



Comparison of whole genome sequencing typing tools for the typing of Belgian *Legionella pneumophila* isolates

Fedoua Echahidi¹ · Subin Park³ · Alaeddine Meghraoui² · Florence Crombé¹ · Oriane Soetens¹ · Denis Piérard¹ · Benoit Prevost² · Ingrid Wybo¹ · Charlotte Michel^{1,2}

Received: 13 August 2024 / Accepted: 6 December 2024 / Published online: 21 December 2024
© The Author(s) 2024

Abstract

Whole genome sequencing (WGS) marks a turning point for outbreak investigations for microorganisms related to public health matters, like *Legionella pneumophila* (*Lp*). Here, we evaluated the available *Lp* WGS typing tools for isolates of previously documented Belgian outbreaks, as well as small groups of related and non-related isolates. One reference strain and 77 clinical and environmental isolates were evaluated. Seven isolates belong to a Sequence Type (ST) 36 outbreak in 1999 and sixteen (ten clinical, two matching environmental and four non-related controls) belong to another ST1 outbreak in 1985–1987. The remaining isolates belong to small groups of related and non-related isolates of diverse ST's. WGS was performed and data were analysed using whole genome (wg) and core genome (cg) multilocus sequence typing (MLST) with “Ridom SeqSphere +” (cgMLST), “Applied Maths-Bionumerics” (wgMLST) and the 50 loci cgMLST (CDC/ESGLI_ESC-MID). Results of the three tools were concordant with the traditional Sequence Based Typing (SBT). The known outbreaks and small clusters could be detected and clear discrimination of ST1 non-related isolates was obtained. In addition, the 50 loci cgMLST allowed to classify the isolates into subtypes because almost all the 50 genes could be called in all the analysed isolates, which was not achieved by the other tools. This is a big advantage in terms of standardisation and comparison between laboratories for future epidemiological investigations. WGS allowed to analyse a large volume of samples and generated more accurate conclusions for outbreak investigations compared to other typing methods due to its higher discriminatory power and throughput.

Keywords *Legionella pneumophila* · Whole genome sequencing

Introduction

Legionella is a gram-negative facultative intracellular pathogenic bacterium. The bacterium is found in air conditioning systems, cooling towers, showers, hot water systems,

fountains and other devices that make aerosols [1]. People become infected by inhalation of contaminated aerosols [1]. The infection starts with the invasion and replication in macrophages and epithelial cells of the lungs [2, 3] and cannot be transmitted from person to person [1]. *Legionella* was first identified as the causative agent of pneumonia in 1977 after an outbreak in Philadelphia during a meeting of American veterans from the war of Korea [4]. It can cause two different diseases; Pontiac fever and Legionnaires' disease (LD) [5]. Pontiac fever is a mild respiratory disease that occurs after a short incubation period (24–48 h) with the bacteria and that can disappear spontaneously. LD is an atypical pneumonia with a variable mortality rate, ranging from less than 1% and up to 80% depending on several factors, mainly the immune status of the host [6, 7]. The incubation period for this disease is 2–10 days, and up to 16 days [4, 8]. The risk groups for this infection are mainly older people and immunocompromised patients [8, 9]. The incidence of LD

✉ Fedoua Echahidi
fedoua.echahidi@uzbrussel.be

¹ Department of Microbiology and Infection Control, National Reference Centre for Legionella Pneumophila, Vrije Universiteit Brussel (VUB), Universitair Ziekenhuis Brussel (UZ Brussel), Laarbeeklaan 101, 1090 Brussels, Belgium

² Department of Microbiology, National Reference Centre for Legionella Pneumophila, Laboratoire Hospitalier Universitaire de Bruxelles - University Laboratory of Brussels (LHUB-ULB), Route de Lennik 808, 1070 Brussels, Belgium

³ Association of Public Health Laboratories (APHL), Silver Spring, MD 20910, USA

follows an increasing trend in Europe and worldwide in the last years [10–13]. The cause of the increased notification rate recently observed in Europe remains unknown. Factors that may explain these increases include changes in national testing policies and surveillance systems; an ageing EU/EEA population; and the design, infrastructure, and maintenance of water systems used in buildings. Changes in climate and weather patterns across Europe and worldwide can also impact both the ecology of *Legionella* in the environment and the exposure to water aerosols containing the bacteria [12]. There are more than 60 *Legionella* species known [14]. *Lp* occurs in 90% of the human isolates with 16 serogroups, from which 90% are *Lp* serogroup 1 [15–17].

Epidemiological typing of *Lp* is performed using different methods [18–20]. The standardised SBT is until now the gold standard for molecular epidemiology studies of LD [21, 22]. SBT, based on 7 loci comparison, is discriminatory and reliable; however, for very ST's such as ST1 (worldwide abundant) or ST47 (very frequent in north-western Europe), it is often challenging to conclude whether environmental and patient isolates belonging to the same frequent ST are related [20, 23]. Other methods, like Pulsed-field Gel Electrophoresis (PFGE) [20] or the spoligotyping [24] are applied to further discriminate such isolates. However, these methods have a very limited value [20, 24].

The WGS has shown a big advantage in the investigation of both nosocomial and community outbreaks for public health important microorganisms, mainly because of its higher discriminatory power and higher throughput in comparison to older methods [25–27]. The first studies applying the WGS tool for *Lp* outbreak investigations used the analysis of single nucleotide polymorphisms (SNPs) [25]. Subsequently, core genome and whole genome multi locus sequence typing cgMLST/wgMLST schemes were developed, based on 1521 core genes (cgMLST), supplemented by 4249 accessory genes for wgMLST [28]. These schemes are based on gene-by-gene comparison approach offering an easier and automated way for *Lp* typing. However, even with a more simplified way of analysis than SNPs-WGS the use of these schemes do not always allow for the assignment of subtypes that would be reproducible and comparable between different laboratories in a similar way as for the classical SBT, as 100% success of the whole set of allelic assignment is indeed not often achievable [28]. These schemes rely on the use of clustering methods using thresholds for relatedness conclusion and should be rerun at every sample supplementation. As opposed to that and according to the study of David et al. [29] a more simplified cgMLST scheme with approximately 50 genes would offer the best compromise between improving the discrimination obtainable by current WGS methods and maintaining a good epidemiological concordance, without the need to use thresholds or clustering methods. In addition, using this simplified scheme would

increase the chance of having successful results since it only needs a few alleles to align. This 50 loci cgMLST scheme has been selected by the ESGLI-NGS working group as the best choice method for the standardisation and implementation of WGS in a routine fashion. It would be the official alternative to SBT and is currently under thorough validation in a cooperative study with the CDC (<https://github.com/CDCgov/LpSubP>).

