HHS Public Access

Author manuscript

Infect Genet Evol. Author manuscript; available in PMC 2022 April 18.

Published in final edited form as:

Infect Genet Evol. 2018 April ; 59: 172–185. doi:10.1016/j.meegid.2018.02.008.

Comparative genome analysis reveals a complex population structure of *Legionella pneumophila* subspecies

Natalia A. Kozak-Muiznieks, **Shatavia S. Morrison**, **Jeffrey W. Mercante**, **Maliha K. Ishaq**, **Taccara Johnson**, **Jason Caravas**, **Claressa E. Lucas**, **Ellen Brown**, **Brian H. Raphael**, **Jonas M. Winchell**[*]

Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA, United States

Abstract

The majority of Legionnaires' disease (LD) cases are caused by *Legionella pneumophila*, a genetically heterogeneous species composed of at least 17 serogroups. Previously, it was demonstrated that *L. pneumophila* consists of three subspecies: *pneumophila, fraseri* and *pascullei*. During an LD outbreak investigation in 2012, we detected that representatives of both subspecies *fraseri* and *pascullei* colonized the same water system and that the outbreak-causing strain was a new member of the least represented subspecies *pascullei*. We used partial sequence based typing consensus patterns to mine an international database for additional representatives of *fraseri* and *pascullei* subspecies. As a result, we identified 46 sequence types (STs) belonging to subspecies *fraseri* and two STs belonging to subspecies *pascullei*. Moreover, a recent retrospective whole genome sequencing analysis of isolates from New York State LD clusters revealed the presence of a fourth *L. pneumophila* subspecies that we have termed *raphaeli*. This subspecies consists of 15 STs. Comparative analysis was conducted using the genomes of multiple members of all four *L. pneumophila* subspecies. Whereas each subspecies forms a distinct phylogenetic clade within the *L. pneumophila* species, they share more average nucleotide identity with each other than with other *Legionella* species. Unique genes for each subspecies were identified and could be used for rapid subspecies detection. Improved taxonomic classification of *L. pneumophila* strains may help identify environmental niches and virulence attributes associated with these genetically distinct subspecies.

Keywords

Legionella pneumophila ; Legionnaires' disease; Outbreak; Subspecies

1. Introduction

Legionella pneumophila (Lp) is responsible for over 90% of cases of legionellosis in the United States and Europe (Mercante and Winchell, 2015). Legionellosis includes two

[*]Corresponding author at: Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Mailstop G03, Atlanta, GA 30033, United States. Jwinchell@cdc.gov (J.M. Winchell).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2018.02.008.

clinical presentations: Legionnaire's disease (LD) and Pontiac fever. LD is a severe form of pneumonia with potentially high fatality rates, reaching up to 46% in the healthcare-associated setting (Soda et al., 2017), while Pontiac fever is typically a self-limiting flu-like illness. The first subtyping method for Lp involved the detection of differences in the lipopolysaccharide (LPS) antigens of Lp bacteria belonging to different serogroups (Ciesielski et al., 1986; McKinney et al., 1979). However, several early studies led to an alternative concept of Lp subtyping that provides a more accurate picture of the genetic diversity within the species.

A 1982 study by Garrity et al. described three *Legionella* isolates grown from potable water samples collected in a large healthcare facility (Garrity et al., 1982). Even though these isolates were Lp serogroup 5 (Lp5), DNA-DNA hybridization studies indicated that these isolates were genetically divergent from the Lp5 type strain, suggesting that Lp isolates of the same serogroup may contain considerable genetic heterogeneity.

