Design and application of a core genome multilocus sequence typing scheme for investigation of Legionnaires' disease incidents

J Moran-Gilad (giladko@post.bgu.ac.il)[1,2,3], K Prior[4], E Yakunin[5], T G Harrison[6], A Underwood[6], T Lazarovitch[7], L Valinsky[5], C Lück[8], F Krux[4], V Agmon[5], I Grotto[1,3], D Harmsen[4]

1. Public Health Services, Ministry of Health, Jerusalem, Israel
2. Surveillance and Pathogenomics Israeli Centre of Excellence, National Institute for Biotechnology in the Negev, Ben-Gurion University of the Negev, Beer-Sheva, Israel
3. Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel
4. Department of Periodontology, University of Münster, Münster, Germany
5. Central Laboratories, Public Health Services, Ministry of Health, Jerusalem, Israel
6. Reference Microbiology Services, Public Health England, London, United Kingdom
7. Department of Clinical Microbiology, Assaf Harofeh Medical Centre, Zerifin, Israel
8. Institute of Medical Microbiology and Hygiene, University of Technology, Dresden, Germany

Citation style for this article:
Haar K, Amato-Gauci AJ. European men who have sex with men still at risk of HIV infection despite three decades of prevention efforts. Euro Surveill. 2015;20(14):pii=21087. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=21087

Article submitted on 20 November 2014 / published on 16 July 2015

Sequence-based typing (SBT) for *Legionella pneumophila* (Lp) has dramatically improved Legionnaires' disease (LD) cluster investigation. Microbial whole genome sequencing (WGS) is a promising modality for investigation but sequence analysis methods are neither standardised, nor agreed. We sought to develop a WGS-based typing scheme for Lp using de novo assembly and a genome-wide gene-by-gene approach (core genome multilocus sequence typing, cgMLST). We analysed 17 publicly available Lp genomes covering the whole species variation to define a core genome (1,521 gene targets) which was validated using 21 additional published genomes. The genomes of 12 Lp strains implicated in three independent cases of paediatric humidifier-associated LD were subject to cgMLST together with three 'outgroup' strains. cgMLST was able to resolve clustered strains and clearly identify related and unrelated strains. Thus, a cgMLST scheme was readily achievable and provided high-resolution analysis of Lp strains. cgMLST appears to have satisfactory discriminatory power for LD cluster analysis and is advantageous over mapping followed by single nucleotide polymorphism (SNP) calling as it is portable and easier to standardise. cgMLST thus has the potential for becoming a gold standard tool for LD investigation. Humidifiers pose an ongoing risk as vehicles for LD and should be considered in cluster investigation and control efforts.

Introduction

Legionellae are Gram-negative rods found in aqueous environments [1]. Humans become infected through exposure to contaminated aerosols originating from man-made water systems, such as spa pools, cooling towers or showering facilities in various settings such as healthcare, public and domestic facilities as well as occupational and travel-related settings. Clinical manifestation varies from a mild illness (Pontiac fever) to potentially fatal pneumonia known as Legionnaires' disease (LD) [2]. Among nearly 70 *Legionella* species described, *Legionella pneumophila* (Lp) causes the vast majority of LD and of 16 known serogroups, Lp serogroup 1 accounts for over 80% of LD cases and almost all clusters and outbreaks [3,4].

LD is a notifiable disease in all European Union and European Economic Area countries. Surveillance coordinated by the European Legionnaires' Disease Surveillance Network (ELDSNet) of the European Centre for Disease Prevention and Control (ECDC) demonstrates a mild increase in incidence of LD since 2001 [5,6]. The standardised sequence-based typing (SBT) scheme for Lp developed by the European Working Group for Legionella Infections (EWGLI, now European Society for Clinical Microbiology Study Group on Legionella Infections, ESGLI) marked an important advancement in the study of the molecular epidemiology of LD [7,8]. Implementation of this typing scheme, similar to multilocus sequence typing (MLST) has yielded useful and comparable data worldwide [9] and has been shown to be applicable in the investigation of unusual LD cases such as humidifier-associated LD [10] and legionellosis outbreaks [11].

The advent of next-generation sequencing (NGS) has revolutionised microbiology by making whole genome sequencing (WGS) of pathogens of public health importance, readily available [12]. The currently most significant role for NGS in microbiology is communicable disease surveillance and outbreak investigation. Many studies have demonstrated that whole genome comparisons provide far greater resolution for outbreak detection and microbial strain tracking than gold standard typing methods of different bacteria [13-15]. While most studies using WGS-based molecular epidemiology have relied on mapping of read data against

Table 1

Finished genomes[a] and assembled raw reads[b] used for *Legionella pneumophila* core genome definition (n=17)

Strain	SBT BAPS cluster	SBT ST (bySanger sequencing)	Mean coverage	NCBI/EBI accession number
Philadelphia[c]	13	36	N/A	NC_002942.5
Lens	2	15[d]	N/A	NC_006369.1
Lorraine	3	47	N/A	NC_018139.1
Paris	6	1[d]	N/A	NC_006368.1
Alcoy	11	578[d]	N/A	NC_014125.1
Corby	11	51	N/A	NC_009494.2
H093620212	1	46[d]	350.32	ERR315646
H090500162	4	611[d]	447.33	ERR315652
RR08000760	5	376[d]	359.47	ERR315654
H093380153	7	179	38.27	ERR315657
H100260089	8	44	486.84	ERR315660
Lansing-3	9	336	354.57	ERR315662
H063280001	10	23	387.44	ERR315663
H070840415	12	59[d]	463.52	ERR315666
H044500045	13	28[e]	520.93	ERR315669
H074360710	14	68[d]	418.93	ERR315671
H091960009	15	707[e]	391.10	ERR315672

BAPS: Bayesian analysis of population structure; EBI: European Bioinformatics Institute; N/A: not applicable; NCBI: National Center for Biotechnology Information; SBT: sequence-based typing; ST: sequence type.