In the past, two large outbreaks occurred in Belgium in 1985–1987 and in 1999 respectively [30, 31]. The first one was a nosocomial wide outbreak of *Lp* serogroup 1 (ST1) involving 32 patients, four (12%) of them died in the hospital. The second outbreak, also caused by *Lp* serogroup 1 (ST36), occurred among visitors of an annual fair that attracted 50,000 visitors. The outbreak involved 93 people among exhibitors and visitors, five (5.4%) of them died because of LD. The aim of this retrospective study is to apply the available gene by gene WGS typing tools for the investigation of isolates from the two most important outbreaks in Belgium as well as small related and sporadic clinical and environmental *Lp* isolates from our routine collection [30–32]. The thoroughly validated 1521 loci cgMLST is taken as a reference for the comparison with the wgMLST and 50 loci MLST schemes.

Materials and methods

Legionella pneumophila isolates and culture

A total of 78 *Lp* isolates were tested, see Table 1. One reference strain (Philadelphia ATCC 33152), taken as a control for the whole process, and 77 clinical or environmental isolates. Out of these 77 isolates, 7 belong to the outbreak from 1999 involving ST1 [31] and 16 (10 clinical, two matching environmental and four non-related isolates taken as controls) belong to the nosocomial outbreak from 1985–1987 involving ST36 [30]. The 54 remaining isolates were, in most cases, originally tested in a context of the investigation of infection source in sporadic cases. They belong to small groups of related and non-related isolates of 21 different ST's consisting each time of one or more patient isolates with the respective one or more environmental isolates presumed to be the source of infection.

Pure cultures of each strain were grown on BCYE Legionella medium (Oxoid) at 35 °C in an aerobic 5% CO₂-humid atmosphere until sufficient growth (at least 4 days) was obtained.

DNA purification

Grown colonies were suspended in the lysis buffer to start DNA purification. The Dneasy blood & tissue kit (Qiagen,

Table 1 List of *Lp* isolates selection from Belgian historical collection used for a retrospective analysis by 3 Whole Genome Sequencing typing tools

Strain number	outbreak/group number	Source	Cluster ID cg-wgMLST	Collection date	SBT ST (sanger sequencing)	Epidemiological information
LEG251	1	Clinical	Cluster 2	1/11/1999	ST36	Outbreak_1999
LEG252	1	Clinical	Cluster 2	1/11/1999	ST36	Outbreak_1999
LEG253	1	Clinical	Cluster 2	1/11/1999	ST36	Outbreak_1999
LEG254	1	Clinical	Cluster 2	1/11/1999	ST36	Outbreak_1999
LEG255	1	Clinical	Cluster 2	1/11/1999	ST36	Outbreak_1999
LEG256	1	Clinical	Cluster 2	1/11/1999	ST36	Outbreak_1999
LEG257	1	Clinical	Cluster 2	1/11/1999	ST36	Outbreak_1999
LEG311	2	Clinical	Cluster 3	11/01/2003	ST110	Clinical isolates from 1 hospitalised patient with the corresponding environmental isolate from the suspected source (Water sample from the hospital). Same hospital as for LEG362-366
LEG312	2	Environmental		11/01/2003	ST196	
LEG321	3	Clinical	Cluster 8	21/11/2003	ST1	Clinical isolates from 2 hospitalised patients at another hospital than the outbreak 1985–87, but in the same city. The isolates at 3 months interval from each other
LEG323	3	Clinical	Cluster 8	26/02/2004	ST1	
LEG339	4	Clinical	Cluster 6	12/10/2006	ST110	2 Clinical isolates from the same patient at a nursing home and the corresponding environmental isolates from the suspected source (Shower in 2 rooms)
LEG342	4	Environmental	Cluster 6	13/10/2006	ST110	
LEG345	4	Environmental	Cluster 6	13/10/2006	ST110	
LEG338	4	Clinical	Cluster 6	13/10/2006	ST110	
LEG362	5	Clinical	Cluster 3	15/10/2008	ST110	Clinical isolates from 2 hospitalised patients with the corresponding environmental isolates from the suspected source (Shower and sink from the patients' room)
LEG363	5	Clinical	Cluster 3	6/10/2008	ST110	
LEG364	5	Environmental	Cluster 3	3/11/2008	ST110	
LEG366	5	Environmental	Cluster 3	3/11/2008	ST110	
LEG370	6	Environmental		2/03/2009	ST1	Clinical isolates from 1 hospitalised patient with the corresponding environmental isolate from the suspected source (Water sample from the patient's room)
LEG369	6	Clinical		27/02/2009	ST345	
LEG379	7	Environmental		16/03/2010	ST1333	Clinical isolates from 1 hospitalised patient with the corresponding environmental isolate from the suspected source (Water sample from the hospital)
LEG378	7	Clinical		29/03/2010	ST68	
LEG414	8	Clinical	Cluster 7	5/01/2012	ST188	From January 2012 to April 2013: 4 consecutive patients with positive <i>Lp</i> at the same hospital, from which only 2 isolates (2 patients) were kept. In April 2013, numerous water samples at different sites of the hospital were <i>L.pn.</i> pos. SBT investigation of water strains showed the presence of several ST's from which ST87 and ST188 matching the 2 patients' strains respectively
LEG435	8	Environmental	Cluster 7	24/10/2012	ST188	
LEG436	8	Environmental	Cluster 7	24/10/2012	ST188	
LEG437	8	Environmental	Cluster 7	24/10/2012	ST188	
LEG417	8	Clinical	Cluster 12	26/03/2012	ST87	
LEG425	8	Environmental	Cluster 12	24/10/2012	ST87	
LEG443	8	Environmental		11/04/2013	ST87	
LEG421	9	Clinical	Cluster 9	20/08/2012	ST1387	Clinical isolates from 1 hospitalised patient with the corresponding environmental isolates from the suspected source (Water samples from the sink at the patient's room)
LEG422	9	Environmental	Cluster 9	24/08/2012	ST1387	
LEG423	9	Environmental	Cluster 9	24/08/2012	ST1387	

Table 1 (continued)

