the outbreak was already widespread (moderate).[27]
8 The sequencing of samples taken consecutively from the same residents of LTCFs showed that viral lineages persisted over an extended period of time despite infection prevention and control measures. It also showed that the sequencing technique was reproducible (moderate).[26]
9 Residents of LTCFs were more likely to be infected with identical genome sequences if their bedrooms were in close proximity to each other (moderate).[27]
10 The mortality rate among residents of LTCFs was high in all facilities, with no link to particular lineages (moderate).[20,21,26]
11) The temporal analysis of genomic data allows for the estimation of when an introduction was likely to have occurred (moderate).[23]

Reproducibility of genomic analysis

12 The genomic studies reviewed commonly misapplied bioinformatics methods (strong).[20,21,25]
13 Minimum quality thresholds set by public archives on SARS-CoV-2 data limit data availability and reproducibility (moderate).[18,19,21]
14 Most studies did not provide adequate epidemiological data or metadata to allow analysis to be reproduced (strong).[16–18,20,21,25,26]

Strength of findings: strong indicates multiple sources of evidence, supported by in-depth analysis or experiments; moderate indicates one or more sources of evidence, supported by analysis or experiments; weak indicates one or more sources of evidence that are potentially contradictory. LTCFs=long-term care facilities.

The intensive sequencing of all residents and staff in an outbreak does not appear to provide additional genomic information after the first few sequences.[17] The strategic subsampling of staff and residents should be adequate to understand the number of clusters and their relative proportions. However, inadequate sampling does have a large effect on the usefulness of genomics.[25] Genomics has reduced usefulness once there is a large outbreak; however, it does provide useful information about how SARS-CoV-2 might enter a care home, such as via staff or patient movements,[18] and continued ongoing monitoring using genomics can identify new sources of infection (new seeding events), which can help to inform policy.[18,27] Because visitors were restricted from visiting LTCFs early in the pandemic, no data are available on their role as a source of introduction.

Genomes sequenced through prospective surveillance have proven to be useful for identifying linked outbreaks that might have been missed otherwise.[10,18,19,24] The limitation is that it might 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.[18,19]

Residents of LTCFs who develop severe COVID-19 are often admitted to hospital, which might be the first indication of an outbreak. Samples that are sequenced as part of surveillance studies can provide early insight into outbreaks in LTCFs.[18,19,23,24] 5·8% of COVID-19 infections in residents of LTCFs were suspected to be acquired in hospital.[18] 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. These findings have important implications for infection control in LTCFs and for public health policies.[18]

So far, most genomic epidemiology studies of SARS-CoV-2 in LTCFs have been done 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 endeavour 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 might be different.

Recommendations

It is clear from the studies summarised here that genomics play a crucial role in understanding the transmission dynamics within LTCFs. 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 3.

All staff working in a LTCF (regardless of their role) should be treated as a single cohort and subject to infection prevention and control measures that are uniform, including the appropriate use of personal protective equipment, regular screening for SARS-CoV-2, and genome sequencing of any positive samples. Genome sequencing has shown 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. The early identification and exclusion of asymptomatic staff through regular screening might reduce the risk of transmission to residents and other staff. However, it should be noted that the regular screening of staff for asymptomatic infections might still sometimes be unable to identify an individual who is infectious, and that could lead to a superspreading event.[37]

Sequencing every genome in an outbreak is not recommended because it provides rapidly diminishing returns. Instead, the strategic sequencing of a subset of

