Epidemiological analysis of Legionnaires' disease in Scotland: a genomic study

Jamie Gorzynski, Bryan Wee*, Melissa Llano, Joana Alves, Ross Cameron, Jim McMenamin, Andrew Smith, Diane Lindsay, J Ross Fitzgerald*

Summary

Background ***Legionella pneumophila*** **is the main cause of a severe pneumonic illness known as Legionnaires' disease and is a global public health threat. Whole-genome sequencing (WGS) can be applied to trace environmental origins of *L pneumophila* infections, providing information to guide appropriate interventions. We aim to explore the evolutionary and epidemiological relationships in a 36-year Scottish *L pneumophila* reference isolate collection.**

Methods **We investigated the genomic epidemiology of Legionnaires' disease over 36 years in Scotland, comparing genome sequences for all clinical *L pneumophila* isolates (1984–2020) with a sequence dataset of 3211 local and globally representative isolates. We used a stratified clustering approach to capture epidemiological relationships by core genome Multi-locus Sequence Typing, followed by high-resolution phylogenetic analysis of clusters to measure diversity and evolutionary relatedness in context with epidemiological metadata.**

Findings **Clustering analysis showed that 111 (57·5 %) of 193 of *L pneumophila* infections in Scotland were caused by ten endemic lineages with a wide temporal and geographical distribution. Phylogenetic analysis of *L pneumophila* identified hospital-associated sublineages that had been detected in the hospital environment up to 19 years. Furthermore, 12 (30·0%) of 40 community-associated infections (excluding a single, large outbreak) that occurred over a 13 year period (from 2000 to 2013) were caused by a single widely distributed endemic clone (ST37), consistent with enhanced human pathogenicity. Finally, our analysis revealed clusters linked by national or international travel to distinct geographical regions, indicating several previously unrecognised travel links between closely related isolates (fewer than five single nucleotide polymorphisms) connected by geography.**

Interpretation **Our analysis reveals the existence of previously undetected endemic clones of *L pneumophila* that existed for many years in hospital, community, and travel-associated environments. In light of these findings, we propose that cluster and outbreak definitions should be reconsidered, and propose WGS-based surveillance as a critical public health tool for real-time identification and mitigation of clinically important endemic clones.**

Funding **Chief Scientist Office, Biotechnology and Biological Sciences Research Council (UK), Medical Research Council Precision Medicine Doctoral Training Programme, Wellcome Trust, and Medical Research Council (UK).**

Lancet Microbe 2022; 3: e835–45

October 11, 2022
https://doi.org/10.1016/S2666-5247(22)00231-2

*Contributed equally

The Roslin Institute, University of Edinburgh, Edinburgh, UK (J Gorzynski MSc, B Wee PhD, J Alves PhD, Prof J R Fitzgerald PhD); **Scottish Microbiology Reference Laboratory, Glasgow Royal Infirmary, Glasgow, UK** (Prof A Smith FRCpath, D Lindsay PhD); **Public Health Scotland, Glasgow, UK** (M Llano MPH, R Cameron MPH, J McMenamin MBChB); **College of Medical, Veterinary & Life Sciences, Glasgow Dental Hospital and School, University of Glasgow, Glasgow, UK** (Prof A Smith)

Correspondence to:
Prof J Ross Fitzgerald, The Roslin Institute, University of Edinburgh, Edinburgh EH25 9RG, UK
ross.fitzgerald@ed.ac.uk

Introduction

Legionella is a genus of globally ubiquitous, intracellular, freshwater bacteria that naturally coexists with protozoa. More than 30 *Legionella* species can invade immune cells and cause opportunistic human respiratory infections.[1] However, over 90% of infections are caused by a single species, *Legionella pneumophila,* which is typically inhaled from contaminated environmental aerosols.[2] Legionellosis varies from a mild flu (Pontiac fever) to Legionnaires' disease, an atypical, often severe form of pneumonia with a 10% mortality rate.[1] Most Legionnaires' disease cases are sporadic and unlinked to other infections, with many acquired during travel. However, outbreaks of Legionnaires' disease occur when a cluster of infections are linked to a common source such as plumbing, air-conditioning systems, or industrial cooling towers.[3] Additionally, commercial compost or soil can be a reservoir for outbreaks of Legionnaires' disease caused by *Legionella longbeachae*.[4] *L pneumophila* bacteria are challenging to eradicate from environmental reservoirs and strains might persist in these environments, posing a threat to public health.[5]

To identify the source of outbreaks or sporadic infections, clinical and environmental isolates are compared on the basis of knowledge of possible exposures.[3] Legionnaires' disease cases are typically classified as either travel-associated Legionnaires' disease (TALD), hospital-associated Legionnaires' disease (HALD), or community-acquired Legionnaires' disease (CALD), depending on the patient's history of potential exposure in the 10 days before disease onset.[6] Although sequence-based typing is considered the current gold standard method for discrimination of *L pneumophila*,[7] whole-genome sequencing (WGS) methods have a much higher level of resolution to distinguish closely related isolates and to determine the source of infections.[7,8] Applied retrospectively, WGS has been used to resolve the source of Legionnaires' disease outbreaks that remained ambiguous with standard

Research in context

Evidence before this study

We conducted a literature search with Google Scholar for genomic studies investigating the epidemiology of Legionnaires' disease between Jan 1, 2010, and Jan 1, 2021, using the search terms "Legionella", "epidemiology", "WGS", and "genomic". No language restrictions were applied to this search. Since 2012, whole-genome sequencing (WGS) has been employed in numerous Legionnaires' disease outbreak investigations to establish the most probable source of infections, including several that remained ambiguous with standard molecular methods. However, no studies within these dates had investigated the broader epidemiology of Legionnaires' disease over extended timeframes or broad geographical areas, with large numbers of environmental isolates. Two studies compared WGS from separate outbreaks over long time periods but their analysis was limited to specific municipal areas. Another study investigated nosocomial infection by a single sequence type using a spatially and temporally varied dataset but only a small number of isolates from non-hospital environmental sources were included. We did not find any WGS studies comparing all historic infections on a national scale and no large-scale WGS studies investigating travel-associated cases.

Added value of this study

In this study, we conducted WGS analysis on a 36-year national collection of reference isolates that included isolates from all Scottish patients with Legionnaires' disease from 1984 to 2020 (n=193) and 229 environmental isolates from a range of environmental reservoirs. We also included 2944 genomes from global sources (The National Center for Biotechnology Information). To our knowledge, this is the largest dataset used for genomic epidemiological analysis of Legionnaires' disease to date and includes a uniquely high number of environmental isolates that provide important context for investigating probable origins of infectious strains. This study provides new insights into the cause of Legionnaires' disease, identifying the major *Legionella pneumophila* clones that have been sampled in hospital and municipal plumbing systems and other man-made water systems in Scotland, which have been the cause of a major proportion of Legionnaires' disease for decades.