Strain number	outbreak/group number	Source	Cluster ID cg-wgMLST	Collection date	SBT ST (sanger sequencing)	Epidemiological information
LEG587	10	Clinical		12/07/2017	ST2458	An outbreak occurred in a small town in Belgium involving few cases. The environmental source could not be defined. These 2 strains were isolated from 2 patients, for a 3rd patient we found ST291 based on SBT directly on sample. These 3 ST's share 5 out of the 7 alleles, a probable association?
LEG586	10	Clinical		17/07/2017	ST2461	
LEG692	11	Clinical	Cluster 5	1/11/2017	ST6	Clinical isolates from 2 hospitalised patients with the corresponding environmental isolates from the suspected source (Water from the patients' room)
LEG693	11	Clinical	Cluster 5	30/10/2017	ST6	
LEG694	11	Environmental	Cluster 5	28/10/2017	ST6	
LEG695	11	Environmental	Cluster 5	28/10/2017	ST6	
LEG738	12	Clinical	Cluster 11	2/01/2018	ST1362	Clinical isolate from 1 hospitalised patient with the corresponding environmental isolates from the suspected source (Water samples from the hospital)
LEG739	12	Environmental	Cluster 11	25/01/2018	ST1362	
LEG741	12	Environmental		25/01/2018	ST1362	
LEG743	13	Clinical	Cluster 10	24/01/2018	ST1443	1 Clinical isolate from a patient at a nursing home and the corresponding environmental isolates from the suspected source (Water sample from the nursing home)
LEG744	13	Environmental	Cluster 10	14/02/2018	ST1443	
LEG745	13	Environmental	Cluster 10	14/02/2018	ST1443	
LEG750	14	Clinical	Cluster 1	5/11/2002	ST1	Clinical isolate from 1 hospitalised patient with the corresponding environmental isolates from the suspected source (Water samples from the hospital). Same hospital as Outbreak_Nosocomial_1985-1987
LEG751	14	Environmental	Cluster 1	2002	ST1	
LEG752	15	Clinical	Cluster 4	19/03/2003	ST6	Clinical isolate from 1 hospitalised patient with the corresponding environmental isolates from the suspected source (Water samples from the hospital)
LEG753	15	Environmental	Cluster 4	2003	ST6	
LEG754	15	Environmental	Cluster 4	2003	ST6	
LEG755	16	Clinical	Cluster 1	20/09/2003	ST1	Clinical isolate from 1 hospitalised patient with the corresponding environmental isolates from the suspected source (Water samples from the hospital). Same hospital as Outbreak_Nosocomial_1985-1987
LEG756	16	Environmental	Cluster 1	2003	ST1	
LEG757	17	Environmental	Cluster 4	2011	ST6	Clinical isolate from 1 hospitalised patient with the corresponding environmental isolates from the suspected source (Water samples from the hospital) Same hospital as for LEG752, LEG753, LEG754
LEG758	17	Clinical	Cluster 4	1/06/2011	ST6	
LEG563	18	Clinical		21/09/2011	ST1	Clinical isolate from 1 hospitalised patient with the presumably corresponding environmental isolates from the suspected source (Water samples from the hospital) Same hospital as Outbreak_Nosocomial_1985-1987
LEG759	18	Environmental	Cluster 1	2011	ST1	
LEG724	19	Clinical	Cluster 13	6/10/2016	ST23	1 Clinical isolate from a patient at a nursing home and the corresponding environmental isolates from the suspected source (Water sample from the nursing home)
LEG729	19	Environmental	Cluster 13	2016	ST23	
LEG523	20	Clinical		1985–1987	ND\$	Control for Outbreak_1985-1987*

Table 1 (continued)

Strain number	outbreak/group number	Source	Cluster ID cg-wgMLST	Collection date	SBT ST (sanger sequencing)	Epidemiological information
LEG508	20	Environmental	Cluster 1	1985–1987	ST1	Outbreak_Nosocomial_1985-1987
LEG509	20	Environmental	Cluster 1	1985–1987	ST1	Outbreak_Nosocomial_1985-1987
LEG510	20	Clinical	Cluster 1	1985–1987	ST1	Outbreak_Nosocomial_1985-1987
LEG511	20	Clinical	Cluster 1	1985–1987	ST1	Outbreak_Nosocomial_1985-1987
LEG512	20	Clinical	Cluster 1	1985–1987	ST1	Outbreak_Nosocomial_1985-1987
LEG514	20	Clinical	Cluster 1	1985–1987	ST1	Outbreak_Nosocomial_1985-1987
LEG515	20	Clinical	Cluster 8	1985–1987	ST1	Outbreak_Nosocomial_1985-1987
LEG516	20	Clinical	Cluster 1	1985–1987	ST1	Outbreak_Nosocomial_1985-1987
LEG517	20	Clinical	Cluster 1	1985–1987	ST1	Outbreak_Nosocomial_1985-1987
LEG519	20	Clinical	Cluster 1	1985–1987	ST1	Outbreak_Nosocomial_1985-1987
LEG521	20	Clinical	Cluster 1	1985–1987	ST1	Outbreak_Nosocomial_1985-1987
LEG522	20	Clinical	Cluster 1	1985–1987	ST1	Outbreak_Nosocomial_1985-1987
LEG513	20	Clinical		1985–1987	ST146	Control for Outbreak_1985-1987*
LEG520	20	Clinical		1985–1987	ST22	Control for Outbreak_1985-1987*
LEG518	20	Clinical		1985–1987	ST42	Control for Outbreak_1985-1987*
LEG533	21	Clinical	Cluster 12	09/09/2015	ST20	2 clinical isolates from the same hospitalised patient isolated at 2 weeks interval
LEG534	21	Clinical	Cluster 12	22/09/2015	ST20	
LEG585	21	Clinical		30/06/2017	ST42	1 clinical isolate from a hospitalised patient
REF134 (ATCC33152)	NA	NA		NA	ST36	Reference strain taken as control

*: The isolates LEG513, LEG518, LEG520 and LEG523 were isolated from patients with community acquired pneumonia hospitalised at the same hospital where the outbreak occurred. These were originally taken as control strains for the analyses of the outbreak related isolates.

\$: The sequence type was not determined for this isolate.

Hilden, Germany) was used to extract and purify the proper amount of genomic DNA needed for the library preparation according to the manufacturer's protocol. DNA concentration was determined using the Qubit dsDNA HS (or BR) assay kit (ThermoFisher Scientific, Waltham, MA, USA) and the Qubit 2.0 Fluorometer (ThermoFisher Scientific, Waltham, MA, USA). The NanoDrop 2000C spectrophotometer (ThermoFisher Scientific, Waltham, MA, USA) was used for extracted DNA quality control verification through 260/280 ratio determination.

Library preparation

Kapa HyperPlus Library Preparation Kit (Kapa Biosystems, Wilmington, MA, USA) was used. Quality control and pooling of the library was performed employing a 2100 BioAnalyzer (Agilent, Santa Clara, CA, USA), a Qubit 2.0 Fluorometer (ThermoFisher Scientific) and the KAPA Illumina Library Quantification Kit (Kapa Biosystems, Wilmington, MA, USA). The library was pooled to a final concentration of 2 nM after denaturation with 0,2N NaOH. A 1% PhiX control library was spiked in, according to instructions for Illumina sequencing (Illumina, San Diego, CA, USA).

Sequencing run

The sequencing was performed at the following Belgian WGS platform: BRIGHTcore (Brussels Interuniversity Genomics & High Throughput core). MiSeq or Hiseq (Illumina, San Diego, CA, USA) sequencing systems were used with the MiSeq Reagent kit v2 (500 cycle): 2 × 250 bp read-length, 39 h runtime, 7.5–8.5 Gb output, aiming for a 100-fold coverage. Sequence quality was assessed with FastQC (version 0.11.4) software (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>).