	Point from Key findings panel	Effect of these measures
Transmission of SARS-CoV-2		
Limiting the spread of SARS-CoV-2 between hospitals, health-care workers, and residents of LTCFs is an urgently needed infection control measure and public health priority	2–5, 8–10	Control transmission
All staff, not just individuals with direct contact with residents, should be treated as one cohort and subject to the same infection prevention and control measures	3	Control transmission
Genomics identifies transmission between staff, between staff and residents, and between care facilities. Findings should direct future control measures	2–4	Control transmission
Clustering based on physical proximity to the bedroom of a resident infected with SARS-CoV-2 supports its use as an additional factor to identify at-risk individuals and prioritise testing	9	Control transmission and resource allocation
LTCF sequencing strategy		
A targeted approach weighted towards sequencing early positive samples in an outbreak coupled with potential epidemiological links can help to highlight the source of introduction; widespread sequencing within a care home is unlikely to yield substantially more information	3–4, 6, 8	Control transmission and resource allocation
Genomic surveillance in a proportion of samples from LTCFs should be done including both patients and staff, allowing the genomic epidemiology of a LTCF to be put into context	3–4, 6–8, 11	Control transmission and resource allocation
Residents with a recent hospital admission who subsequently test positive should have their genome sequenced to identify the hospital seeding of outbreaks in LTCFs	2, 5	Control transmission and resource allocation
Ongoing community surveillance with SARS-CoV-2 sequencing allows outbreaks in LTCFs to be better characterised	1–2, 3–4	Control transmission and resource allocation
Recommendations for future research		
Modelling of subsampling strategies within LTCFs is needed to optimally use genomic surveillance	6	Control transmission and research need
Epidemiological and genomic data should be released to public archives with sufficient metadata to enable genomic epidemiology	13–14	Control transmission and research need
Appropriate and validated bioinformatics methods should be applied to genomic analysis with domain experts reviewing results to avoid erroneous results	12	Control transmission and research need
A focus on rapid integrated epidemiological and genomic analysis will have the most clinical benefit	4–5, 7–10, 14	Control transmission and resource allocation

LTCFs=long-term care facilities.

Table 3: **Recommendations for measures derived from the use of SARS-CoV-2 genomics in LTCFs**

Search strategy and selection criteria

The studies we included in this Review were identified by searches of PubMed, Web of Science, and Scopus from Jan 1 to Nov 3, 2020. We used the search terms ("COVID-19" OR "SARS-CoV-2") and ("long term care facility", "care home", "skilled nursing facility", "nursing home" or "residential home") and ("sequenc*" or "genom*" or "WGS") to identify relevant English-language publications and preprints since January, 2020. Because of the limitations of the systematic search functionality on medRxiv and bioRxiv, these servers were hand-searched for additional papers that met the inclusion criteria. We focused particularly on studies where genomic epidemiology was used to enhance the interpretation of outbreaks. Articles that did not use genomic sequencing as a method were excluded. Studies were screened by authors NMT and AJP. The subsequent reported outcomes were extracted (where documented): the location of the study, the time period, number of long-term care facilities involved, the total number of positive cases broken down by staff and residents, the total number of genomes sequenced, the wider effect, the sample preparation methods, the sequencing equipment used, the genome creation method, the phylogeny, and the data availability. We report on the practical difficulties if reanalysis was attempted, and do not formally attempt to reproduce the analyses presented. Papers were classified as follows: outbreak investigation, surveillance, first case, and genomic reanalysis of public data. Because this topic is a rapidly emerging field of research, preprints were included but it should be noted that these are not peer-reviewed.

samples should be undertaken. The strategy for sequencing positive samples should be weighted towards staff rather than residents because they are at risk of community acquisition and subsequent transmission, whereas residents are less likely to have external contact. The modelling of subsampling strategies within LTCFs is needed.

Once a resident in a LTCF tests positive, other residents with bedrooms in close proximity should be considered to be at a high risk of infection, regardless of contact patterns and other infection control measures, because 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 residents in LTCFs, health-care workers, and hospitals should be a key target for infection control and prevention.

Raw sequencing data and consensus or assembled genomes should be made available in the public archives in a timely manner with the internationally recommended minimal set of metadata to enable genomic epidemiological analysis at local, national, and international levels.[34] 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 reduce the probability of missing one of these domain-specific considerations, we recommend the use of validated and tested SARS-CoV-2 pipelines, with the involvement of domain-specific experts to assist with the analysis and review of results.