Implications of all the available evidence

Our analysis indicates the existence of previously unrecognised endemic clones of *L pneumophila* that existed for many years in Scottish hospital-associated, community-associated, and travel-associated environments. In light of these findings, we propose that cluster and outbreak definitions should be reconsidered. Furthermore, we propose that routine regular environmental sampling should be used to facilitate near-real-time WGS-based identification of epidemiological links, attribution of outbreak sources, and to inform public health measures targeting endemic clones that are an ongoing threat to public health.

molecular methods, and provides enhanced resolution to cases affected by recent genetic recombination events that complicate phylogenetic reconstruction.[8,9]

It has been observed that *L pneumophila* isolates from infections in distinct geographical locations or occurring years apart can have a nearly identical genotype.[10,11] As such, the genetic similarity of epidemiologically related and unrelated isolates might overlap, presenting an ongoing challenge for source attribution.[7] Conversely, *L pneumophila* associated with a single outbreak can represent multiple genotypes coexisting in an infection source or the result of strain diversification over time in the environment prior to the outbreak.[11–13] Accordingly, definitive source attribution is not always possible but a high confidence of prediction may be achieved by WGS of extensive environmental samples.[10,11]

See Online for appendix 1

The frequency of Legionnaires' disease infection is increasing and our understanding of the epidemiology of infections remain unclear,[14] particularly the relationship between community-associated, hospital-associated, and travel-associated clinical and environmental *L pneumophila*. Although clustering of sporadic legionellosis cases in Scotland has been observed from epidemiological data, to date, only a single outbreak has been investigated using a WGS approach.[12,15] Here, we aimed to examine the cause and environmental origins of historical episodes of Legionnaires' disease in Scotland, using WGS of all clinical isolates and a selection of environmental *L pneumophila* isolates from 1984 to 2020, in combination with epidemiological data provided by routine Legionnaires' disease surveillance.

See Online for appendix 2

Methods

Culture collection and whole-genome sequencing

In total, we analysed quality-filtered genome sequences for 3397 *L pneumophila* isolates, including 453 obtained from our reference isolate collection and 2944 representative global isolates from 25 countries obtained from the National Center for Biotechnology Information (NCBI). Publicly available WGS reads for global isolates were downloaded (May 20, 2020) and assembled as described in appendix 1 (p 10). Previously assembled *Legionella* sequences in Genbank or RefSeq were also downloaded (Feb 18, 2020) and the collective data were passed through a quality-filtering pipeline (appendix 1 p 10) to remove highly divergent sequences, poor-quality assemblies, redundant copies (from different databases), and assemblies with a high proportion of sequences not from *L pneumophila* (evidence of culture contamination). A complete list of the quality-filtered isolates, with corresponding accession numbers and countries of isolation is included in appendix 2.

The reference collection isolates included all those collected from patients in Scotland with Legionnaires'

disease between 1984 and 2020 (n=193) and 229 environmental isolates from the Scottish reference laboratory archive over the same timeframe, which we selected to represent the breadth of spatial and temporal diversity of environmental isolates in the archive. The archive contains isolates that were collected as part of outbreak investigations, enhanced surveillance, and general submissions from Scottish water testing laboratories and included 16 isolates from ships or from outside Scotland. We previously sequenced 431 Scottish isolates (NCBI BioProject Accessions: PRJEB31628 [n=395], PRJEB6631 [n=25], and PRJEB1828 [n=11]),[16] and 22 (5%) were newly sequenced for the current study, with reads uploaded to the NCBI Short Read Archive under the BioProject accession PRJEB50423.

Core-genome multilocus sequence typing (cgMLST) and cluster definition

Adapting a previously published cgMLST scheme for *L pneumophila*,[17] profiles were assigned to the quality-filtered data using ChewBBACA (v2.0.17.2),[18] including appropriate controls for atypical gene calls such as missing loci (appendix 1 p 11). Genomes were clustered hierarchically by the number of core gene variants between cgMLST profiles in R (median-link, v3.6.3), inferred using 1469 loci present in at least 95% of the data (see gene list in appendix 2). cgMLST clusters were defined as clusters of isolates that differ at less than 115 cgMLST loci, determined using the max-clade method of TreeCluster (v1.0.0).[19] This conservative threshold was chosen to identify clusters of isolates that might have diverged over long timescales, informed by the findings of the outbreak in Edinburgh in 2012.[12] This outbreak comprised a diverse infecting population of ST191 subtypes that had been persisting in the environment for a long time before the outbreak[12] and contained strains that differed at up to 115 cgMLST loci. We then performed phylogenetic analysis of each cgMLST cluster to identify phylogenetic clades (subclusters), in which the majority of isolates shared epidemiological metadata.

Phylogenetic analysis

Maximum-likelihood core single-nucleotide polymorphism (SNP) phylogenies were constructed for each cluster containing at least one Scottish isolate, using the General Time Reversible model with 1000 bootstrap replicates. Phylogenies were generated from core SNP alignments with recombinant regions removed using Gubbins (v2.3.4; appendix 1 p 11), selecting the highest N50-valued assembly as a reference. Phylogenetic analysis of the entire dataset was performed using the same method but without removing recombinant SNPs and using the Philadelphia-1 reference genome to call variants (GenBank accession AE017354.1). For select clusters, rooted trees were constructed by the same method with the additional inclusion of an outgroup sequence chosen based on the location of clustered isolates in the whole-dataset phylogeny. Phylogenies were visualised using GrapeTree (v1.5.0)[20] and the Interactive Tree of Life web visualisation tool (v6.5.7),[21] with additional formatting and details added using InkScape (v1.0).

For more on **InkScape** see https://inkscape.org/

Patient metadata

Permission to access patient epidemiological data was sought and obtained from the electronic data research and innovation services (part of the Information Services Division of Public Health Scotland) through the public benefit and privacy panel for health and social care. We obtained epidemiological data on the Scottish population from 1984 to 2020. Limits were applied so that individual cases could not be identified. However, information on broad regions of home and travel were available, as well as age range, sex, whether patients smoked or had common symptoms, and whether the case was travel-associated, community-associated, or hospital-associated. The project adhered to the ethical principles in the Research Governance Framework for Health and Community Care, Second Edition, 2006, and applicable legal and regulatory requirements. Associated metadata for the sequences downloaded from NCBI was collected using e-utilities (v12.3) and processed using an in-house pipeline. Sequence-based typing profiles were estimated with MLST (v2.19.0), using an archived copy of the sequence type profile information from Public Health England[22] and resolving composite allele calls using Python (v3.6.5). Where the sequence type could not be determined due to the presence of novel or duplicated typing alleles, sequences were classified as undetermined. Additional metadata was sourced from the relevant scientific literature (appendix 1 pp 11–12). Unless otherwise stated, isolates with missing metadata were classified as unknown for the category in which data was unavailable.