Data analysis

WGS data were further analysed using 3 analysis schemes. From the two commercially available data analysis tools the following schemes were used: “Ridom SeqSphere +” cgMLST (v.5.1.1) (input data: SPADES preassembled WGS data), “Applied Maths-Bionumerics” (wgMLST) (v.8.1) (input data: FASTQ files). cgMLST and wgMLST are both based on the publication of Moran Gilad et al. [28]. For cgMLST above four allelic differences [28], the strains were considered not to cluster together and for wgMLST above

10 allelic differences (Default parameters used). In addition, the 50 loci cgMLST scheme (Still in validation process by CDC/ESGLI_ESCMID) was also used. This tool initially computed allelic difference distance matrices using binary scoring at each locus. Identical alleles were assigned a score of '0', while different alleles were assigned a score of '1'. To qualify for an allele call, a minimum threshold of 75% sequence coverage and 85% sequence identity was required. Post allele calling, normalization procedures were implemented to alleviate the influence of the differing number of comparisons. The normalization process ensured the genetic distances obtained were consistent regardless of the number of loci being compared. The normalized scores, which were combined for the 50 loci, quantified the genetic differences between the genomes. The distances were utilized as input for an Unweighted Pair Group Method with Arithmetic Mean (UPGMA) clustering analysis. This analysis produced a Newick format (nwk) file that depicted the genetic relationships derived from the cgMLST data. The nwk file was inputted into the GrapeTree software to generate a Minimum Spanning Tree (MST) visualization. This visualization effectively illustrates the clonal relationships between bacterial isolates based on the genetic distances obtained from the cgMLST analysis.

Results

WGS typing using “Ridom SeqSphere +” cgMLST

See Fig. 1 for the MST obtained with the Ridom Seq-Sphere+™ cgMLST scheme. According to this scheme all the 78 isolates data had 98 to 100% of successful 1521 targets (the cut-off of acceptable data being 95% [28]. The MST was constructed based on the 1412 targets present in all analysed genomes.

As expected, the 7 ST36 isolates of the outbreak from 1999 were classified within the same cluster (cluster 2). All but one of the 12 isolates of the ST1 outbreak from 1985–1987 were classified within the same cluster (cluster 1). Leg515 from the outbreak 1985–1987 clustered by Leg321 and Leg323 (cluster 8) isolated from two patients at four months interval in 2003–2004 at another hospital also located in the same city as the nosocomial outbreak from 1985–1987. Leg750/Leg751 and Leg755/Leg756 were clinical/environmental pairs isolated during hospitalisation in 2002 and 2003 respectively in the same hospital of the outbreak from 1985–1987, now classified within the epidemic cluster. Leg759 is the environmental isolate presumed to be the source of infection during a

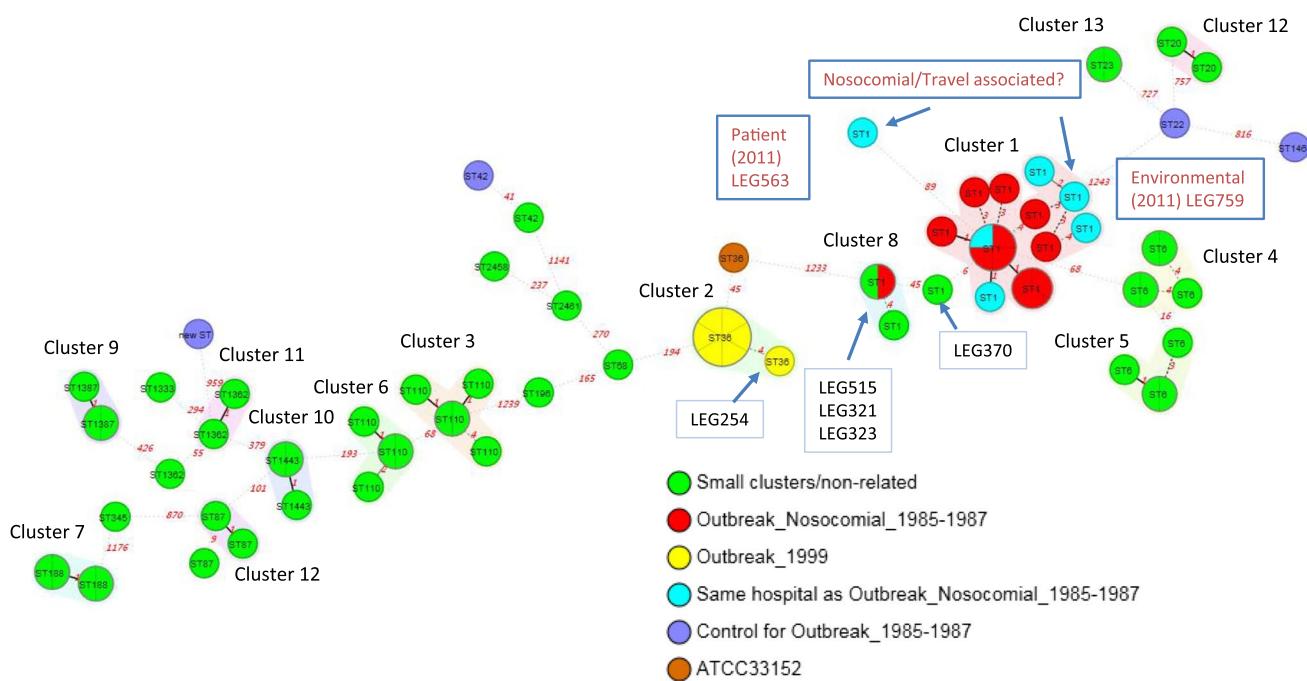


Fig. 1 Minimum spanning tree generated by cgMLST ridom seq-sphere+ *Legionella pneumophila* scheme based on allelic profiles of 1412 genes present in all 78 analysed genomes. Circles correspond to the sequenced strains. Inside the circles, the strains are labelled by their ST. The number of allelic differences are indicated beside the branches. Strains with no allelic differences are classified in one cir-

cle. Strains classified in the same cluster are connected by a coloured background. For ST1 strains: Based on this MST, it could be shown that the ST1 environmental strain isolated from the hospital in 2011 was not the source of contamination of the patient hospitalised at that period (LEG759/LEG563)

patient hospitalisation in 2011 also at the same hospital of the outbreak from 1985–1987. However, the patient's isolate Leg563, initially related to the outbreak, has shown 89 allelic differences compared to the presumed environmental source. The last tested ST1 isolate (Leg370) was originally the presumed environmental source for the patient's isolate Leg369. However, both SBT and cgMLST agreed on them being unrelated isolates.

Regarding the two ST110 clusters (Cluster 3 and cluster 6), the first one is linked to a hospital in a small town in the north of Belgium (three clinical isolates and two environmental isolates). Four of the five isolates were isolated in 2008 while clinical isolate (Leg311) was found earlier in 2003. The second ST110 cluster corresponds to a nosocomial context from a nursing home in Brussels (two clinical isolates and two environmental isolates). On the MST analysis, the two clusters of the two different locations (hospital/nursing home) were classified in two ST110 separate clones respectively.