In 1985, Selander et al. published the first analysis of the Lp population structure (Selander et al., 1985). Electrophoretic mobilities of 22 enzymes were used to place nearly 300 Lp isolates into 62 distinctive electrophoretic types (ET). Analysis of the ETs showed that some Lp isolates were genetically very different from the rest and could be grouped into distinct “species”. Thus the authors proposed to divide the Lp population into three groups: *L. pneumophila*, species 1 and species 2. The *L. pneumophila* group was the most numerous, included 50 ETs and was represented by well characterized strains such as Philadelphia-1, OLDA and Pontiac 1. Species 1 encompassed 9 ETs and was represented by strains Dallas 1E, Los Angeles 1 and Lansing 3. Species 2 was the smallest group, consisting of only 3 closely related ETs, each represented by a single isolate. Coincidently, the three isolates that formed species 2 were the same as those described by Garrity et al. (1982) three years earlier. Selander et al. (1985)) did not observe any concordance between the ETs and serogroups and thus concluded that serotyping did not provide an accurate picture of Lp genetic structure.

Finally, in 1988 Brenner et al. published a study in which they analyzed genetic relatedness within a group of 60 Lp isolates using DNA-DNA hybridization and also characterized these isolates biochemically and serologically (Brenner et al., 1988). Based on hybridization results, the 60 Lp isolates were separated into three groups, in agreement with earlier enzyme mobility data (Selander et al., 1985). Yet, these groups were practically indistinguishable biochemically or serologically. Brenner et al. proposed to denote these groups as *L. pneumophila* subspecies (Brenner et al., 1988). *L. pneumophila* subsp. *pneumophila* corresponded to the “*L. pneumophila*” group in the Selander et al. study (Selander et al., 1985) and included Lp serogroups 1–14, with Philadelphia-1 being the type strain. *L. pneumophila* subsp. *fraseri* corresponded to the previously defined “species 1” (Selander et al., 1985) and included Lp serogroups 1, 4, 5 and Lansing 3 (now serogroup 15) with Los Angeles 1 as the type strain. Finally, *L. pneumophila* subsp. *pascullei* corresponded to “species 2” in the Selander et al. study (Selander et al., 1985), but only included three isolates, all serogroup 5, with the type strain U8W.

More recent molecular typing methods have similarly distinguished between Lp subspecies. For example, a 2002 study by Ko et al. showed that phylogenetic analysis of partial sequences of a housekeeping gene, *rpoB*, and a virulence gene, *dotA*, could separate Lp isolates belonging to subsps. *pneumophila* and *fraseri* (Ko et al., 2002). In 2008, Edwards et al. used sequences of six gene fragments based on an early form of the sequence based typing (SBT) scheme to investigate the population structure of *L. pneumophila*. Phylogenetic trees constructed using 127 unique allelic profiles indicated the distinct separation of both the *fraseri* and *pascullei* subspecies clades from the other Lp strains (Edwards et al., 2008). In a separate study using whole genome sequencing data (WGS), SNP-based phylogenetic trees of 32 Lp genomes indicated that a clade consisting of subsp. *fraseri* had an exceptionally long branch length and was clearly separated from other Lp genomes (Underwood et al., 2013).

A multi-year LD outbreak at a healthcare facility in Pennsylvania (PA) that ultimately resulted in five definite cases, 16 probable cases and five deaths was investigated in 2012 (Demirjian et al., 2015). During the CDC-assisted investigation, several Lp strains were recovered from multiple environmental sources, including faucets and showers in case rooms and an operating theater (Supplemental data; Table S1 and (Demirjian et al., 2015)). The majority of environmental isolates were of the same sequence type (ST), ST1395, which had not been previously identified. These environmental isolates had the same ST as clinical isolates obtained from three LD cases with documented exposures at the PA healthcare facility. Moreover, three environmental isolates obtained from the same healthcare facility 30 years earlier and originally described by Garrity et al. (1982) belonged to ST1335, which had a SBT profile similar to the PA12-associated ST1395 (Supplemental data; Table S1). Coincidently, the ST1335 isolates were the only representatives of the Lp subsp. *pascullei* (Brenner et al., 1988). The complete genomes of ST1395 clinical and environmental isolates from the 2012 PA outbreak as well as of one of the ST1335 environmental isolates from the 1980s were recently sequenced (Kozak-Muiznieks et al., 2016). The alignment of the complete genomes as well as analysis of 16S and average nucleotide identity (ANI) data indicated that the ST1395 strains from the 2012 PA outbreak were more closely related to subsp. *pascullei* than to subsp. *pneumophila* (Kozak-Muiznieks et al., 2016). Hence, we concluded that the ST1395 isolates were new members of subsp. *pascullei*.