A follow-up of the study done in London, UK, by Ladhani and colleagues[38] using serological testing showed that by 5 weeks, most individuals had seroconverted, including 66·4% of staff and 67·0% of residents who were asymptomatic and tested negative by RT-PCR.[38] This finding highlights the need to combine various surveillance methods, including genomic epidemiology, to accurately characterise the dynamics of transmission within LTCFs; this is planned in a prospective study across 105 care homes in the UK.[39]

Ultimately, genomics provides the most clinical benefit and insight if it is integrated with detailed epidemiological data in a timely fashion. Meredith and colleagues[10] established weekly infection prevention control meetings, combining phylogenetic analysis to assist outbreak investigation and contact-tracing efforts at a health-care facility. Furthermore, although the routine genomic surveillance of hospital patients and staff, residents and staff at LTCFs, and community cases will provide greater insight into transmission dynamics, integrating additional epidemiological information such as hospital discharges, patient movements, and discharge locations would provide a much more informative approach. 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 for leveraging the clinical benefits of sequencing.

Conclusions

We have presented findings from multiple genomic epidemiology studies of transmission in LTCFs in an evolving pandemic. We have amalgamated the data to provide clinical recommendations from the findings and recommendations for refining methods for such studies in the future. Genomics can help to 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, indicating nosocomial infection. Placing these outbreaks in the context of the wider circulating lineages

in the locality also provides information about routes of transmission. For example, this method can separate local community transmission from other routes of transmission, which informs policy and helps to limit future outbreaks. The genome sequencing of SARS-CoV-2 has been proven to provide useful insights into the transmission and dynamics of outbreaks. Prospective genomic surveillance provides a backbone of information, helping to inform outbreak analysis. Hidden transmission links are uncovered using genomics that help with the interpretation of epidemiology and with contract-tracing efforts. Consecutive sampling provides yet more insight 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 is done repeatedly. The ability to integrate epidemiological and genomic analysis in a clinically actionable timeframe is a major challenge to realising the clinical benefits of genomics.

Contributors

All authors read the manuscript and consented to its publication. AJP led the Review and wrote the first draft of the manuscript. DA and AJP extensively re-drafted the manuscript. DA did enhanced analysis for the UK studies and provided clinical oversight. NMT, AJP, and DA contributed to the literature search. WLH, IG, MET, and DA provided additional insight into the east of England study (by Hamilton and colleagues). LS provided insight into the study by Krutikov and colleagues. JO'G, AJP, and EJM provided insight into the Norfolk study (by Page and colleagues). TC, MC, TB, CSB, MZ, DTB, MB, DLS, and SR provided public health insight and guidance. RM undertook reanalysis of most UK studies. SJP instigated the Review and provided overall leadership.

Declaration of interests

We declare no competing interests.

Acknowledgments

We thank members of the COVID-19 Genomics UK Consortium for their contributions to generating the data used in some of these studies. We thank Judith Pell for critically assessing and improving this manuscript. DA is a Clinical PhD Fellow and gratefully supported by the Wellcome Trust (grant number 222903/Z/21/Z). AJP and JO'G 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 (project number BB/R012504/1) and its constituent project (project number BBS/E/F/000PR10352), and also the Quadram Institute Bioscience BBSRC-funded Core Capability grant (project number BB/CCG1860/1). SR, DLS, and MB gratefully acknowledge support from Research England's Expanding Excellence in England Fund. IG is a Wellcome Senior Fellow and supported by the Wellcome Trust (grant number 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 National institutes of Health Research Cambridge Biomedical Research Centre. The COVID-19 Genomics UK Consortium is supported by funding from the Medical Research Council part of UK Research & Innovation and the National Institute of Health Research and Genome Research, 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. The UK studies were done as part of surveillance for COVID-19 infections under the auspices of Section 251 of the National Health Service Act 2006 or Regulation 3 of The Health Service (Control of Patient Information) Regulations 2002, or both. They therefore did not require individual patient consent or ethical approval. The COVID-19 Genomics UK Consortium study protocol was approved by the Public Health England Research Ethics Governance Group (reference number R&D NR0195).