WGS typing using “Applied Maths-Bionumerics” wgMLST

See Supplementary Table 1 for the dataset quality scores determined by Applied Maths-Bionumerics software. Good WGS quality scores were obtained for all the 78 isolates. Regarding the Applied Maths-Bionumerics’ wgMLST scheme, all the 78 isolates WGS data had only 48 to 52% successfully assigned targets from the total of 5770 targets. MST was constructed based on the successful targets present in the analysed genomes. See Fig. 2 for the MST obtained with the “Applied Maths-Bionumerics” wgMLST scheme.

According to this scheme, the isolates from the two outbreaks (1985–1987 and 1999) were as expected classified in two clusters respectively. For the remaining isolates the same number of clusters as with Seqsphere scheme could be found, with for some isolates a slightly higher/lower number of allelic differences were observed. For instance, LEG370 stands within cluster 1 while it was outside Cluster 1 with the cgMLST scheme and LEG254 is located within cluster 2 with no allelic difference with other outbreak isolates while for cgMLST scheme, LEG254 has 4 allelic differences with the other isolates of the outbreak.

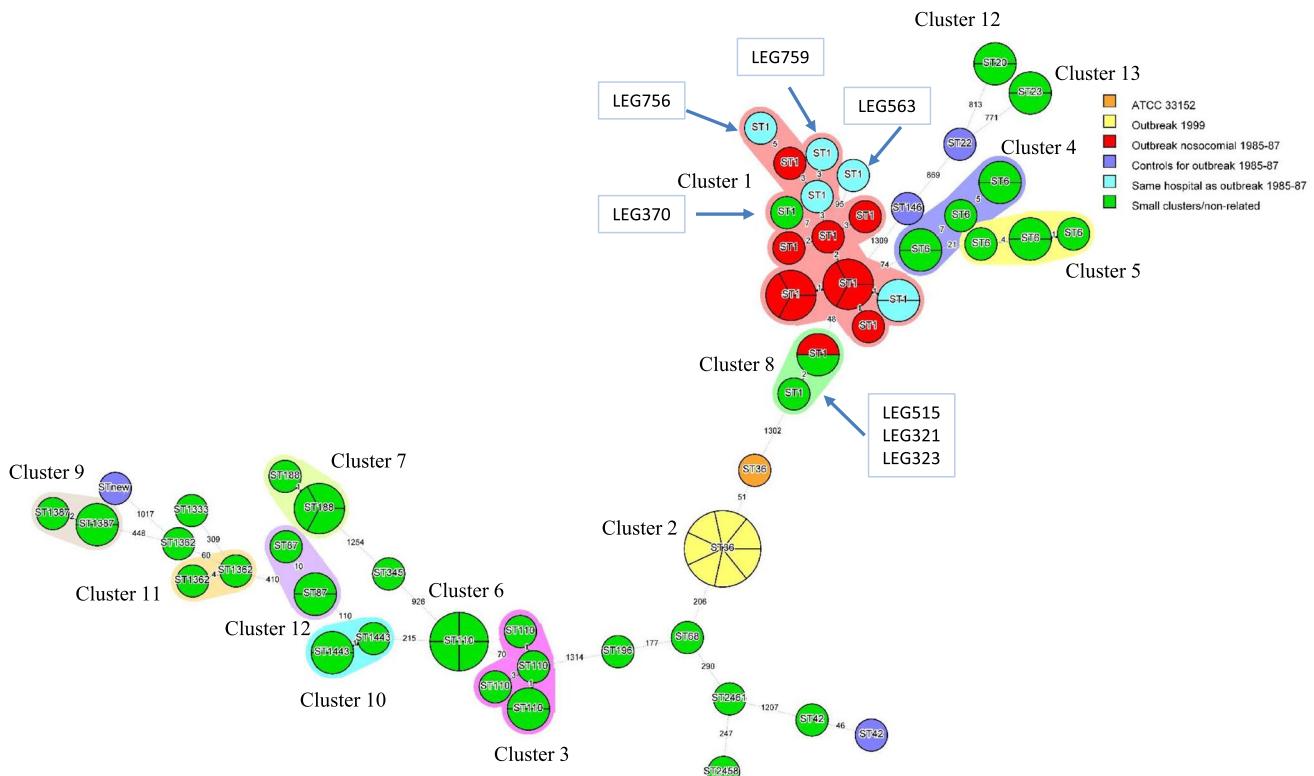


Fig. 2 Minimum spanning tree generated by wgMLST Applied Maths, Bionumerics *Legionella pneumophila* scheme based on allelic profiles of all 78 analysed genomes. Circles correspond to the sequenced strains. Inside the circles, the strains are labelled by their

strain ST. The number of allelic differences are indicated beside the branches. Strains with no allelic differences are classified in one circle. Strains classified in the same cluster are grouped in circles with coloured background

WGS typing using 50 loci cgMLST scheme

According to the MST generated with the 50 loci cgMLST scheme, see Fig. 3, the isolates from the two outbreaks (1985–1987 and 1999) were again as expected classified in two clusters respectively. For the remaining isolates very similar results as the two other schemes were obtained regarding the relatedness/non-relatedness of the analysed isolates. For ST1 isolates, in comparison to the 1521 loci cgMLST; Leg756 and Leg759 (isolates from the same hospital as for the outbreak from 1985–1987) were classified within the epidemic cluster in contrast to Leg755 which was outside the cluster but still very close (1 allele difference). The ST1 isolates from the presumed pair Leg759/Leg563 were classified at 8 allele differences, in agreement with the two other typing tools. In addition to the MST, the 50 loci cgMLST scheme allowed for the division and classification of the isolates into subtypes, see Fig. 4 and Supplementary Table 2.

For a genomic profile analysis involving 50 loci with allele ID, refer to Supplementary Table 3. The assessment covered 78 genomes subjected to cgMLST analysis, where 19 of these genomes exhibited non-called alleles at one or more loci because they did not meet the allele calling criteria of 75% coverage and 85% sequence identity. In total,

26 non-called alleles were identified across these genomes. Despite this, all isolates could be assigned to a subtype using the 50 loci cgMLST scheme. The presence of non-called alleles at any locus was considered a unique allele state in the analysis, thus contributing to the allele profile of the genomes. The cgMLST analysis effectively differentiated the 21 traditional sequence types (STs) into 34 subtypes, based on observed allelic variations, inclusive of the non-called alleles.

Discussion

In this study, concordant results were obtained by the three used WGS typing tools for clustering analysis of *Lp* isolates with epidemiological relatedness. These results were also concordant with the results of traditional SBT method. Genomes of isolates belonging to the same ST were clustering near to each other and no discordant results were observed. Moreover, WGS typing allowed for a better discrimination within ST's.