[a] Finished genomes were from the NCBI database.

[b] Assembled raw reads were from EBI.

[c] Reference genome.

[d] Extraction of *mompS* allele from genomic data not possible due to multi-copy occurrence.

[e] Whole genome sequencing analysis corrected erroneous allelic profile of ST707 compared with original publication [24].

a reference followed by analysis of single nucleotide polymorphisms (SNPs), a de novo assembly and genome-wide gene-by-gene approach looking at allelic differences in core genome (cg) genes (cgMLST, or MLST+ as called in the SeqSphere+ software used for analysis) has been suggested as an alternative to SNP mapping [16,17].

There are limited data regarding the application of WGS for investigation of LD. Moreover, current experience is limited to analyses of SNPs and thus there is an unmet need for a cgMLST typing scheme for Lp that would enable a portable global nomenclature. Therefore, the goal of the current study was to set up, validate and apply a cgMLST scheme for Lp.

Methods

Standard laboratory work up of Legionella pneumophila strains

Isolates were cultured on BCYEα plates (Oxoid, Basingstoke, United Kingdom) for 48–72h at 35 °C before phenotypic and molecular tests were performed. Presumptive identification as Lp was confirmed using MonoFluo *Legionella pneumophila* indirect fluorescent antibody (IFA) Test (Biorad, Hemel Hempstead, United Kingdom). Lp serogroups and immunological subgroups (for selected serogroup 1 isolates) were determined using the Dresden panel of monoclonal antibodies [18]. Strains not readily confirmed as Lp were identified to species level by sequencing the *mip* gene as described by Ratcliff et al. [19] and comparing the sequence to the *mip* database [20].

Between two to three single colonies per Lp strain were selected and DNA extracted using the InstaGene Matrix (Biorad, Hemel Hempstead, United Kingdom). The genotype of each strain was determined using the M13 modification of the ESGLI SBT method by Sanger sequencing [21]. All alleles and sequence types (ST) were determined using the Legionella Sequence Quality Tool [22,23]. SBT was attempted on sputum in culture-negative cases using the direct or nested-SBT approach [21].

Whole genome sequencing and assembly

Whole genome shotgun sequencing was performed on 15 strains recovered from clinical and environmental samples in Israel. High molecular weight and quality DNA was extracted using the Wizard DNA purification kit (Promega, Madison, WI, United States). Sequencing libraries were prepared using the Nextera chemistry (Illumina Inc., San Diego, California, United States) for a 250 bp paired-end sequencing run on an Illumina MiSeq sequencer. Samples were sequenced to aim for a minimum coverage of 75-fold using Illumina's recommended standard protocols. All generated raw reads were submitted to the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena/) of the European Bioinformatics Institute under the study accession number PRJEB7140. After sequencing, the reads were quality-trimmed using the CLC Genomics Workbench software version 6.0 (CLC bio, Aarhus, Denmark) and then assembled de novo using CLC Genomics Workbench with default settings. The resulting assembly files were exported as ACE files and imported into SeqSphere+ software version 2.1 (Ridom GmbH, Germany).

Core genome multilocus sequene typing scheme definition and validation

For determining a cgMLST or MLST+ target set we aimed to cover the whole Lp species variation. By Bayesian Analysis of Population Structure (BAPS) based on more than 800 SBT STs, Underwood et al. recently reported 15 such Lp BAPS SBT clusters [24]. Therefore, we used for the cgMLST scheme definition, six finished genomes available from GenBank and 11 raw read datasets from the ENA archive that cover all BAPS SBT clusters (Table 1). ENA raw read data were again de novo assembled into draft genomes with CLC Genomics Workbench. The genome of strain Philadelphia (NC_002942.5) was used as a reference. To determine the cgMLST target gene set, a genome-wide gene-by-gene comparison was performed using the MLST+ target definer function of SeqSphere+ with default parameters. These

Table 2

Finished genomes[a] and assembled raw reads[b] used for *Legionella pneumophila* core genome validation (n=21)

Strain	SBT BAPS cluster	SBT ST (by Sanger sequencing)	Mean coverage	% MLST+ good targets	NCBI/EBI accession number
Thunder Bay	13	187[c]	N/A	99.34	NC_021350
HL06041035	7	734[c]	N/A	98.29	NC_018140
ATCC43290	13	187	N/A	99.67	NC_016811
LPE509	Not known	New ST (3,10,1,1,–[c],9,1)[d]	N/A	99.67	NC_020521
H053260229	1	74	72.66	97.76	ERR315647
H043940028	2	84	379.34	98.42	ERR315648
LP617	3	47	83.46	98.82	ERR164430
H064180002	3	62	73.28	96.98	ERR315651
H065000139	3	54	283.37	97.57	ERR315650
H063920004	3	47	271.57	98.82	ERR315649
H071260094[e]	5	87	485.82	98.29	ERR315653
LP423	6	1	46.26	98.75	ERR164431
EUL00013	6	5	364.53	98.75	ERR315655
H074360702	6	152[c]	343.23	98.62	ERR315656
RR08000517	7	337[c]	339.81	97.24	ERR315658
LC6774	9	154[c]	356.65	96.32	ERR315661
LC6451	10	78	74.39	97.63	ERR315664
H091960011	11	454[c]	433.78	98.62	ERR315665
H075160080	12	188	388.03	99.01	ERR315667
H034680035	13	37	84.00	97.96	ERR315668
RR08000134	14	34	435.23	99.80	ERR315670

BAPS: Bayesian analysis of population structure; EBI: European Bioinformatics Institute; N/A: not applicable; NCBI: National Center for Biotechnology Information; SBT: sequence-based typing; ST: sequence type.