The finding that subsp. *pascullei* caused a large and long term healthcare-associated outbreak encouraged us to scrutinize Lp subspecies using a whole genome sequencing approach. Our objectives were to evaluate a genomic definition of Lp subspecies, identify methods that could accurately assign Lp strains to subspecies and to understand what makes each subspecies unique.

2. Materials and methods

2.1. Legionella isolates

Thirteen “historic isolates” used in this study were obtained from the CDC *Legionella* culture collection (Table 1). The isolation and characterization of PA healthcare facility-associated isolates U8W and MICU-B is described in previous reports (Garrity et al., 1982;

Selander et al., 1985). The collection and processing of the 2012 outbreak isolates from this facility is described elsewhere (Demirjian et al., 2015) (Supplemental data, Table S1).

2.2. Sequence based typing (SBT)

The Lp isolates described in this study were typed by sequencing seven gene fragments (*flaA, pilE, asd, mip, mompS, proA, neuA*), obtaining allelic profiles and determining their corresponding STs (Gaia et al., 2005; Ratzow et al., 2007). SBT was performed according to the European Society of Clinical Microbiology and Infectious Diseases Study Group for *Legionella* Infections (ESGLI) SBT protocol for epidemiological typing of *L. pneumophila* with M13-tagged primers (Mentasti and Fry, 2012). In cases when *flaA* and *neuA* gene fragments failed to amplify, the alternative *flaA-L-N* and *flaA-R-N* (Ginevra et al., 2009) and *neuAh* primer pairs (Mentasti et al., 2014) were used instead. Novel alleles and STs were submitted to the ESGLI SBT database (http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php).

2.3. Identification of Legionella serogroups

For those *Legionella* isolates that were identified by multiplex PCR (Benitez and Winchell, 2013) as non-serogroup 1 Lp, direct fluorescent antibody staining with *L. pneumophila* serogroup-specific fluorescein isothiocyanate labeled antibody was used to determine serogroup (Cherry et al., 1978).

2.4. Genomic DNA extraction and Illumina-compatible library construction and sequencing

Legionella genomic DNA (gDNA) was extracted from pure isolate cultures using the Epicentre Masterpure DNA Purification Kit (Epicentre, Madison, WI), following the manufacturer's instructions. The Qubit Fluorometric Quantitation system (Life Technologies, Carlsbad, CA), combined with the dsDNA broad range assay kit, was used to measure DNA concentration at all steps of the extraction and NGS library preparation process.

Illumina-compatible shotgun libraries were prepared as previously described (Mercante et al., 2016), with some minor protocol modifications. Briefly, 2 μg of genomic DNA was sheared using a Covaris M220 ultrasonicator (Life Technologies, Carlsbad, CA) to a target size of 600 bp. A Zephyr Molecular Biology Workstation (Perkin Elmer, Waltham, MA) was then used for library preparation with the NEBNext Ultra II DNA Library Preparation Kit (New England Biolabs, Ipswich, MA) and NEBNext Multiplex Oligos (Dual Index Primers Set1). Pooled libraries were sequenced on an Illumina MiSeq instrument using v2 reagent chemistry and a 2 × 250 bp paired-end protocol.