Regarding the rate of successfully assigned loci needed for analysis, the 50 loci cgMLST scheme, harboured only 0.67% missing alleles across all tested genomes. For the 1521 loci cgMLST, only 7% of the genes failed to be

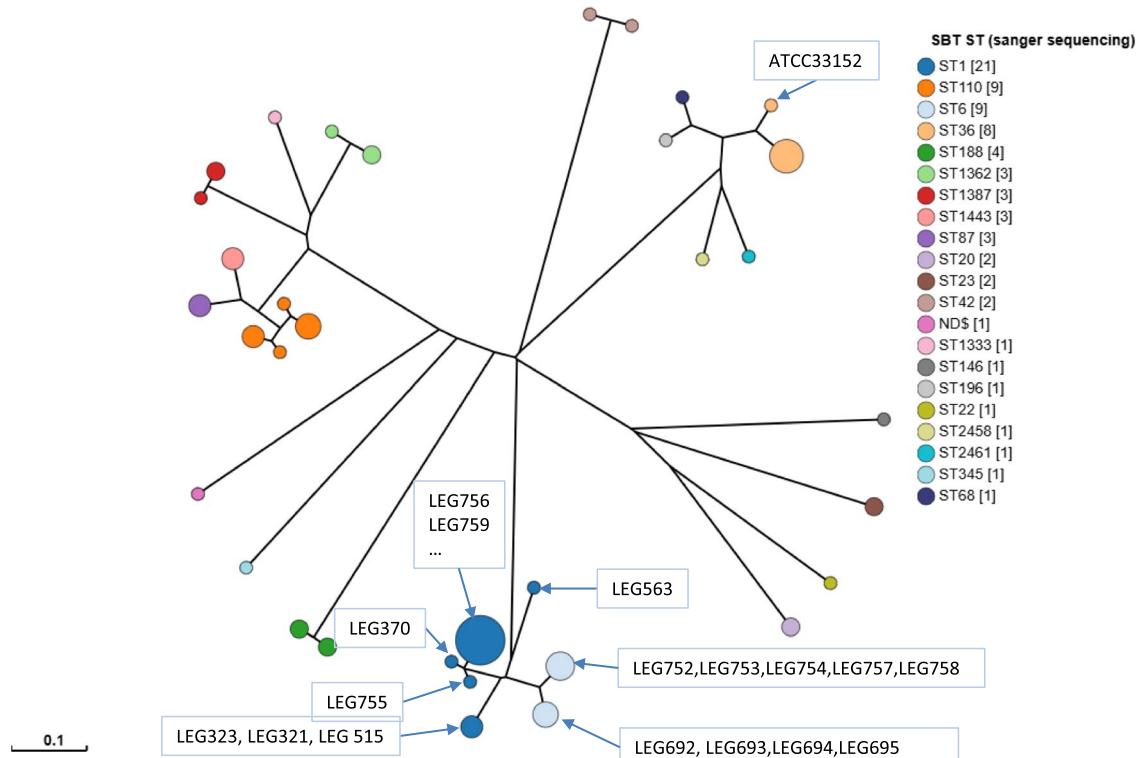
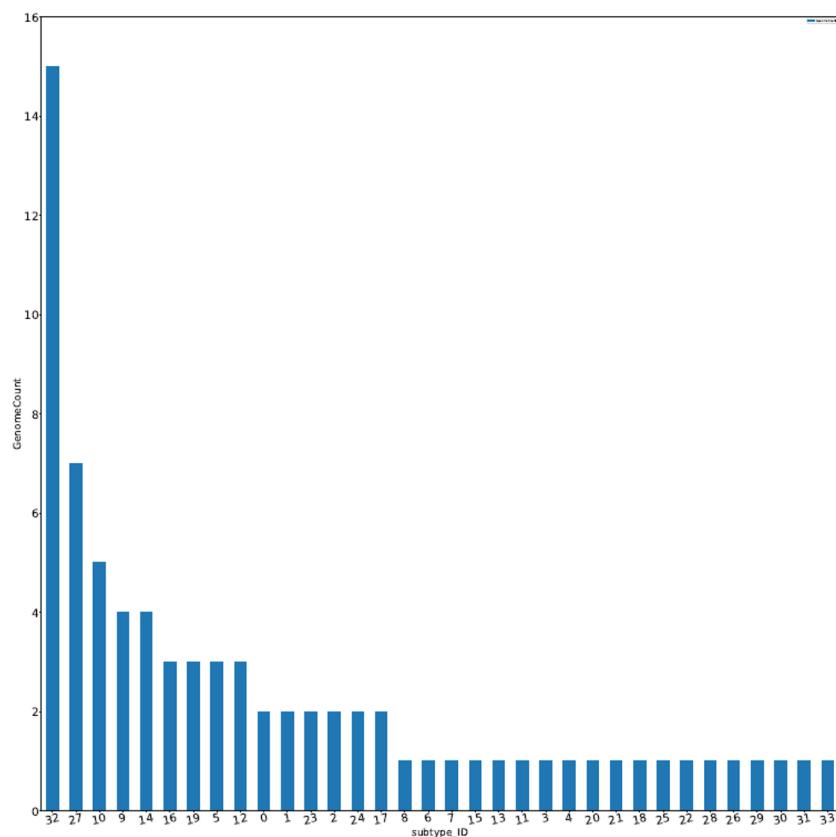


Fig. 3 MST of the 50 loci cgMLST for the 78 genomes. Minimum spanning tree generated by the 50 loci cgMLST *Legionella pneumophila* scheme based on allelic profiles of all 78 analysed genomes.

Circles correspond to the sequenced strains. The scores beside the branches correspond to the normalized dissimilarity. Strains with no allelic differences are classified in one circle

Fig. 4 The 78 isolates belonging to 21 Sequence Types (ST) were classified into 34 subtypes as shown in this graphic (Genome count/subtype_ID) by using the 50cgMLST scheme. As follows, a list of the subtypes included in each ST: ST1: 5, 6, 7, 32, 33; ST6: 9, 10; ST22: 4; ST42: 25, 26; ST87: 16; ST110: 11, 12, 13, 14; ST188: 0, 1; ST196: 28; ST20: 24; ST23: 23; ST36: 27, 31; ST68: 29; ST1333: 20; ST1362: 2, 3; ST1387: 17, 18; ST1443: 19; ST2458: ST 8; ST2461: 30; ST345: 15; ST146: 22; and one strain with unknown ST: 21



assigned, but for the 5770 loci wgMLST, the failure affected about 50% of the genes. The high failure rate of the 5770 loci wgMLST has resulted in a low difference in discrimination between this scheme and the 1521 loci cgMLST, as one would expect a higher discrimination rate when the clustering analysis is based on a much larger number of the compared genes. Since not all accessory genes are present in all the isolates, this can explain the obtained result and tends to give more advantages to a core genome analysis.

In addition to the good discrimination for clustering, the 50 loci was able to assign a subtype to all the 78 genomes. This is a big advantage in terms of standardisation and comparison between different laboratories for future epidemiological investigations. Results of current study also confirmed what was expected by David S et al. [29], meaning that the 50 loci cgMLST would offer the best compromise between improving the discrimination obtainable by WGS methods and maintaining good epidemiological concordance, without the need to use thresholds or clustering methods as the isolates comparison would be based on subtype numbers just like SBT method. Furthermore, as this 50 loci cgMLST scheme is also containerized within Docker as a portable analysis tool, we could expect the quick application of this scheme without excessive computer power.