[a] Finished genomes were from the NCBI database.

[b] Assembled raw reads were from EBI.

[c] Extraction of *mompS* allele from genomic data not possible due to multi-copy occurrence.

[d] Ordered in accordance with SBT scheme [21]: *flaA, pilE, asd, mip, mompS, proA, neuA*.

[e] Wrongly stated as LC6677 in Underwood et al. [24].

parameters comprised the following filters to exclude certain genes of the Philadelphia reference genome from the MLST+ scheme: a 'Minimum length filter' that discards all genes shorter than 50 bp; a 'Start codon filter' that discards all genes that contain no start codon at the beginning of the gene; a 'Stop codon filter' that discards all genes that contain no stop codon, more than one stop codon or if the stop codon is not at the end of the gene; a 'Homologous gene filter' that discards all genes with fragments that occur in multiple copies within a genome (with identity 90% and >100 bp overlap); and a 'Gene overlap filter' that discards the shorter gene from the MLST+ scheme if the affected two genes overlap>4 bp. The remaining genes were then used in a pairwise comparison using Basic Local Alignment Search Tool (BLAST) version 2.2.12 (parameters used were: 'Word size: 11', 'Mismatch penalty: -1', 'Match reward: 1', 'Gap open costs: 5', and 'Gap extension costs: 2') with the 16 query Lp chromosomes [25]. All genes of the reference genome that were common in all query genomes with a sequence identity≥90% and 100% overlap, and with the default parameter 'Stop codon percentage filter' turned on (this discards all genes that have internal stop codons in more than 20% of the query genomes) formed the final MLST+ scheme (downloadable from SeqSphere+ software).

To validate the applicability and representativeness of the Lp MLST+ target gene set, a total of 21 published high-quality genomes [24,26] – four finished genomes and 17 raw read ENA datasets that were first de novo assembled – representing 12 of the 15 BAPS SBT clusters were chosen for SeqSphere+ cgMLST analysis (Table 2) performed as below. It was assumed that a well-defined cgMLST scheme should reach at least 95% of the MLST+ genes present in each of the 21 validation genomes.

Core genome multilocus sequence typing analysis of humidifier related cases

To calibrate the cgMLST scheme for micro-evolutionary change, 15 newly generated Lp genomes (Table 3) representing three epidemiologically unrelated humidifier-associated paediatric LD clusters from Israel were analysed together with the finished Philadelphia and Paris strain genomes.

Thus SeqSphere+ extracted the defined MLST+ core genome genes from each assembly with default parameters, mainly consisting of the following settings: (i) processing options: 'Ignore contigs shorter than 200 bases'; (ii) scanning options: 'Matching scanning thresholds for creating targets from assembled genomes' with 'required identity to reference sequence of 90%' and 'required aligned to reference sequence with 100%'; (iii) BLAST options: 'Word size: 11', 'Mismatch penalty: -1', 'Match reward: 1', 'Gap open costs: 5', and 'Gap extension costs: 2'. In addition, the MLST+ scheme genes were assessed for quality, i.e. the absence of premature stop codons, ambiguous nucleotides, and support of variants to reference sequence by 75% or more read nucleotide.

A core genome gene was considered a 'good target' only if all of the above criteria were met, in which case complete sequence was analysed in comparison to the Philadelphia reference and SeqSphere+ assigned a numerical allele type. The combination of all core genome alleles in each strain formed an allelic profile per the proposed new scheme. From these allelic profiles a minimum spanning tree was calculated and drawn using SeqSphere+. In order to maintain backwards compatibility with Lp SBT, sequences of the seven genes comprising the allelic profile of the SBT schemes were separately extracted from finished genomes and WGS data and then queried against the SBT database in order to assign classic STs in silico.

Table 3

Whole genome sequencing data of *Legionella pneumophila* strains included in the study[a]

Strain	Source	Epidemiological context	SBT ST (by Sanger sequencing)	Mean coverage	Conting count	MLST* good targets %	ENA accession number
Lp-001	Clinical	Unrelated case; ST40 'outgroup' strain	40	131.51	43	99.54	ERR593560
Lp-012	Clinical	Unrelated case	23[b]	48.09	69	98.75	ERR593561
Lp-032	Environmental	Routine inspection; ST1 'outgroup' strain	1	43.61	70	98.29	ERR593562
Lp-56207	Clinical	Case 1; epidemiologically linked to strain Lp-2002694p8	1	93.20	66	98.55	ERR594281
Lp-2002694p7	Environmental	Case 1; concurrent isolate from humidifier; unrelated 'innocent bystander'	40	50.53	39	99.74	ERR593569
Lp-2002694p8	Environmental	Case 1; isolate from humidifier; last stage in transmission chain	1	74.11	57	98.62	ERR593570
Lp-119	Environmental	Case 2; isolate from humidifier; last stage in transmission chain	1	77.40	367	98.29	ERR632205, ERR632206
Lp-120	Environmental	Case 2, isolate from domestic water filtering device; middle stage in transmission chain	1	49.73	92	98.49	ERR593565
Lp-121	Environmental	Case 2; isolate from domestic water; initial stage in transmission chain	1	34.62	89	98.49	ERR593566
Lp-122	Environmental	Case 2, isolate from domestic water filtering device's filter; middle stage in transmission chain	1	109.62	409	98.22	ERR593567, ERR593568
Lp-282-1	Environmental	Case 3; isolate from domestic water; middle stage in transmission chain	1	68.83	113	98.75	ERR593571
Lp-283	Environmental	Case 3; isolate from domestic water; initial stage in transmission chain	1	42.20	77	98.22	ERR593572
Lp-284	Environmental	Case 3, isolate from domestic water filtering device's filter; middle stage in transmission chain	1	52.66	233	97.57	ERR593573
Lp-285	Environmental	Case 3, isolate from domestic water filtering device's filter; middle stage in transmission chain	1	55.89	284	98.49	ERR593574
Lp-286-1	Environmental	Case 3; isolate from humidifier; last stage in transmission chain	1	122.55	87	98.55	ERR593575