2.5. Pacific Bioscience-compatible library construction and sequencing

The Pacific Biosciences (Menlo Park, CA, USA) single molecule real-time (SMRT) sequencing method was followed to prepare 10 kb or 20 kb SMRTbell template libraries with Blue Pippin size selection system (Sage Science, Beverly, MA). Briefly, genomic DNA was sheared into fragments using a Covaris g-TUBE (Woburn, MA, USA) and the SMRTbell Template Prep Kit 1.0 was used to ligate hairpin adapters onto DNA templates.

DNA/Polymerase Binding Kit P6 v2 was employed for the annealing and binding reactions and sequencing was performed using the P6C4 chemistry on the PacBio RSII sequencing instrument with a single SMRT Cell v3 over a 240 min movie period. Sequencing analysis was performed using the hierarchical genome-assembly process (HGAP) through MRT Analysis version 2.3.0.

2.6. Pacific Bioscience assembly with Illumina read mapping

The Pacific Biosciences HGAP3 assembler (Chin et al., 2013) was used to construct completely closed Lp genomic sequences (Table 1). Parameters for the HGAP3 assembler included an expected genome size of 3.4 Mb and 15× target genome coverage. A previously described (Mercante et al., 2016) approach was used to reduce the genome depth of coverage and identify nucleotide discrepancies between PacBio and Illumina sequencing data. Complete, error-corrected genomes sequenced in this study were deposited as NCBI under BioProjects PRJNA344070, PRJNA345024 and PRJNA376177. The raw Illumina data were uploaded to SRA (Table 1).

2.7. Gene prediction with Prokka

All completely closed genomic sequences were reoriented to begin with the *dnaA* gene using an in-house Python script. Prokka version 1.8 (Seemann, 2014) was used to predict protein coding sequences, ribosomal RNA genes, and transfer RNA genes. In addition, all NCBI reference genomes were processed through Prokka to minimize gene sites discrepancies for downstream comparative analyses.

2.8. Core SNP tree

kSNP version 3.0.0 (Gardner et al., 2015) was used to construct a core SNP tree with 38 genome sequences included in this study (Table 1). The kChooser script was used to identify an optimal kmer value of 31 to construct the best representation of a core-SNP phylogenetic tree. All other parameters were left at default values.

2.9. Core genome SNP pairwise comparison for the PA healthcare facility isolates

Roary version 3.6.1 (Page et al., 2015) was used to identify the core genomes between D-7119, U8W, and Dallas 1E isolates. Independent gene alignments were performed with Clustal Omega version 1.2 (Sievers et al., 2011) to prevent gene rearrangement during alignment step. SNPs were identified with in-house Perl script that detects non-consensus sites. All sites in the alignment containing gaps or “N”s were ignored.

2.10. Recombination phylogenetic analysis: whole genome alignment of Illumina sequences

An approach similar to one previously described (Chewapreecha et al., 2014) was used to construct the input data set for the recombination analysis. Isolates with Illumina data were mapped to the Lp Philadelphia-1 genome (CP013742). We used FreeBayes (Garrison and Marth, 2012) to call variants; indel positions were ignored. Also, sites with < 25× depth coverage were masked by replacing the site nucleotide with character ‘N’. The reference-based whole genome alignment was used as input into the fastGear application

(Mostowy et al., 2017) downloaded in April 2017 to identify the recent recombination events. The recent recombination genomic coordinates identified with fastGear were used as input into an in-house script to mask those additional regions in each consensus whole genome alignment sequence. All removed low-coverage (< 25×) and recombination masked consensus sequences were aligned with Mafft version 7.215 (Yamada et al., 2016) to produce a multiple sequence alignment. A maximum-likelihood phylogenetic tree was constructed with RAxML (Stamatakis, 2014) version 7.3.0-PTHREAD using the GTRGAMMA model with 1000 bootstrapping replicates.