The WGS typing by the three tools showed that one isolate, Leg515, from the outbreak 1985–1987 did not cluster

within the epidemic clone but clustered by Leg321 and Leg323 recovered more than 15 years later from two patients at another hospital also located in the same city as the nosocomial outbreak from 1985–1987. Originally, this isolate gave different MAbs subtyping results as well as a slightly different genotyping profile than the remaining epidemic isolates by the typing methods available at that time [30]. So, the current result by using the new WGS typing tools was a confirmation of what was rather expected. Regarding the clustering of this isolate with isolates recovered many years later at a completely different location is intriguing, however, similar results were obtained in a previous study where it was demonstrated that identical environmental isolates can be found in different sampling locations, with no direct water pipe connection [33].

Leg750/leg751 and Leg755/756 were clinical/environmental pairs isolated during hospitalisation in 2002 and 2003 in the same hospital where the 1985–1987 outbreak took place. They could now be classified within the epidemic cluster with more certainty (Figs. 1 and 2). Regarding the *Lp* ST110 (cluster 3), the same observation as for cluster 1 can be made, as four of the five isolates were isolated in 2008 while the remaining isolate was found five years earlier, in 2003. This suggests that the *Lp* epidemic clone remained present in the water system of the hospital for several years despite thorough decontamination

procedures, as previously demonstrated in nosocomial cases investigations [34]. This could be either due to sub-optimal disinfection or other conditions favouring the growth and transmission of this strain. Despite the multiple observation of such clone reappearance over time, few is known about the triggers for *Lp* proliferation in pipes [35]. However, in this ST110 case, WGS typing tools could clearly discriminate the ST110 nonrelated isolates from the two different locations (hospital/nursing home). This showed once more the ability of the WGS tools to allow for accurate epidemiological conclusions due to the higher resolution in comparison to SBT.

Another very important advantage that could be demonstrated in the current study regarding the WGS typing tools is their ability to discriminate ST1 unrelated isolates. The following case demonstrates this outcome very clearly: Leg759 is the environmental isolate originally presumed to be the source of infection during a patient hospitalisation in 2011 at the hospital where the outbreak occurred in 1985–1987. Based on SBT results only, we could not conclude whether this infection was nosocomial, or travel associated. By using the current WGS tools Leg563 has been classified far (89 allelic difference using the 1521 cgMLST) from the presumed environmental source while the Leg759 has been clustered by the old epidemic isolates. This result has now elucidated the remaining question regarding the source of the patient infection. Nevertheless, despite a better discrimination of ST1, the lack of clear cutoff for clustering can lead to remaining unclarified cases and should be further explored by increasing ST1 WGS sequencing in routine [36].

In conclusion, concordant results were obtained by the three WGS typing tools. WGS tools demonstrated that isolates belonging to previously documented nosocomial outbreak clones remain present in the same hospital for many years. In addition, a non-elucidated case by SBT could now be clearly answered as being travel associated instead of nosocomial. These results were also concordant with the results of traditional SBT method but much more discriminatory. Particularly, the 50 loci cgMLST as it was stated in the study of David et al. [29], gave enough resolution for the typing of *Lp* isolates despite the lower number of the analysed alleles. These results are encouraging for its use as a standardized scheme for *Lp* genotyping. The known outbreaks and clusters could be detected and for the panel of *Lp* isolates tested in the current study, clear discrimination of ST1 non-related isolates was obtained. However, more studies with a higher number of isolates should be performed to verify ST1 discrimination ability by these three WGS typing tools.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10096-024-05013-4>.

Author contributions FE, SP, CM, FC and DP drafted the paper. FE and SP performed WGS data analyses. AM, OS, BP and IW contributed to the revision of the paper.

Funding This work was performed in the frame of the Belgian National Reference Centre for Legionella supported by the Belgian Ministry of Social Affairs through a fund within the Health Insurance System.

Data availability All the obtained WGS raw data were submitted to the European nucleotide archive (<http://www.ebi.ac.uk/ena/>) of European molecular biology laboratory (EMBL) European Bioinformatics Institute (EBI) under the study access number PRJEB52784.

Declarations

Ethical statement Epidemiological data were collected anonymously in the frame of Decision No 2119/98/EC of the European Parliament and of the Council, concerning the epidemiological surveillance and control of communicable diseases in the Community, as completed by Decision No 1082/2013/EU.

Competing Interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- Fraser DW (1980) Legionellosis: evidence of airborne transmission. Ann N Y Acad Sci 353:61–66. <https://doi.org/10.1111/j.1749-6632.1980.tb18906.x>
- Horwitz MA, Silverstein SC (1980) Legionnaires' disease bacterium (*Legionella pneumophila*) multiples intracellularly in human monocytes. J Clin Invest 66(3):441–450. <https://doi.org/10.1172/JCI109874>
- Horwitz MA (1983) The Legionnaires' disease bacterium (*Legionella pneumophila*) inhibits phagosome-lysosome fusion in human monocytes. J Exp Med 158(6):2108–2126. <https://doi.org/10.1084/jem.158.6.2108>
- Fraser DW et al (1977) Legionnaires' disease: description of an epidemic of pneumonia. N Engl J Med 297(22):1189–1197. <https://doi.org/10.1056/NEJM197712012972201>
- Glick TH, Gregg MB, Berman B, Mallison G, Rhodes WW Jr, Kassanoff I (1978) Pontiac fever. An epidemic of unknown etiology in a health department: I. Clinical and epidemiologic aspects. Am J Epidemiol 107(2):149–160. <https://doi.org/10.1093/oxfordjournals.aje.a112517>
- Heath CH, Grove DI, Looke DF (1996) Delay in appropriate therapy of *Legionella* pneumonia associated with increased mortality.