ENA: European Nucleotide Archive; MLST: multilocus sequence typing; SBT: sequence-based typing; ST: sequence type.

[a] ENA study number PRJEB7140.

[b] Extraction of *mompS* allele from whole genome sequence data not possible due to multi-copy occurrence.

Results

Setting up and validation of core genome multilocus sequence typing for Legionella pneumophila

Six finished genomes available from GenBank and 11 raw read datasets from ENA that cover all BAPS SBT clusters (Table 1) were used for cgMLST definition. ENA raw read data were de novo assembled into draft genomes. The Philadelphia strain (NC_002942.5) served as reference for core genome gene definition. The resulting cgMLST scheme consisted of 1,521 genes (ca 47.2% of the complete Philadelphia strain chromosome nucleotide; list of core genes available upon request from the authors). The SBT alleles were extracted from the genomes and generated correct ST

Table 4

Characteristics of paediatric humidifier-associated Legionnaires' disease cases included in the study

Case number / Year	Outcome	Setting	Recovered strains	Comments
1 / 2012 [10]	Fatal	Domestic free-standing cold-water humidifier serving as vehicle	ST1 from clinical (sputum) and environmental (humidifier) samples; ST40 concurrently recovered from environmental (humidifier) sample	Infection diagnosed by sputum culture; ST40 considered an innocent bystander not implicated in infection
2 / 2013	Mild	Domestic free-standing cold-water humidifier serving as vehicle; humidifier filled with water from a filtrating machine	ST1 from various environmental samples (domestic water, domestic water filtering device's filter, domestic water filtering device's water and humidifier)	Infection diagnosed by urinary antigen
3 / 2013	Severe Severe Legionnaires' disease	Domestic free-standing cold-water humidifier serving as vehicle; humidifier filled with water from a filtrating machine	ST1 (and also ST93) from various environmental samples (domestic water, domestic water filtering device's filter, domestic water filtering device's water and humidifier)	Infection diagnosed by PCR on sputum; direct sequencing on sputum confirmed ST1 infection; ST93 co-infection documented

PCR: polymerase chain reaction; ST: sequence type.

designations for nine strains. In eight strains, six of seven alleles were called correctly but the allele number for the *mompS* gene could not be determined due to presence of more than one copy of *mompS* in the genome (Table 1).

The cgMLST scheme was validated using 21 additional genomes derived from recent publications (Table 2). All 21 strains showed›96% good MLST⁺targets and resulted on average in 98.4% MLST⁺targets. Of 20 strains for which the ST designation was available, 14 were fully extracted from the WGS data, whereas in the six remaining strains, only six of seven alleles were called correctly due to multiple *mompS* gene copies (Table 2).