For more on **e-utilities** see https://github.com/Klortho/edirect

For more on **MLST** see https://github.com/tseemann/mlst

Case definitions

In accordance with the European Centre for Disease Prevention and Control (ECDC) definition of an Legionnaires' disease cluster,[23] isolates were considered epidemiologically linked if they were in the same cluster and closely associated in space (eg, isolated from the same street or building) and time (within a 6-month timeframe). We defined isolates as probably related if they were from the same geographical region, differed by 16 SNPs or less, and differed by fewer SNPs than with other isolates unconnected in space. Infections were categorised following the definitions used by the ECDC. Patients were classified with TALD if they had stayed at or visited a commercial accommodation site in the 14 days before the onset of illness.[24] Similarly, if patients were admitted to hospital 2–10 days before onset of illness, infections were classified as HALD.[25] Community acquisition was assigned for cases in which the patient did not meet either the

For more on **Gubbins** see https://github.com/nickjcroucher/gubbins

	People with Legionnaires' disease (n=193)
Age	
30–39 years	4 (2·1%)
40–49 years	26 (13·5%)
50–59 years	60 (31·3%)
60–69 years	59 (30·7%)
70–79 years	26 (13·5%)
80–89 years	4 (2·2%)
Unknown	14 (7·3%)
Sex	
Male	134 (69·4%)
Female	51 (26·4%)
Unknown	8 (4·1%)
Legionnaires' disease classification	
TALD	102 (52·8%)
CALD	54 (28·0%)
HALD	18 (9·3%)
Unknown	19 (9·8%)
Smoking status	
Unknown	92 (47·7%)
Smoker	72 (37·3%)
Non-smoker	29 (15·0%)
Symptoms	
Pneumonia	190 (98·4%)
Cough	67 (34·7%)
Shortness of breath	58 (30·1%)
Increased temperature	54 (27·9%)
Diarrhoea	38 (19·7%)
Confusion	34 (17·6%)
Lethargy	30 (15·5%)
Myalgia	23 (11·9%)
Chest pain or consolidation	22 (11·4%)
Headache	17 (8·8%)

Data are n (%). CALD=community-acquired Legionnaires' disease. HALD=Hospital-acquired Legionnaires disease. TALD=travel-acquired Legionnaires' disease.

Table: **Patients with culture-confirmed Legionnaires disease in Scotland between 1984 and 2020**

criteria for HALD or TALD, and cases with insufficient data collected were classed as unknown. We defined sporadic cases as clinical isolates that did not cluster with other clinical isolates at the threshold analysed.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

From 1984 to 2020, there was a total of 1138 cases of Legionnaires' disease identified in Scotland (appendix 1 p 2). Of these, 193 (17·0%) were culture-positive for *L pneumophila* and were analysed in this study (table).

To investigate the epidemiological relationships between isolates, we performed high resolution WGS-based clustering on isolates from all 193 patients with Legionnaires' disease and 229 environmental isolates from our reference archive, as well as 2944 isolates from global sources in the NCBI. For isolates with computationally determined sequence types, the clusters identified by cgMLST were largely consistent with sequence-based typing (appendix p 7). We identified 60 cgMLST clusters that contained at least one isolate from Scotland and only 29 (15·0%) of 193 Scottish clinical isolates did not cluster with other clinical isolates, indicative of sporadic cases by our definition (figure 1A). A further 13 Scottish clinical isolates clustered only with isolates from locations outside Scotland. Only ten clusters accounted for 111 (57·5%) Scottish clinical isolates and 57 (29·5%) Scottish clinical isolates belonged to one of three clusters that correspond to previously described sequence types known to be associated with human disease (ST1, ST37, and ST42; appendix 1 p 7). Within large clusters, we observed frequent overlap of clinical isolates from patients with HALD, CALD, and TALD and only four clusters containing three or more Scottish clinical isolates were restricted to a single infection category (figure 1B). Phylogenetic analysis of the entire dataset revealed that clusters containing HALD, CALD, and TALD patient isolates had emerged from diverse phylogenetic backgrounds across the species (figure 1C). The clusters identified often corresponded to paraphyletic groups, and we hypothesised that this reflected the effect of recombination on core gene profiles. To address this potential problem, recombination analysis (Gubbins) was performed on 43 clusters containing at least four isolates in total and including one or more Scottish isolates. Of these clusters, 33 (76·7%) contained regions of recombination representing more than 90% of SNPs identified (range 90·3–99·8%), potentially obscuring detection of both long-term and short-term epidemiological linkage (appendix 1 p 8). Taken together, our analysis revealed a small number of endemic clones that were responsible for 111 historical Scottish infections from a mixture of hospital-associated, community-associated, and travel-associated environments.

To investigate the relationship between Legionnaires' disease and travel, we analysed the geographical relationships in clusters containing TALD isolates. Only 17 (16·7%) of 102 Scottish TALD isolates were singletons (ie, did not form clusters with other sequences in the dataset), and we identified 13 clusters that contained more than one TALD isolate, accounting for 64 (62·7%) TALD isolates in total. Scottish TALD isolates that clustered together were typically from patients sharing travel history to the same geographical region, often clustering with clinical or environmental isolates from those regions (figure 2). Of the 102 TALD patients, at least 85 (83·3%) had travelled to destinations outside

Figure 1: **Clustering (cgMLST) and phylogenetic analysis of *Legionella pneumophila* from Scotland in relation to isolates of global origin**
A) Bar plot showing the number of Scottish isolates in each cgMLST cluster of *L pneumophila* (represented by distinctly coloured circles) that contained one or more isolates from Scotland. Bars are split into different colours to indicate the proportion of isolates classified as HALD, CALD, TALD, environmental, or unknown (clinical). Environmental isolates from ships are not included. Isolates that were not part of a cluster (singletons) are shown as a single bar on the left. For the three clusters that closely corresponded to major sequence types, the consensus sequence type of the clustered isolates for which a sequence type was determined (appendix 1 p 7) is shown in text above the bar. (B) A Venn diagram showing the number of clusters with Scottish clinical isolates that contained each possible overlap between infection categories. Infection category colours are consistent with figure 1A. Both the number and percentage of isolates in the clusters with each overlap are shown and singletons are included as separate, individual clusters. (C) A maximum likelihood core single-nucleotide polymorphism phylogeny of the entire dataset, with the cgMLST clusters containing one or more isolates from Scotland indicated by the same colours used for the circles in figure 1A. cgMLST=core-genome multilocus sequence typing. CALD=community-acquired Legionnaires' disease. HALD=hospital-associated Legionnaires' disease. TALD=travel-associated Legionnaires' disease.

the UK, predominantly within Europe (n=68; 66·7%), and eight clusters containing TALD isolates reflected either travel within the UK or on ships (figure 2A, B). Of note, 11 clusters contained both international TALD isolates and non-TALD isolates or UK environmental isolates (figure 2B, D), making it plausible that travel infections within these clusters were acquired locally. Our analysis supported six epidemiological links that had previously been inferred in historical cases where the putative source of infection was identified (figure 2A–C). We also identified 12 TALD patient isolates that were probably related to other isolates associated with the same geographical regions, differing by fewer SNPs than with isolates from other geographical regions and by only 0–14 SNPs in total (figure 2D, E; appendix 1 p 8). In most cases, further epidemiological inference was limited by a lack of available metadata relating to a precise source of isolation. However, in two cases involving European TALD linked to hotels, the associated isolates were located in distinct phylogenetic clades within broader lineages of European isolates (figure 2C, D). Of note, we identified cases where