References

1 Abrams HR, Loomer L, Gandhi A, Grabowski DC. Characteristics of U.S. nursing homes with COVID-19 cases. *J Am Geriatr Soc* 2020; **68:** 1653–56.

2 Fisman DN, Bogoch I, Lapointe-Shaw L, McCready J, Tuite AR. Risk factors associated with mortality among residents with coronavirus disease 2019 (COVID-19) in long-term care facilities in Ontario, Canada. *JAMA Netw Open* 2020; **3:** e2015957.

3 Burton JK, Bayne G, Evans C, et al. Evolution and impact of COVID-19 outbreaks in care homes: population analysis in 189 care homes in one geographic region. *medRxiv* 2020; published online July 10. https://doi.org/10.1101/2020.07.09.20149583 (preprint).

4 Jordan RE, Adab P, Cheng KK. Covid-19: risk factors for severe disease and death. *BMJ* 2020; **368:** m1198.

5 Office for National Statistics. Deaths registered weekly in England and Wales, provisional: week ending 16 October 2020. Oct 27, 2020. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregisteredweeklyinenglandandwalesprovisional/weekending16october2020 (accessed Nov 3, 2020).

6 Public Health England. COVID-19: number of outbreaks in care homes—management information. April 29, 2020. https://www.gov.uk/government/statistical-data-sets/covid-19-number-of-outbreaks-in-care-homes-management-information (accessed Nov 3, 2020).

7 Al-Tawfiq JA, Rodriguez-Morales AJ. Super-spreading events and contribution to transmission of MERS, SARS, and SARS-CoV-2 (COVID-19). *J Hosp Infect* 2020; **105:** 111–12.

8 Swinkels K. SARS-CoV-2 superspreading events database. Jun 12, 2020. https://kmswinkels.medium.com/covid-19-superspreading-events-database-4c0a7aa2342b (accessed Nov 3, 2020).

9 Centers for Disease Control and Prevention. Testing residents. Jan 7, 2021. https://www.cdc.gov/coronavirus/2019-ncov/hcp/nursing-homes-testing.html (accessed May 25, 2021).

10 Meredith LW, Hamilton WL, Warne B, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis* 2020; **20:** 1263–71.

11 Lucey M, Macori G, Mullane N, et al. Whole-genome sequencing to track severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission in nosocomial outbreaks. *Clin Infect Dis* 2020; **71:** e727–35.

12 Paltansing S, Sikkema RS, de Man SJ, Koopmans MPG, Oude Munnink BB, de Man P. Transmission of SARS-CoV-2 among healthcare workers and patients in a teaching hospital in the Netherlands confirmed by whole-genome sequencing. *J Hosp Infect* 2021; **110:** 178–83.

13 Aggarwal D, Warne B, Jahun A S, et al. Genomic epidemiology of SARS-CoV-2 in a UK university identifies dynamics of transmission. *Res Sq* 2021; published online May 19. https://doi.org/10.21203/rs.3.rs-520627/v1 (preprint).

14 Deng X, Gu W, Federman S, et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* 2020; **369:** 582–87.

15 Villabona-Arenas CJ, Hanage WP, Tully DC. Phylogenetic interpretation during outbreaks requires caution. *Nat Microbiol* 2020; **5:** 876–77.

16 Dautzenberg M, Eikelenboom-Boskamp A, Drabbe M, et al. Healthcare workers in elderly care: a source of silent SARS-CoV-2 transmission? *medRxiv* 2020; published online Sept 9. https://doi.org/10.1101/2020.09.07.20178731 (preprint).