Investigation of humidifier-associated Legionnaires' disease

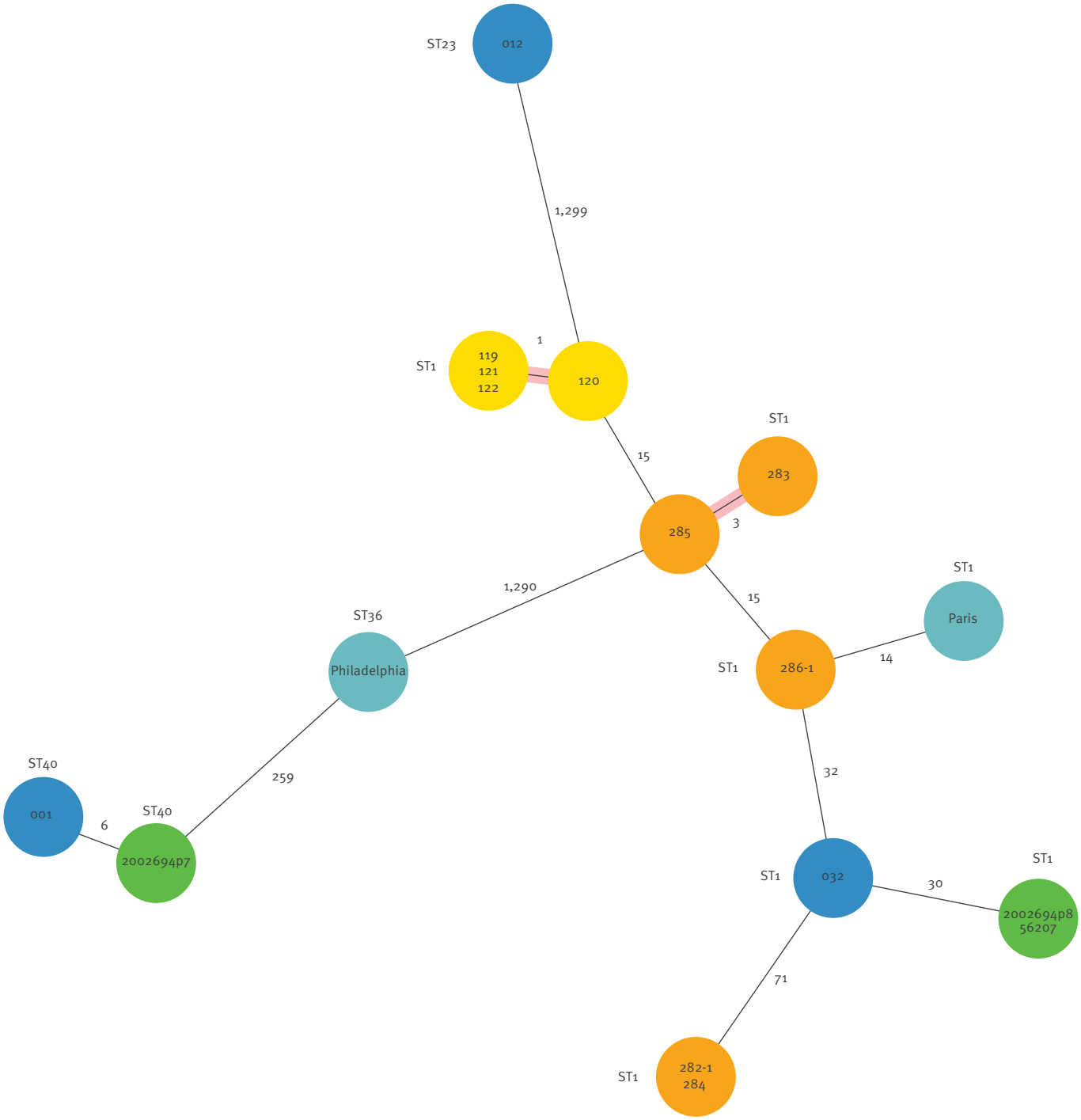
To calibrate the cgMLST scheme for micro-evolutionary change and to define a cluster type (CT) threshold, 15 newly generated Lp genomes representing three epidemiologically unrelated humidifier-associated paediatric LD clusters from Israel were analysed together with the finished Philadelphia and Paris strain genomes. Characteristics of sequenced strains are summarised in Table 3. Analysis involved 11 ST1 strains from the three incidents, a concurrent ST40 strain and three 'outgroup' strains including unrelated ST1, ST40 and ST23 strains. The median coverage was 55.9 (range: 34.6–131.5) and on average 98.6% of the MLST⁺targets could be called. The SBT ST was called complete and correct for 14 of the 15 draft genomes.

The three incidents are described in Table 4. All three cases involved children below one year of age exposed to domestic free-standing cold-water humidifiers. In case 1 the humidifier was filled with tap water whereas in cases 2 and 3 humidifiers were filled with water dispensed through domestic filtrating machines (water bars) that used charcoal filters and ultraviolet light. In case 1 Lp ST1 was detected by culture and polymerase chain reaction (PCR) of the patient's sputum and Lp ST1 was also recovered from humidifier residual water. Notably, environmental sampling revealed a ST40 strain which was not present in clinical samples. In case 2, diagnosis was made using urinary antigen testing and no sputum was available for analysis. Multiple environmental samples obtained from the water system, water filtrating machine, and humidifier were all positive for Lp ST1. In case 3, sputum was culture negative but PCR was positive for Lp. Direct SBT performed on sputum revealed a co-infection with Lp serogroup 1 ST1 and Lp serogroup 3 ST93. Environmental samples obtained from the water system, water filtrating machine and humidifier were all positive for Lp ST1 and some were also ST93 positive.

All 17 analysed genomes (including the 15 from Israel as well as the Philadelphia and Paris strain) shared in total 1,446 of the 1,521 defined core genome genes (data not shown – allelic profiles available upon request). From these allelic profiles SeqSphere⁺ calculated and drew a minimum spanning tree where the number of differing alleles is given along the branches (Figure). For case 1, identical clinical and environmental ST1 strains (Lp-2002694p8 and Lp-56207) were found (no differing alleles) and a concurrent ST40 (Lp-2002694 p7), which as expected, did not cluster with implicated ST1 strains. This ST40 strain exhibited a difference of six alleles (of the 1,521 core genome genes) from an unrelated ST40 strain serving as an 'outgroup' for ST40. Of the four environmental strains representing

Figure

Use of a minimum spanning tree generated from allelic profiles of 1,446 core genome genes shared by 17 *Legionella pneumophila* strains analysed, to investigate paediatric humidifier-associated Legionnaires' disease cases



Lp: *Legionella pneumophila*; ST: sequence type.

The Lp strain numbers are described inside the circles. Finished genomes of the 'Philadelphia' (ST36) and 'Paris' (ST1) strains obtained from the National Center for Biotechnology Information (NCBI) were used as reference (turkoise blue). Strains corresponding to the three ST1 humidifier-associated epidemiological clusters are designated in green (epidemiological cluster 1; includes also one ST40 'bystander' strain), yellow (epidemiological cluster 2) and orange (epidemiological cluster 3). Epidemiologically unrelated strains, including ST1 and ST40 'outgroup' strains (Lp-032 and Lp-001, respectively) and a ST23 (Lp-012) are designated in dark blue. The number of differing alleles is stated along the branches of the tree. Lines connecting strains within cluster type distance are highlighted by pale red background shading. The lengths of the branches reflecting distances between strains are drawn in a logarithmic scale.

the chain of transmission in case 2, three were identical ST1 strains (Lp-119, Lp-121 and Lp-122) and one had only one differing allele (Lp-120). Case 3 demonstrated more complex clustering of environmental strains into

two pairs (one identical pair formed by Lp-282-1 and Lp-284, and one pair with three differing alleles formed by Lp-285 and Lp-283) and a fifth distinct strain (Lp-286-1) which on epidemiological grounds was considered the most likely cause of infection (Lp strain recovered from humidifier residual water and therefore most likely to have been aerosolised during humidifier use). A subsequent cloning experiment performed on the patient's sputum extract, revealed sequences unique to at least two of the environmental strains (Lp-286–1 and Lp-285), suggesting co-infection (data not shown).