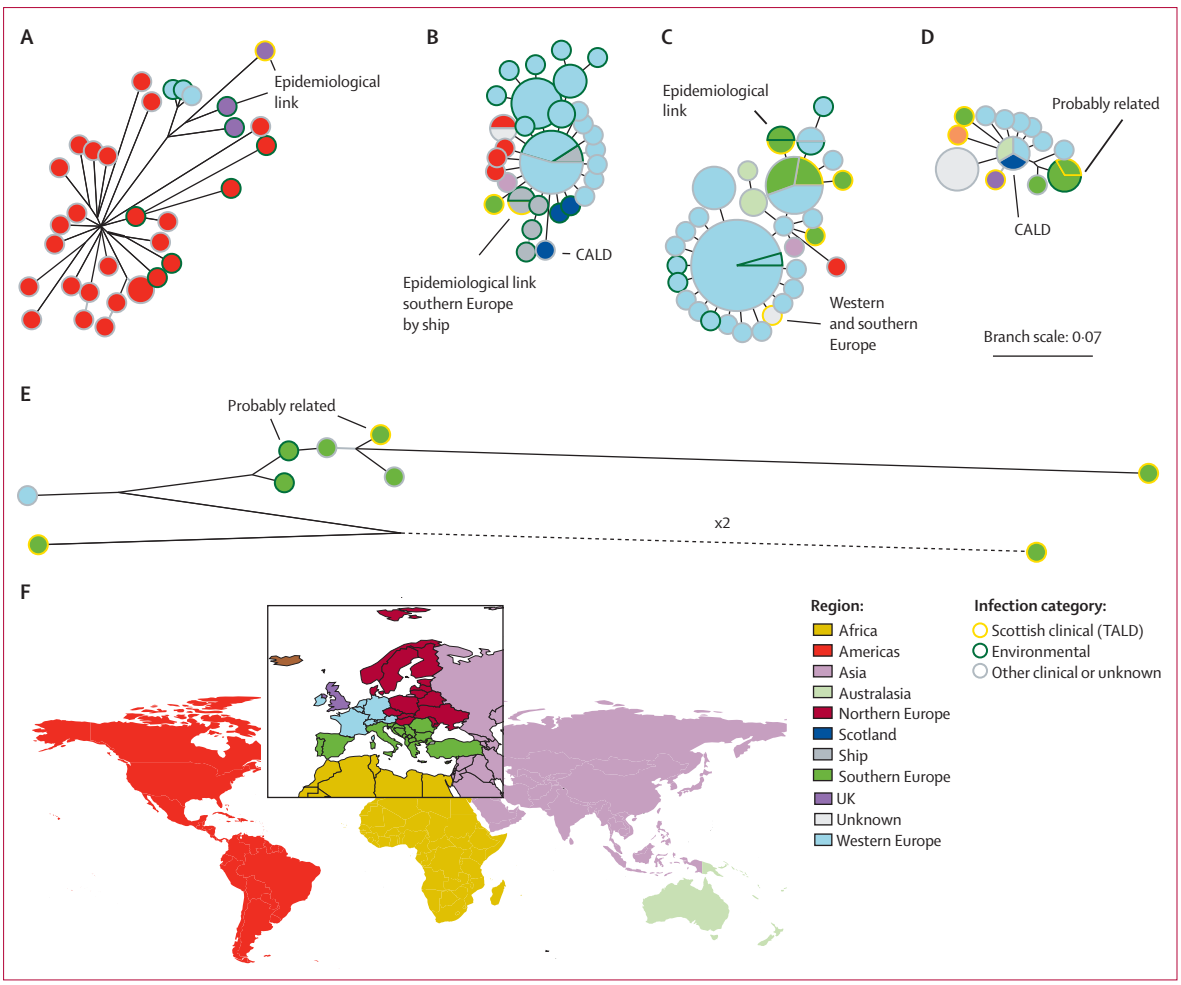

Figure 2: **Identification of national and international TALD-associated clusters.**
Maximum likelihood core single-nucleotide polymorphism phylogenies of five clusters containing TALD isolates (indicated by nodes with orange outlines), visualised using GrapeTree.[20] Clusters were selected to exemplify the types of geographical connections referred to in the text. (A) An isolate from a patient with TALD in Scotland with a national travel history is in a distinct phylogenetic clade with an epidemiologically linked environmental isolate. (B) An isolate from a patient with TALD in Scotland with a travel history involving a ship is in a distinct phylogenetic clade with an epidemiologically linked environmental isolate. The cluster contains additional isolates from ships and other isolates from Scotland. (C) Several isolates from patients with TALD in Scotland with travel history to southern Europe are in a distinct phylogenetic clade with other isolates from southern and western Europe. An isolate from one patient with TALD in Scotland is located in a nested phylogenetic clade with an epidemiologically linked environmental isolate. (D) An isolate from a patient with TALD in Scotland with international travel history to southern Europe is located in a distinct phylogenetic clade with a probably related environmental isolate. The cluster contains additional isolates from southern Europe and other isolates from Scotland. (E) Several isolates from patients with TALD in Scotland with travel history to southern Europe are located in a cluster with other clinical and environmental isolates from southern and western Europe. The cluster includes a pair of isolates that are probably related. Branch lengths are drawn to scale, branches of length of less than 0·00015 are collapsed and node sizes drawn proportionally to the number of isolates. Nodes are coloured by geographical region, with pie charts used to indicate the geographical distribution of isolates at collapsed nodes. Environmental isolates and Scottish TALD isolates are indicated by node border colour. (F) A map of geographical regions used to analyse the data is provided. For patients with TALD with travel history to multiple geographical regions or with a history of international travel on ships, additional known travel destinations are indicated by text. CALD=community-acquired Legionnaires' disease. TALD=travel-associated Legionnaires' disease.

probably related isolates were collected 4–17 years apart (appendix 1 p 9). For example, a TALD isolate and environmental isolates from the same travel destination region differing by only 4–5 SNPs were collected 14– 17 years apart, suggesting long-term persistence of an environmental clone with potential to cause TALD (figure 2E). To our knowledge, this is the first report of an international travel link between closely related isolates over such a long timeframe. Taken together, these data demonstrate the capacity for WGS to detect short-term and long-term epidemiological clusters associated with international travel.