17 van den Besselaar JH, Sikkema RS, Koene FMHPA, et al. Are presymptomatic SARS-CoV-2 infections in nursing home residents unrecognized symptomatic infections? Sequence and metadata from weekly testing in an extensive nursing home outbreak. *Age Ageing* 2021; published online May 7. https://doi.org/10.1093/ageing/afab081.

18 Hamilton WL, Tonkin-Hill G, Smith ER, et al. Genomic epidemiology of COVID-19 in care homes in the east of England. *eLife* 2021; **10:** e64618.

19 Page AJ, Mather AE, Le-Viet T, et al. Large-scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management. *Microb Genom* 2021; **7**.

20 Graham NSN, Junghans C, Downes R, et al. SARS-CoV-2 infection, clinical features and outcome of COVID-19 in United Kingdom nursing homes. *J Infect* 2020; **81:** 411–19.

21 Ladhani SN, Chow JY, Janarthanan R, et al. Investigation of SARS-CoV-2 outbreaks in six care homes in London, April 2020. *EClinicalMedicine* 2020; **26:** 100533.

22 Ladhani SN, Chow JY, Janarthanan R, et al. Increased risk of SARS-CoV-2 infection in staff working across different care homes: enhanced CoVID-19 outbreak investigations in London care homes. *J Infect* 2020; **81:** 621–24.

23 Lemieux JE, Siddle KJ, Shaw BM, et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* 2020; **371:** eabe3261.

24 Zhang W, Govindavari JP, Davis BD, et al. 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 Netw Open* 2020; **3:** e2024191.

25 Gallichote EN, Quicke KM, Sexton NR, et al. Longitudinal surveillance for SARS-CoV-2 RNA among asymptomatic staff in five Colorado skilled nursing facilities: epidemiologic, virologic and sequence analysis. *medRxiv* 2020; published online Nov 5. https://doi.org/10.1101/2020.06.08.20125989 (preprint).

26 Taylor J, Carter RJ, Lehnertz N, et al. 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 Morb Mortal Wkly Rep* 2020; **69:** 1288–95.

27 Arons MM, Hatfield KM, Reddy SC, et al. Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. *N Engl J Med* 2020; **382:** 2081–90.

28 Rambaut A, Holmes EC, Hill V, et al. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv* 2020; published online April 19. https://doi.org/10.1101/2020.04.17.046086 (preprint).

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

30 Farr B, Rajan D, Betteridge E, et al. COVID-19 ARTIC v3 Illumina library construction and sequencing protocol V.3. May 22, 2020. https://doi.org/10.17504/protocols.io.bgq3jvyn (accessed Nov 3, 2020).

31 Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020; **37:** 1530–34.

32 Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 2010; **59:** 307–21.

33 Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol Biol Evol* 2018; **35:** 1547–49.

34 Griffiths EJ, Timme RE, Page AJ, et al. The PHA4GE SARS-CoV-2 contextual data specification for open genomic epidemiology. *Preprints (Basel)* 2020; published online Aug 9. https://doi.org/10.20944/preprints202008.0220.v1 (preprint).

35 Cochrane G, Karsch-Mizrachi I, Takagi T. Sequence database collaboration IN. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res* 2016; **44:** D48–50.

36 Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* 2017; **22:** 30494.

37 Bedford T, Logue JK, Han PD, et al. Viral genome sequencing places White House COVID-19 outbreak into phylogenetic context. *medRxiv* 2020; published online Nov 13. https://doi.org/10.1101/2020.10.31.20223925 (preprint).

38 Ladhani SN, Jeffery-Smith A, Patel M, et al. High prevalence of SARS-CoV-2 antibodies in care homes affected by COVID-19: prospective cohort study, England. *EClinicalMedicine* 2020; **28:** 100597.

39 Krutikov M, Palmer T, Donaldson A, et al. Study protocol: understanding SARS-Cov-2 infection, immunity and its duration in care home residents and staff in England (VIVALDI). *Wellcome Open Res* 2021; **5:** 232.