- Eur J Clin Microbiol Infect Dis 15(4):286–290. <https://doi.org/10.1007/BF01695659>
- 7. Benin AL, Benson RF, Besser RE (2002) Trends in legionnaires' disease, 1980–1998: declining mortality and new patterns of diagnosis. Clin Infect Dis 35(9):1039–1046. <https://doi.org/10.1086/342903>
 - 8. J. W. Den Boer et al. (2002) A large outbreak of Legionnaires' disease at a flower show, the Netherlands, 1999. Emerg Infect Dis 8(1):37–43. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed/11749746>
 - 9. Carratala J et al (1994) Risk factors for nosocomial Legionella pneumophila pneumonia. Am J Respir Crit Care Med 149(3 Pt 1):625–629. <https://doi.org/10.1164/ajrccm.149.3.8118629>
 - 10. Beaute JN (2017) The European legionnaires' disease surveillance, "legionnaires' disease in Europe, 2011 to 2015. Euro Surveill22(27). <https://doi.org/10.2807/1560-7917.ES.2017.22.27.30566>
 - 11. Barskey AE, Derado G, Edens C (2022) Rising incidence of legionnaires' disease and associated epidemiologic patterns, United States, 1992–2018. Emerg Infect Dis 28(3):527–538. <https://doi.org/10.3201/eid2803.211435>
 - 12. ECfDp a. C.-. ECDC (2023) Legionnaires' disease Annual Epidemiological Report for 2021 [Internet]. Stockholm, Sweden," Available from: <https://www.ecdc.europa.eu/sites/default/files/documents/legionnaires-disease-annual-epidemiological-report-2021.pdf>
 - 13. Jong HLB (2021) European Surveillance of Legionnaires' Disease. Curr Issues Mol Biol 42:81–96. <https://doi.org/10.21775/cimb.042.081>
 - 14. Jomehzadeh N, Moosavian M, Saki M, Rashno M (2019) Legionella and Legionnaires' disease : An overview. J Acute Disease 13
 - 15. Rizzardi K, Winiecka-Krusnell J, Ramliden M, Alm E, Andersson S, Byfors S (2015) Legionella norrlandica sp. nov., isolated from the biopurification systems of wood processing plants. Int J Syst Evol Microbiol 65(Pt 2):598–603. <https://doi.org/10.1099/ijst.0.068940-0>
 - 16. Yu VL et al (2002) Distribution of Legionella species and serogroups isolated by culture in patients with sporadic community-acquired legionellosis: an international collaborative survey. J Infect Dis 186(1):127–128. <https://doi.org/10.1086/341087>
 - 17. Edelstein PH, Luck C (2015) Legionella (Chapter 49) Manual of Clinical Microbiology (11th ed. pp. 887). Eds. Jorgensen JH, Carroll KC, Pfaller MA, Landry ML, Richter SS, Warnock DW. ASM Press
 - 18. Helbig JH et al (Oct2002) Pan-European study on culture-proven Legionnaires' disease: distribution of Legionella pneumophila serogroups and monoclonal subgroups. Eur J Clin Microbiol Infect Dis 21(10):710–716. <https://doi.org/10.1007/s10096-002-0820-3>
 - 19. Luck PC, Kohler J, Maiwald M, Helbig JH (May1995) DNA polymorphisms in strains of Legionella pneumophila serogroups 3 and 4 detected by macrorestriction analysis and their use for epidemiological investigation of nosocomial legionellosis. Appl Environ Microbiol 61(5):2000–2003. <https://doi.org/10.1128/AEM.61.5.2000-2003.1995>
 - 20. Luck C, Fry NK, Helbig JH, Jarraud S, Harrison TG (2013) Typing methods for legionella. Methods Mol Biol 954:119–148. https://doi.org/10.1007/978-1-62703-161-5_6
 - 21. Gaia V et al (May2005) Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of Legionella pneumophila. J Clin Microbiol 43(5):2047–2052. <https://doi.org/10.1128/JCM.43.5.2047-2052.2005>
 - 22. Ratzow S, Gaia V, Helbig JH, Fry NK, Luck PC (Jun2007) Addition of neuA, the gene encoding N-acylneuraminate cytidylyl transferase, increases the discriminatory ability of the consensus sequence-based scheme for typing Legionella pneumophila serogroup 1 strains. J Clin Microbiol 45(6):1965–1968. <https://doi.org/10.1128/JCM.00261-07>
 - 23. Phin N et al (Oct2014) "Epidemiology and clinical management of Legionnaires' disease," (in English). Lancet Infectious Diseases 14(10):1011–1021. [https://doi.org/10.1016/S1473-3099\(14\)70713-3](https://doi.org/10.1016/S1473-3099(14)70713-3)
 - 24. Ginevra C et al (Mar2012) Legionella pneumophila sequence type 1/Paris pulsotype subtyping by spoligotyping. J Clin Microbiol 50(3):696–701. <https://doi.org/10.1128/JCM.06180-11>
 - 25. Reuter S et al. (2013) A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak. BMJ Open 3(1). <https://doi.org/10.1136/bmjopen-2012-002175>
 - 26. Gilchrist CA, Turner SD, Riley MF, Petri WA Jr, Hewlett EL (Jul2015) Whole-genome sequencing in outbreak analysis. Clin Microbiol Rev 28(3):541–563. <https://doi.org/10.1128/CMR.00075-13>
 - 27. Tran A, Rowlinson M-C (2019) Application of next-generation sequencing in public health epidemiology and outbreak investigation. Adv Mol Pathol 2:89–97
 - 28. Moran-Gilad J et al. (2015) Design and application of a core genome multilocus sequence typing scheme for investigation of Legionnaires' disease incidents. Euro Surveill 20(28). <https://doi.org/10.2807/1560-7917.es2015.20.28.21186>
 - 29. David S et al (2016) Evaluation of an Optimal Epidemiological Typing Scheme for Legionella pneumophila with Whole-Genome Sequence Data Using Validation Guidelines. J Clin Microbiol 54(8):2135–2148. <https://doi.org/10.1128/JCM.00432-16>
 - 30. Struelens MJ et al (1992) Genotypic and phenotypic methods for the investigation of a nosocomial Legionella pneumophila outbreak and efficacy of control measures. J Infect Dis 166(1):22–30. <https://doi.org/10.1093/infdis/166.1.22>
 - 31. De Schrijver K et al (Mar2003) An outbreak of Legionnaire's disease among visitors to a fair in Belgium in 1999. Public Health 117(2):117–124. [https://doi.org/10.1016/S0033-3506\(02\)00011-2](https://doi.org/10.1016/S0033-3506(02)00011-2)
 - 32. Vekens E et al. (2012) Sequence-based typing of Legionella pneumophila serogroup 1 clinical isolates from Belgium between 2000 and 2010. Euro Surveill 17(43):20302. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed/23137466>
 - 33. Wuthrich D et al. (2019) Air-conditioner cooling towers as complex reservoirs and continuous source of Legionella pneumophila infection evidenced by a genomic analysis study in 2017, Switzerland. Euro Surveill 24(4). <https://doi.org/10.2807/1560-7917.ES.2019.24.4.1800192>
 - 34. David S et al (2017) Seeding and Establishment of Legionella pneumophila in Hospitals: implications for genomic investigations of nosocomial legionnaires' disease. Clin Infect Dis 64(9):1251–1259. <https://doi.org/10.1093/cid/cix153>
 - 35. Gleason JA, Cohn PD (2022) A review of legionnaires' disease and public water systems - scientific considerations, uncertainties and recommendations. Int J Hyg Environ Health 240:113906. <https://doi.org/10.1016/j.ijeh.2021.113906>
 - 36. Michel C et al (2024) From investigating a case of cellulitis to exploring nosocomial infection control of ST1 Legionella pneumophila using genomic approaches. Microorganisms 12(5). <https://doi.org/10.3390/microorganisms12050857>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.