To investigate the cause of HALD infections in Scotland, we examined the relatedness of previously defined HALD isolates to environmental isolates from nine different Scottish hospitals (denoted A–I). Isolates collected from seven hospitals belonged to more than one cluster, indicating the presence of multiple *L pneumophila*

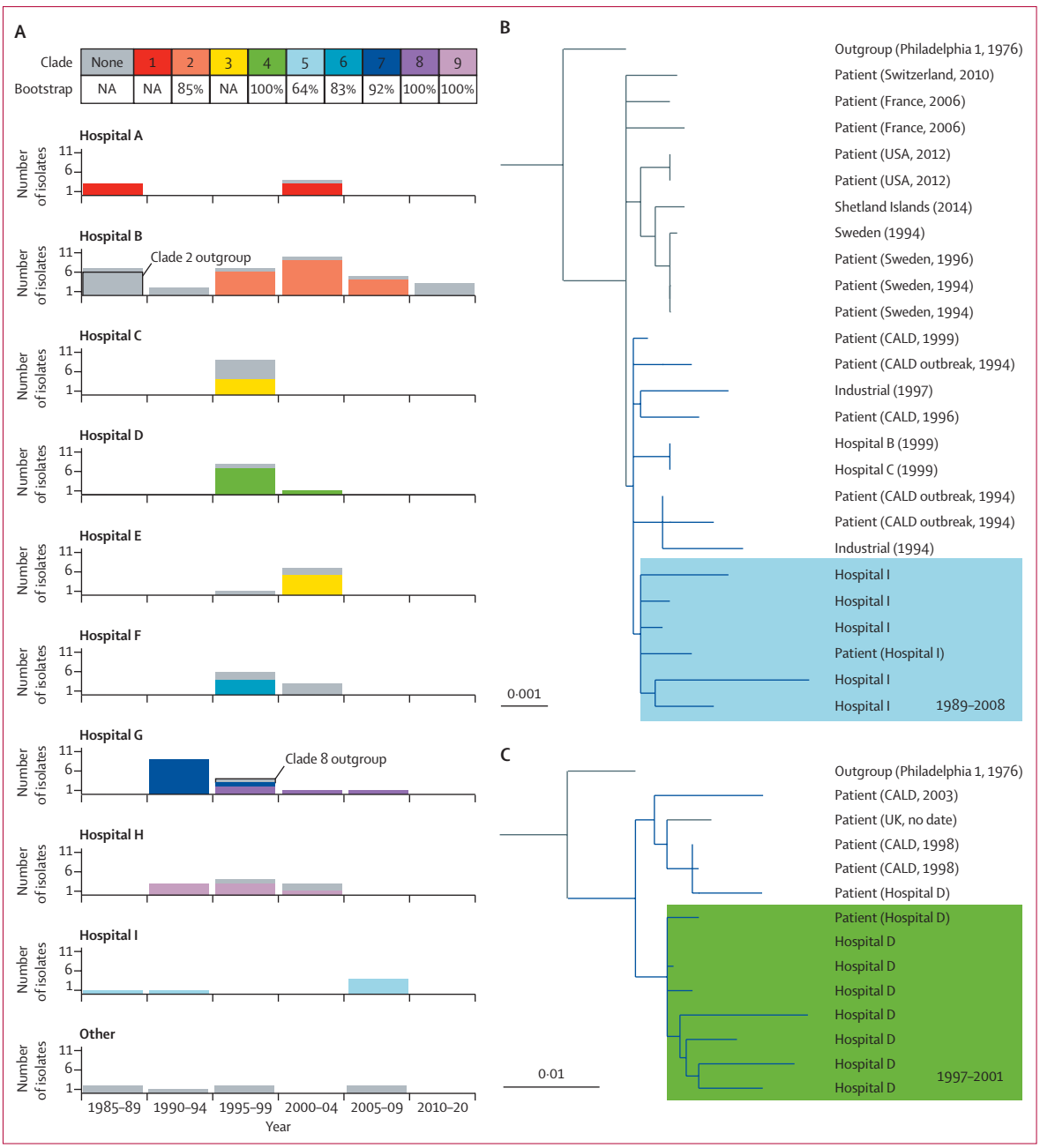

Figure 3: **Temporal and phylogenetic analysis reveals hospital-associated lineages of *Legionella pneumophila***
(A) The temporal distribution of hospital-associated Legionnaires' disease and environmental isolates associated with different hospitals over the timeframe of the study. Individual hospitals are indicated by letters and hospitals with less than four environmental isolates are grouped together and labelled as other. Isolates are coloured by hospital-associated clade, with grey used to indicate isolates that were not part of a hospital-associated clade. Clinical isolates within close outgroups to hospital-associated clades are indicated by text. Bootstrap support values are shown for each clade that did not constitute an entire cluster (or an entire cluster excluding one isolate). (B) Maximum likelihood core SNP phylogeny of the cluster containing hospital clade 5, with the branch scale indicated. Near-monophyletic clades of Scottish isolates (allowing up to one exception) are indicated by blue branches, with the exceptions marked in grey to match the other parts of the tree. An outgroup was selected as described in the methods. The exact date range of the hospital-associated clades is indicated by bold text. (C) Maximum likelihood core SNP phylogeny of the cluster containing hospital clade 4, shown as in 3B. CALD=community-acquired Legionnaires' disease. SNP=single-nucleotide polymorphism.

lineages. However, phylogenetic analysis revealed nine distinct hospital-associated clades (defined as clades in which most isolates, and at least four in total, were environmental isolates from the same hospital; appendix 1 pp 3–6). Of the 18 patients diagnosed with HALD, isolates from seven (38·9%) patients belonged to hospital-associated clades that were specific to an individual hospital (clades 4, 5, 7, and 9), consistent with