Discussion

WGS-analysis is emerging as the optimal molecular epidemiology tool for microbial genotyping but its application and implementation are limited by challenges in timely analysis of data and standardised integration into scalable classification schemes. While most WGS-based epidemiological investigations published to date have relied on mapping of SNPs, extension of the classical MLST approach [27] to a gene-by-gene typing scheme based on the entire core genome [16] is a promising approach for a standardised, portable and expandable typing method. The current study presents a novel core-genome allele-based typing scheme for Lp based on a standardised analysis of WGS of an internationally representative and biologically diverse Lp collection of genomes. The proposed scheme follows several recently proposed cgMLST schemes for pathogens of public health importance including *Staphylococcus aureus, Listeria monocytogenes, Escherichia coli, Neisseria meningitidis* and *Mycobacterium tuberculosis* [13-15,28,29].

Only a few clusters of LD have been investigated to date using a WGS-based approach. One study provided a retrospective analysis of a community-acquired LD cluster in which WGS yielded comparable results to that of conventional SBT [26]. Notably, WGS could not identify the most likely source of infection. The two clinical and three environmental isolates analysed in that study were not more than 15 SNPs apart [26]. Another study provided a real time investigation of a nosocomial LD cluster involving two patients [30] in which WGS had a greater resolution as compared with conventional typing and was able to link the two cases with an environmental strain and possibly to a past case. Related strains in that study were 17 SNPs apart. The reliance on SNP mapping makes those two reports difficult to reproduce and to compare, especially given the differences in reference genomes and bioinformatics pipelines used, as well as software parameter selections. Nevertheless, both papers contribute to the proof of concept of harnessing WGS for Lp investigation.

In our report, WGS of Lp strains related to three independent LD incidents was successful in demonstrating the phylogeny of implicated clinical and environmental strains. We chose to focus on Lp ST1 which is one of the most abundant ST globally and by far the most common cause of LD in Israel [31] and thus conventional tools such as SBT are not always powerful enough for epidemiological purposes. Moreover, it has recently been shown that Lp ST1 could be further characterised using additional typing methods such as spoligotyping [32]. Using the 'Paris' ST1 type strain and an 'outgroup' ST1 strain our analysis shows that the cgMLST scheme of ca 1,500 genes has an adequate discriminatory power and could resolve clustering of multiple strains in the ST1 complex. Discriminatory power in concordance with epidemiological data are among the most important performance criteria set to evaluate proposed typing methods [33]. In that respect, the observed clustering pattern of our study isolates suggests that a difference of up to four alleles between strains may serve as a preliminary threshold value for defining a WGS cluster. Nevertheless, this should be further evaluated and fine-tuned as additional genomic epidemiology data on Lp accumulates.

We included in our analysis LD cases diagnosed by three accepted laboratory modalities, being sputum culture, urinary antigen and sputum PCR in order to demonstrate the usefulness of WGS-based typing in all typical epidemiological scenarios. Of note is that case 3 was more difficult to resolve as strain clustering yielded three unrelated ST1 groups. While spontaneous mutations could provide a possible explanation, we believe that this is the result of infection with multiple ST1 strains. This reflects the inherent limitation of culture-based methods used in water testing for Lp, where picking out a single colony from similar morphotypes may overlook the presence of multiple strains. This limitation could be resolved by liberal use of SBT target screening before colony picking and in the future via metagenomic approaches.

One notable hindrance to routine application of WGS for Lp genotyping is the failure to determine the *mompS* allele number (and as a result determine the ST) for some strains, regardless of whether SNP mapping or cgMLST is being used. This phenomenon results from the presence of multiple copies of the *mompS* gene in many Lp strains, which are commonly non-identical, a fact not known when the SBT scheme was initially designed [7]. Current SBT primers used for Sanger sequencing amplify only a single copy of the gene due to sequence variation in the noncoding flanking region and thus generate consistent ST designations [7]. Therefore, in the future, tools for extraction of the correct *mompS* allele from finished genomes harbouring multiple gene copies must take synteny information, e.g. the primer sequences, into consideration to choose the correct gene copy for allele calling. Remediation of the problem for draft genomes is more difficult to achieve as the rather short second generation sequencing reads from both copies are assembled or mapped into a single contig. Notably, resolution of this limitation would be highly desirable for routine WGS application for Lp as backwards compatibility would be maintained.

Our report also highlights humidifier-associated paediatric LD as a continuously emerging risk for LD. After the first paediatric case was acknowledged and reported in Eurosurveillance [10], additional cases have been reported in Europe including Spain [34] and Cyprus [35], the latter involving a nosocomial outbreak in a nursery. The public health response to the first case in Israel was coordinated by the National Programme for Legionellosis Prevention. As part of this response, the Israeli Ministry of Health released specific guidance to professionals and members of the public. Thereafter, scheduled press releases occur every winter. As mandated, cold water humidifiers sold in the country are also labelled with a safety hazard warning through the National Standards Institute of Israel. Moreover, the Israeli Paediatrics Association has released a position paper regarding domestic humidifier use highlighting the benefits and risks. Nevertheless, paediatric humidifier-associated LD is still a public health challenge and deserves more attention.