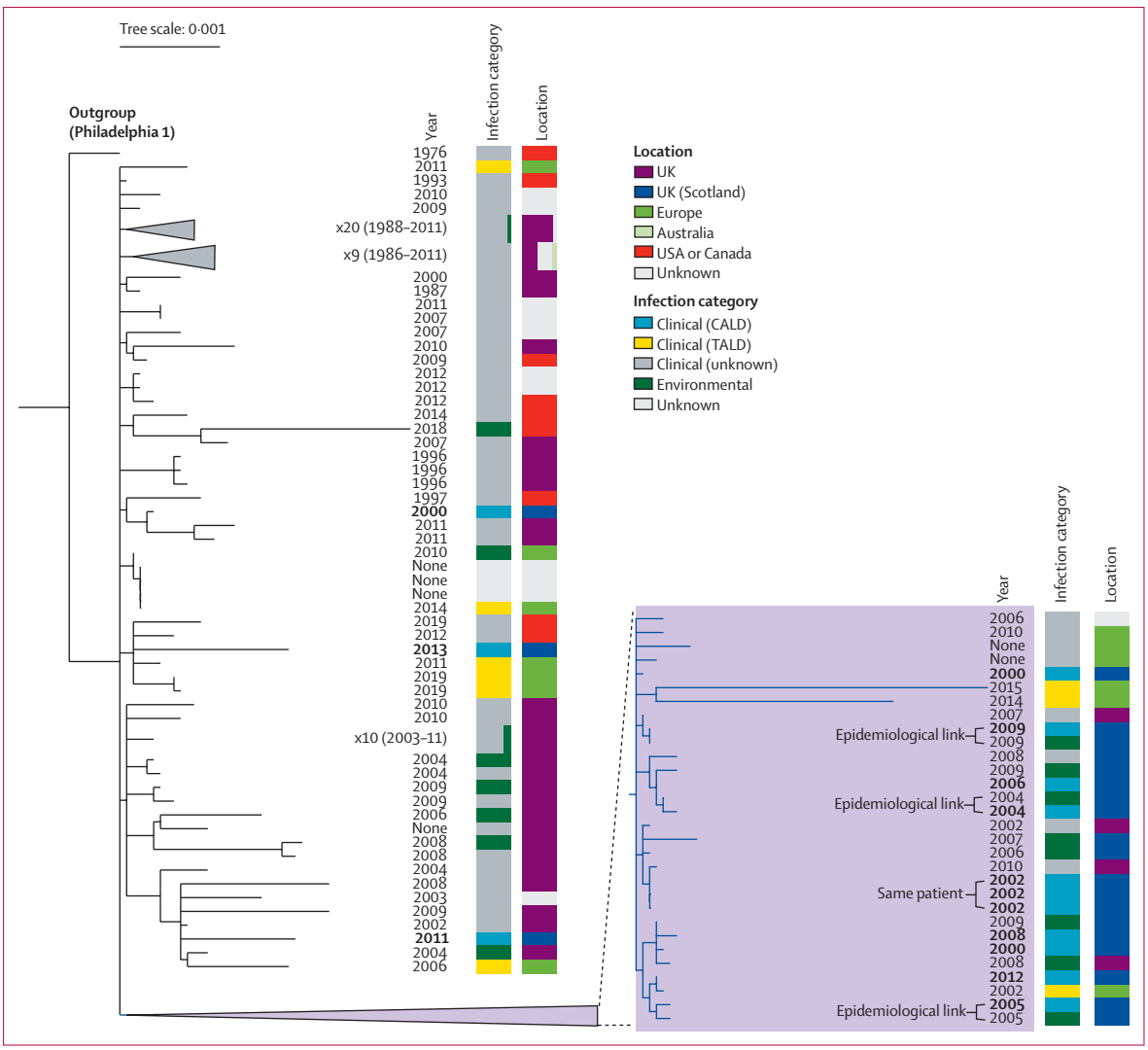

Figure 4: **Identification of an *Legionella pneumophila* lineage responsible for numerous CALD cases over 13 years**
Maximum-likelihood core SNP phylogeny of a single cluster (120 isolates, including 29 from Scotland) generated from non-recombinant SNPs and including the Philadelphia-1 reference genome as a closely related outgroup to root the tree. Three large clades containing no isolates from this study have been collapsed. A specific, predominantly Scottish clade (69% bootstrap support, described in text) is collapsed and shown separately (marked in blue). Isolates are labelled with the year of isolation. Bold text is used to indicate isolates from 12 historic CALD cases referred to in text. The infection category and isolation location countries are indicated by colour. Three epidemiologically linked pairs of isolates are indicated. CALD=community-acquired Legionnaires' disease. SNP=single-nucleotide polymorphism. TALD=travel-associated Legionnaires' disease.

their HALD classification based on epidemiological information. Another seven (38·9%) HALD isolates were within close outgroups to hospital-associated clades, differing by up to 25 non-recombinant SNPs from associated hospital isolates (2 and 8; appendix 1 pp 3–4). Importantly, temporal analysis showed that closely related clinical and environmental isolates associated with individual hospitals had been sampled over 1–19 years, between 1989 and 2008 (figure 3A). Our phylogenetic analysis indicated that *L pneumophila* populations in hospitals are often closely related to those in other environmental reservoirs, presenting a challenge for source attribution (figure 3B; appendix pp 3–6). Hospital-associated clades 2, 4, 5, and 7 had emerged from broader, diverse clades of Scottish isolates (with one exception in clade 4) that included isolates from a variety of environmental reservoirs (figure 3b, C; appendix 1 pp 3, 5). The hospital-associated clade 5 was closely related to Scottish CALD isolates from an outbreak more than 30 km from hospital I (figure 3B). Even though the outbreak isolates were collected more than a decade earlier, they only differed from the hospital clade isolates by 6 to 18 core alleles. Similarly, the cluster containing hospital clade 4 also included two CALD isolates (figure 3C), which were spatially and temporally clustered with each other but were isolated more than 200 km from

the hospital. Despite their wide geographical range, the isolates in this cluster differed by less than 50 non-recombinant SNPs, suggesting a relatively recent common ancestor. Collectively, these data indicate that clinically relevant Scottish endemic clones of *L pneumophila* might be geographically widespread in a variety of reservoirs. Furthermore, hospital-associated strains probably emerged following the introduction of more widely disseminated clones into hospital water systems, followed by persistence or regular re-seeding over a long timeframe.

Finally, to investigate the epidemiology of CALD in Scotland, we examined the epidemiological relationships in clusters containing CALD isolates. Excluding the Edinburgh outbreak in 2012, our data included single CALD isolates from each of 40 patients, of which 15 (37·5%) sequences were located as singletons or in independent clusters. In contrast, isolates from 12 (30·0%) of 40 patients with CALD, collected over a 13 year time period (from 2000–13), belonged to a single cluster, which consisted of 120 global isolates of primarily ST37, including from nine (28·1%) of the 32 patients in the younger age range (<50–59 years; figure 1A). Nine CALD isolates in this cluster, collected over a 9-year period from locations up to 35 km apart, were segregated in the same phylogenetic clade (69% bootstrap support), which also contained one Scottish clinical isolate of unknown provenance, three from the rest of the UK, and six associated with other European countries, either by isolation or travel (figure 4). Notably, at least seven of the eight environmental isolates in the clade were from a defined urban area within Scotland and six were from different private homes. Sequenced isolates from three of these patients with CALD were included in our dataset and segregated together with corresponding environmental isolates in the phylogeny (figure 4). Taken together, the data suggest that most CALD cases are sporadic and have a diverse clonal cause. However, we have revealed the existence of a single major lineage of *L pneumophila* responsible for at least 12 historic CALD cases in Scotland.

Discussion

In this study, we carried out a historical analysis of all clinical isolates of *L pneumophila* in Scotland spanning a 36-year timeframe. Our genomic clustering and phylogenetic analysis revealed that many Scottish Legionnaires' disease infections are caused by a small number of widely disseminated *L pneumophila* clones, found in the community, travel, and hospital environments. It has previously been demonstrated that some *L pneumophila* sequence types are more likely to cause human infections, which might reflect enhanced pathogenicity.[26] In particular, we identified one lineage that contained a quarter of all community-associated infections in Scotland, spanning a 13-year timeframe within a fairly small, 35 km geographical area. All isolates in the lineage belonged to ST37, a major global disease-associated clone, and at least three cases had confirmed links to environmental isolates cultured from residential potable water isolates, suggesting that the lineage might have existed in Scottish drinking water. It should be pointed out that *Legionella* spp are ubiquitously found in water, but our findings highlight the potential for environmental screening to identify strains with increased pathogenicity.

Our analysis revealed the existence of specific sublineages of *L pneumophila* associated with repeated infections in Scotland over many years, and inclusion of clinical and environmental isolates from national and global sources revealed previously cryptic epidemiological relationships, such as closely related TALD isolates that were associated with similar European travel destinations. Consistent with previous analyses,[5,10] we also found that specific lineages of *L pneumophila* were sampled from Scottish hospitals over many years. Phylogenetic analysis indicated that these hospital-associated strains had typically emerged from diverse, widely disseminated clones and that hospital plumbing might be colonised or re-seeded from integrated municipal systems with specific subtypes of *L pneumophila*. Of note, there have been no hospital cases of Legionnaires' disease in Scotland in the last 10 years, probably due to improved water safety management and treatment regulations.[27]

Clustering of TALD cases in Europe has been reported previously but our findings indicate that such epidemiological links might exist over a much longer timeframe than previously appreciated.[28] For most TALD cases, there was a scarcity of environmental isolates to make definitive links to the travel destination. However, for two confirmed cases of European TALD, environmental isolates formed close monophylies with clinical isolates and were nested within larger European clusters, highlighting the value of wide environmental surveillance sampling. In previous analysis of 64 environmental isolates from the same urban area, a clear geographical signature was observed, which provided sufficiently high genetic resolution to train predictive models for source attribution.[11] However, culturing *L pneumophila* from water samples has a low rate of success[29] and our analysis highlighted large gaps in the sampled diversity of environmental *L pneumophila* that should be addressed for more effective source attribution.

A limitation of *L pneumophila* WGS studies is that culture-positive cases account for only 10–20% of all clinical samples (17% in the current study). Theoretically, this low frequency of culture success could introduce a bias towards strains that are more easily cultured in vitro. However, previous studies have not identified differences in the culture success rate of different sequence types, resulting in similar distributions of sequence-based types among samples subjected to culture-dependent and culture-independent typing.[26,30] Accordingly, our sequenced clinical isolates probably represent a reasonable snapshot of historical Scottish clinical *L pneumophila*. Furthermore, our main discovery of

specific clones that are responsible for recurrent infections of Legionnaires' disease over many years in Scotland would be unaffected by such a bias.

As with other studies,[10,11] we often observed close genomic clustering of spatially and temporally unrelated isolates, presenting implications for outbreak investigation. For many bacterial pathogens, there is a clear correlation between genomic and epidemiological relatedness, and genetic distances can be used to define clustering thresholds that capture epidemiological links.[31] However, studies in *L pneumophila* have shown a large overlap between the similarity of epidemiologically linked and unlinked isolates, owing to diverse routes of infection and complex biological factors.[3,7,10] Importantly, depending on the environment, *L pneumophila* might either proliferate rapidly or undergo long periods of dormancy resulting in variable evolutionary rates of different populations.[32] Accordingly, the appropriate genetic distance thresholds required to capture epidemiological links might be unclear, and will depend on the dynamics of individual outbreaks, as proposed for *Salmonella*.[33]

Accordingly, we propose the application of conservative distance thresholds that are informed by our understanding of the biology of Legionnaires' disease outbreaks to capture epidemiological links, though definitive validation of proposed epidemiological links might then require deep environmental sampling and detailed epidemiological metadata analysis. Our data also highlight the importance of prospective sampling to inform appropriate preventive measures. For example, HALD isolates in two clusters, which were collected years before environmental isolates from the hospital were obtained, appeared as outliers of respective phylogenetic clades.[10] The broad timeframe of up to 17 years for isolates within many clusters supports a recent proposal to re-evaluate the European Legionnaires' Disease Surveillance Network cluster definition to consider accommodation sites associated with multiple cases regardless of the time elapsed between them.[34] To support this proposal, we advocate the importance of ongoing environmental surveillance for *L pneumophila* in industrial and public water systems to support the rapid delineation of Legionnaires' disease clinical cases and outbreaks. Taken together, our findings support the proposal that routine regular environmental sampling is required to facilitate the WGS-based identification of epidemiological links and the attribution of outbreak sources, and to inform public health measures targeting endemic clones that are an ongoing threat to public health.

Contributors

DL, BW, AS, and JRF designed the study. DL, AS, and RC collected the genomic data for Scottish isolates. DL, ML, JM, and RC collected the epidemiological metadata. AS, ML, DL, RC, and JM had access to patient epidemiological data and only permitted variables were shared with the other authors. All authors had full access to the data reported in this study. JG, BW, DL, and JA analysed the data and JRF and AS assisted with interpretation. DL and JG have accessed and verified all the data in the study. JG, BW, DL, AS, and JRF wrote the manuscript. All authors reviewed, revised, and approved the final report.

Declaration of interests

JM reports being the chair of the Scottish National Incident Management team for COVID-19; the Head of Health Protection (Infection Services) within Public Health Scotland; previously being the Clinical Director of Health Protection Scotland and the chair of the Scientific Steering Committee of a European Horizon 2020 project additionally supported by European Centre for Disease Prevention and Control and WHO (IMOVE and IMOVE-COVID-19 projects) on vaccine effectiveness of seasonal influenza and COVID-19, respectively; a member of the UK New and Emerging Respiratory Virus Threat Advisory Group, which advises on infection threat; previously being a member of the Scottish COVID-19 Scientific Advisory Group and of the UK Scientific Advisory Group on Emergencies advising on COVID-19 infections. All these positions have been undertaken within JM's National Health Service salary and have not included any additional payment. All other authors declare no competing interests.

Data sharing

All data, including de-identified participant data, can be made available upon reasonable request from the corresponding author.

Acknowledgments

This work was supported in part by grants awarded to JRF, AS, and DL from the Chief Scientist's Office, Scotland (ETM/421); to JRF from the Biotechnology and Biological Sciences, Research Council (BBSRC) ISP (BBS/E/D/20002173) and BBSRC (UK; BBS/E/D/20002174); and to JRF from the Wellcome Trust (201531/Z/16/Z) and the Medical Research Council (MRC; UK; MR/N02995X/1). JG was supported by an MRC Precision Medicine Doctoral Training Programme award (MR/N013166/1). The authors would like to thank staff within the Scottish Microbiology Reference Laboratory, Glasgow, for sequencing isolates after 2015 and the electronic data research and innovation services for their help and support with the Public Benefit and Privacy Panel application process.

References

1 Cianciotto NP, Hilbi H, Buchrieser C. Legionnaires' disease. In: The prokaryotes. Berlin Heidelberg: Springer-Verlag, 2013: 147–217.

2 Baskerville A, Fitzgeorge RB, Broster M, Hambleton P, Dennis PJ. Experimental transmission of legionnaires' disease by exposure to aerosols of *Legionella pneumophila*. *Lancet* 1981; **2:** 1389–90.