In conclusion, we devised a WGS-based cgMLST scheme for typing of Lp, which provided high-resolution analysis of Lp strains within the same clonal complex. cgMLST appears to have satisfactory discriminatory power for LD cluster analysis and is advantageous over mapping followed by SNP calling as it is easier for standardisation and dissemination. cgMLST thus has the potential for becoming a gold standard tool for LD investigation. Humidifiers pose an ongoing risk as vehicles for LD and should be considered in cluster investigation and control efforts.

Acknowledgements

The genomics work of this study was funded by the European Community's Seventh Framework Programme (grant number FP7/2007-2013 to DH) under Grant Agreement N° 278864 in the framework of the EU PathoNGenTrace project.

Conflict of interest

D. Harmsen has declared a potential conflict of interest. He is one of the developers of the Ridom SeqSphere+ software mentioned in the manuscript, which is a development of the company Ridom GmbH (Münster, Germany) that is partially owned by him. All other authors have declared that no competing interests exist.

Author's contribution

JM-G initiated the study, interpreted data and drafted the manuscript; KP performed genome sequencing, bioinformatics analysis, and contributed to manuscript drafting; FK also conducted genome sequencing; EY, TL, LV and VA collected strains and performed traditional and molecular laboratory analyses; TGH performed microbiological analyses and participated in creation of typing scheme and drafting of manuscript; AU performed BAPS analysis and characterisation of Lp strains. CL participated in creation of typing scheme; IG contributed to interpretation of data and related public health policy; DH conceptualized the analysis of the genomic data, the creation of the Lp typing scheme, and contributed to manuscript drafting.

References

1. Fields BS, Benson RF, Besser RE. Legionella and Legionnaires' disease: 25 years of investigation. Clin Microbiol Rev. 2002;15(3):506-26. http://dx.doi.org/10.1128/CMR.15.3.506-526.2002 PMID:12097254
2. Diederen BM. Legionella spp. and Legionnaires' disease. J Infect. 2008;56(1):1-12. http://dx.doi.org/10.1016/j.jinf.2007.09.010 PMID:17980914
3. Yu VL, Plouffe JF, Pastoris MC, Stout JE, Schousboe M, Widmer A, et al. Distribution of Legionella species and serogroups isolated by culture in patients with sporadic community-acquired legionellosis: an international collaborative survey. J Infect Dis. 2002;186(1):127-8. http://dx.doi.org/10.1086/341087
4. European Centre for Disease Prevention and Control (ECDC). Annual Epidemiological Report 2012. Reporting on 2010 surveillance data and 2011 epidemic intelligence data. Stockholm: ECDC; 2013. Available from: http://ecdc.europa.eu/en/publications/Publications/Annual-Epidemiological-Report-2012.pdf
5. European Centre for Disease Prevention and Control (ECDC). Legionnaires' disease in Europe, 2011. Stockholm: ECDC; 2013. Available from: http://ecdc.europa.eu/en/publications/Publications/legionnaires-disease-in-europe-2011.pdf
6. Beauté J, Zucs P, de Jong B; European Legionnaires' Disease Surveillance Network. Legionnaires disease in Europe, 2009-2010. Euro Surveill. 2013;18(10):20417. PMID:23515061
7. Gaia V, Fry NK, Afshar B, Lück PC, Meugnier H, Etienne J, et al. Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of Legionella pneumophila. J Clin Microbiol. 2005;43(5):2047-52. http://dx.doi.org/10.1128/JCM.43.5.2047-2052.2005
8. Ratzow S, Gaia V, Helbig JH, Fry NK, Lück PC. 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. 2007;45(6):1965-8. http://dx.doi.org/10.1128/JCM.00261-07 PMID:17409215
9. Harrison TG, Afshar B, Doshi N, Fry NK, Lee JV. Distribution of Legionella pneumophila serogroups, monoclonal antibody subgroups and DNA sequence types in recent clinical and environmental isolates from England and Wales (2000-2008). Eur J Clin Microbiol Infect Dis. 2009;28(7):781-91. http://dx.doi.org/10.1007/s10096-009-0705-9 PMID:19156453
10. Moran-Gilad J, Lazarovitch T, Mentasti M, Harrison T, Weinberger M, Mordish Y, et al. Humidifier-associated paediatric Legionnaires' disease, Israel, February 2012. Euro Surveill. 2012;17(41):pii=20293.
11. Coetzee N, Duggal H, Hawker J, Ibbotson S, Harrison TG, Phin N, et al. An outbreak of Legionnaires' disease associated with a display spa pool in retail premises, Stoke-on-Trent, United Kingdom, July 2012. Euro Surveill. 2012;17(37):pii=20271.
12. Bertelli C, Greub G. Rapid bacterial genome sequencing: methods and applications in clinical microbiology. Clin Microbiol Infect. 2013;19(9):803-13. http://dx.doi.org/10.1111/1469-0691.12217 PMID:23601179
13. Leopold SR, Goering RV, Witten A, Harmsen D, Mellmann A. Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. J Clin Microbiol. 2014;52(7):2365-70. http://dx.doi.org/10.1128/JCM.00262-14 PMID:24759713
14. Schmid D, Allerberger F, Huhulescu S, Pietzka A, Amar C, Kleta S, et al. Whole genome sequencing as a tool to investigate a cluster of seven cases of listeriosis in Austria and Germany, 2011-2013. Clin Microbiol Infect. 2014;20(5):431-6. http://dx.doi.org/10.1111/1469-0691.12638
15. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, et al. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PLoS ONE. 2011;6(7):e22751. http://dx.doi.org/10.1371/journal.pone.0022751
16. Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013;11(10):728-36. http://dx.doi.org/10.1038/nrmicro3093

17. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, et al. Updating benchtop sequencing performance comparison. Nat Biotechnol. 2013;31(4):294-6. http://dx.doi.org/10.1038/nbt.2522
18. Helbig JH, Bernander S, Castellani Pastoris M, Etienne J, Gaia V, Lauwers S, et al. Pan-European study on culture-proven Legionnaires' disease: distribution of Legionella pneumophila serogroups and monoclonal subgroups. Eur J Clin Microbiol Infect Dis. 2002;21(10):710-6. http://dx.doi.org/10.1007/s10096-002-0820-3
19. Ratcliff RM, Lanser JA, Manning PA, Heuzenroeder MW. Sequence-based classification scheme for the genus Legionella targeting the mip gene. J Clin Microbiol. 1998;36(6):1560-7. PMID:9620377
20. Health Protection Agency. mip gene sequence database. London: HPA. [Accessed 14 Jul 2015]. Available from: http://webarchive.nationalarchives.gov.uk/20140714084352/http://www.hpa.org.uk/web/HPAweb&HPAwebStandard/HPAweb_C/1195733805138
21. Health Protection Agency. Legionella pneumophila Sequence based typing. [Accessed 14 Jul 2015]. London: HPA. Available from: http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php
22. Underwood AP, Bellamy W, Afshar B, Fry NK, Harrison TG. Development of an online tool for the European Working Group for Legionella Infections sequence-based typing, including automatic quality assessment and data submission. In: Cianciotto NP, Abu Kwaik Y, Edelstein PH, Fields BS, Geary DF, Harrison TG, Joseph CA, Ratcliff RM, Stout JE, Swanson MS, eds. Legionella: state of the art 30 years after its recognition. ASM Press: Washington; 2006: Chapter 44. pp. 250-2.
23. Health Protection Agency. Legionella SBT quality assessment. London:HPA. [Accessed 14 Jul 2015]. Available from: http://www.hpa-bioinformatics.org.uk/cgi-bin/legionella/sbt/seq_assemble_legionella1.cgi.
24. Underwood AP, Jones G, Mentasti M, Fry NK, Harrison TG. Comparison of the Legionella pneumophila population structure as determined by sequence-based typing and whole genome sequencing. BMC Microbiol. 2013;13(1):302. http://dx.doi.org/10.1186/1471-2180-13-302 PMID:24364868
25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10. http://dx.doi.org/10.1016/S0022-2836(05)80360-2 PMID:2231712
26. Reuter S, Harrison TG, Köser CU, Ellington MJ, Smith GP, Parkhill J, et al. A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak. BMJ Open. 2013;3(1):e002175. http://dx.doi.org/10.1136/bmjopen-2012-002175
27. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA. 1998;95(6):3140-5. http://dx.doi.org/10.1073/pnas.95.6.3140
28. Vogel U, Szczepanowski R, Claus H, Jünemann S, Prior K, Harmsen D. Ion torrent personal genome machine sequencing for genomic typing of Neisseria meningitidis for rapid determination of multiple layers of typing information. J Clin Microbiol. 2012;50(6):1889-94. http://dx.doi.org/10.1128/JCM.00038-12 PMID:22461678
29. Kohl TA, Diel R, Harmsen D, Rothgänger J, Walter KM, Merker M, et al. Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. J Clin Microbiol. 2014;52(7):2479-86. http://dx.doi.org/10.1128/JCM.00567-14
30. Graham RM, Doyle CJ, Jennison AV. Real-time investigation of a Legionella pneumophila outbreak using whole genome sequencing. Epidemiol Infect. 2014;142(11):2347-51. http://dx.doi.org/10.1017/S0950268814000375 PMID:24576553
31. Moran-Gilad J, Mentasti M, Lazarovitch T, Huberman Z, Stocki T, Sadik C, et al.; ESCMID Study Group for Legionella Infections (ESGLI). Molecular epidemiology of Legionnaires' disease in Israel. Clin Microbiol Infect. 2014;20(7):690-6. http://dx.doi.org/10.1111/1469-0691.12425
32. Ginevra C, Jacotin N, Diancourt L, Guigon G, Arquilliere R, Meugnier H, et al. Legionella pneumophila sequence type 1/Paris pulsotype subtyping by spoligotyping. J Clin Microbiol. 2012;50(3):696-701. http://dx.doi.org/10.1128/JCM.06180-11
33. van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, Fry NK, et al.; European Society of Clinical Microbiology and Infectious Diseases (ESCMID) Study Group on Epidemiological Markers (ESGEM). Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin Microbiol Infect. 2007;13(Suppl 3):1-46. http://dx.doi.org/10.1111/j.1469-0691.2007.01786.x
34. Bonilla Escobar BA, Montero Rubio JC, Martínez Juárez G. Neumonía por Legionella pneumophila asociada al uso de un humidificador doméstico en una niña inmunocompetente. Med Clin (Barc). 2014;142(2):70-2. http://dx.doi.org/10.1016/j.medcli.2013.02.042 PMID:24022027‹/jrn›
35. Yiallouros PK, Papadouri T, Karaoli C, Papamichael E, Zeniou M, Pieridou-Bagatzouni D, et al. First outbreak of nosocomial Legionella infection in term neonates caused by a cold mist ultrasonic humidifier. Clin Infect Dis. 2013;57(1):48-56. http://dx.doi.org/10.1093/cid/cit176