3 Orkis LT, Harrison LH, Mertz KJ, Brooks MM, Bibby KJ, Stout JE. Environmental sources of community-acquired legionnaires' disease: a review. *Int J Hyg Environ Health* 2018; **221:** 764–74.

4 Bacigalupe R, Lindsay D, Edwards G, Fitzgerald JR. Population genomics of *Legionella longbeachae* and hidden complexities of infection source attribution. *Emerg Infect Dis* 2017; **23:** 750–57.

5 Rangel-Frausto MS, Rhomberg P, Hollis RJ, et al. Persistence of *Legionella pneumophila* in a hospital's water system: a 13-year survey. *Infect Control Hosp Epidemiol* 1999; **20:** 793–97.

6 Beauté J, Plachouras D, Sandin S, Giesecke J, Sparén P. Healthcare-Associated Legionnaires' disease, Europe, 2008–2017. *Emerg Infect Dis* 2020; **26:** 2309–18.

7 David S, Mentasti M, Tewolde R, et al. Evaluation of an optimal epidemiological typing scheme for *Legionella pneumophila* with whole-genome sequence data using validation guidelines. *J Clin Microbiol* 2016; **54:** 2135–48.

8 Raphael BH, Baker DJ, Nazarian E, et al. Genomic resolution of outbreak-associated *Legionella pneumophila* serogroup 1 isolates from New York State. *Appl Environ Microbiol* 2016; **82:** 3582–90.

9 Schjørring S, Stegger M, Kjelsø C, et al. Genomic investigation of a suspected outbreak of *Legionella pneumophila* ST82 reveals undetected heterogeneity by the present gold-standard methods, Denmark, July to November 2014. *Euro Surveill* 2017; **22:** 30558.

10 David S, Afshar B, Mentasti M, et al. Seeding and establishment of *Legionella pneumophila* in hospitals: implications for genomic investigations of nosocomial Legionnaires' disease. *Clin Infect Dis* 2017; **64:** 1251–59.

11 Buultjens AH, Chua KYL, Baines SL, et al. A supervised statistical learning approach for accurate *Legionella pneumophila* source attribution during outbreaks. *Appl Environ Microbiol* 2017; **83:** e01482-17. .

12 McAdam PR, Vander Broek CW, Lindsay DS, et al. Gene flow in environmental *Legionella pneumophila* leads to genetic and pathogenic heterogeneity within a Legionnaires' disease outbreak. *Genome Biol* 2014; **15:** 504.

13 David S, Mentasti M, Lai S, et al. Spatial structuring of a *Legionella pneumophila* population within the water system of a large occupational building. *Microb Genom* 2018; **4:** e000226.

14 Herwaldt LA, Marra AR. Legionella: a reemerging pathogen. *Curr Opin Infect Dis* 2018; **31:** 325–33.

15 Bhopal RS, Diggle P, Rowlingson B. Pinpointing clusters of apparently sporadic cases of Legionnaires' disease. *BMJ* 1992; **304:** 1022–27.

16 Wee BA, Alves J, Lindsay DSJ, et al. Population analysis of *Legionella pneumophila* reveals a basis for resistance to complement-mediated killing. *Nat Commun* 2021; **12:** 7165.

17 Moran-Gilad J, Prior K, Yakunin E, et al. Design and application of a core genome multilocus sequence typing scheme for investigation of Legionnaires' disease incidents. *Euro Surveill* 2015; **20:** 21186.

18 Silva M, Machado MP, Silva DN, et al. chewBBACA: a complete suite for gene-by-gene schema creation and strain identification. *Microb Genom* 2018; **4:** e000166.

19 Balaban M, Moshiri N, Mai U, Jia X, Mirarab S. TreeCluster: clustering biological sequences using phylogenetic trees. *PLoS One* 2019; **14:** e0221068.

20 Zhou Z, Alikhan N-F, Sergeant MJ, et al. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* 2018; **28:** 1395–404.

21 Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 2021; **49:** W293–96.

22 European Working Group for Legionella Infections, Public Health England, and the European Centre for Disease Prevention and Control. *Legionella pneumophila* sequence-based typing. 2019. http://bioinformatics.phe.org.uk/legionella/legionella_sbt/php/sbt_homepage.php (accessed April, 2016).

23 European Centre for Disease Prevention and Control. Legionnaires' disease outbreak investigation toolbox. https://legionnaires.ecdc.europa.eu/?pid=206 (accessed Dec 2, 2021).

24 European Centre for Disease Prevention and Control. European legionnaires' disease surveillance network (ELDSNet): operating procedures for the surveillance of travel associated legionnaires' disease in the EU/EEA: 2017. 2017 https://data.europa.eu/doi/10.2900/485245 (accessed Dec 21, 2020).

25 GOV.UK. Legionnaires' disease: case definitions. https://www.gov.uk/government/publications/legionnaires-disease-clinical-case-definitions/legionnaires-disease-case-definitions (accessed Aug 16, 2022).

26 David S, Rusniok C, Mentasti M, et al. Multiple major disease-associated clones of *Legionella pneumophila* have emerged recently and independently. *Genome Res* 2016; **26:** 1555–64.

27 NHS National Services Scotland. Scottish health technical memorandum 04-01 water safety for healthcare premises part B: operational management. 2014. https://www.nss.nhs.scot/media/1807/shtm-04-01-part-b-v20-jul-2014.pdf (accessed Aug 2022).

28 Rota MC, Cano Portero R, Che D, Caporali MG, Hernando V, Campese C. Clusters of travel-associated Legionnaires disease in Italy, Spain and France, July 2002–June 2006. *Euro Surveill* 2007; **12:** E3–4.

29 International Organization for Standardization. ISO 11731:2017(en), Water quality — enumeration of Legionella. 2017. https://www.iso.org/obp/ui/#iso:std:iso:11731:ed-2:v1:en (accessed Dec 3, 2021).

30 Mentasti M, Fry NK, Afshar B, Palepou-Foxley C, Naik FC, Harrison TG. Application of *Legionella pneumophila*-specific quantitative real-time PCR combined with direct amplification and sequence-based typing in the diagnosis and epidemiological investigation of Legionnaires' disease. *Eur J Clin Microbiol Infect Dis* 2012; **31:** 2017–28.

31 Meehan CJ, Moris P, Kohl TA, et al. The relationship between transmission time and clustering methods in *Mycobacterium tuberculosis* epidemiology. *EBioMedicine* 2018; **37:** 410–16.

32 Wells M, Lasek-Nesselquist E, Schoonmaker-Bopp D, et al. Insights into the long-term persistence of Legionella in facilities from whole-genome sequencing. *Infect Genet Evol* 2018; **65:** 200–09.

33 Besser JM. Salmonella epidemiology: a whirlwind of change. *Food Microbiol* 2018; **71:** 55–59.

34 Rota MC, Bella A, Caporali MG, et al. Travel-associated Legionnaires' disease: would changing cluster definition lead to the prevention of a larger number of cases? *Epidemiol Infect* 2018; **147:** e62.