2.11. 16S rRNA gene, mip and gyrB trees

RNAmmer 1.2 (Lagesen et al., 2007) was used to identify 16S sequences in the 38 Lp genomes (Table 1). The highest scoring 16S sequence per genome was selected for comparison. For identifying *mip* sequences, the full length of the *mip* gene *lpg0791* from Lp1 strain Philadelphia-1 was used as a query sequence for blastn searches of the Prokka annotated Lp genomes. Of note, the length of *lpg0791* is 708 base pairs (bp), whereas the length of Prokka annotated genes with the highest identity to the *lpg0791* is 702 bp. The *gyrB* gene was identified in 37 Prokka-annotated genomes as a 2418 bp long gene. In D-5265 the *gyrB* gene had a premature stop codon at position 730 and for this strain the concatenated sequence of 733 bp and 1689 bp genes was used for the alignment. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016) using the Neighbor-Joining method. The evolutionary distances were computed using the Maximum Composite Likelihood method and 500 bootstrap replications were used. All positions containing gaps and missing data were eliminated.

2.12. Average nucleotide identity (ANI)

ANI for 38 complete genomes was calculated using reciprocal best hit (two-way ANI) between two genomics datasets approach that employed the stand alone Ruby script provided by the authors (Goris et al., 2007).

2.13. Pangenome analysis for identification of shared and unique genes within the 38 Lp genomes

Roary version 3.6.1 identified a pangenome using the following options: -i 75, -s, where sequence identity was set at 75% and paralogs were not permitted to be split. All other parameters were used at the default setting. The gene_presence_absence.csv output file was used to select genes unique for each subspecies. A set of 212 Lp “test” genomes was used to verify that the identified unique genes were present in a select subspecies only. The ‘test’ genomes were assigned to specific subspecies based on the subspecies consensus SBT patterns and verified by ANI values calculated against the subspecies type strain genomes (Supplementary information; Tester_Genomes.xlsx). It was assumed that Lp genomes that did not contain other subspecies’ SBT consensus profiles and had ANI with *L. pneumophila* strain Philadelphia-1 of > 96% belonged to subsp. *pneumophila*. Each “unique” gene was tested using BLAST with the NCBI nucleotide database and a local database of 212 test genome sequences. Genes determined to be present in strains of non-corresponding subspecies were omitted.

2.14. eBURST analysis

Investigation of the phylogenetic relationship between STs was conducted using the eBURST v3 (http://eburst.mlst.net). We used the default eBURST setting for identifying clonal complexes, which are groups of related STs. According to this default definition (which is the most conservative), all members of the clonal complex share identical alleles at 6/7 loci with at least one other member of the group (Feil et al., 2004).

3. Results

3.1. Comparison of Lp strains obtained from the same PA healthcare facility in 1981 and 2012

The SBT profile of the subsp. *pascullei* ST1395 strain that was responsible for the 2012 PA healthcare-associated outbreak differed from the profile of the ST1335 Lp isolates recovered from the same healthcare facility 30 years earlier in 2 out of 7 loci (Supplemental data; Table S1). These loci were *mip* (alleles 18 and 10) and *neuA* (alleles 201 and 2). The alleles *mip* 18 and *mip* 10 shared 96% nucleotide identity (384/402 bp) and mismatches between the two alleles were located throughout the locus (Supplemental data, Fig. S1A). The alleles *neuA* 201 and *neuA* 2 shared only 62.7% nucleotide identity (222/354 bp) with mismatches evenly distributed throughout the locus (Supplemental data, Fig. S1B). In fact, *neuA* 201 belongs to a group of alternative *neuAh* alleles that have been only recently recognized, share low nucleotide identity with the original *neuA* locus, and require a special set of primers for amplification (Mentasti et al., 2014). In addition to the differences in SBT profiles, the 1981 and 2012 isolates belonged to different serogroups; the 2012 PA outbreak strain was Lp serogroup 1 (Lp1), whereas the 1981 isolates were Lp5.

During the 2012 PA outbreak investigation, two other Lp1 environmental strains belonging to sequence types ST8 and ST154 were isolated but were later determined not to be associated with cases of disease (Supplemental data; Table S1). To investigate the relatedness between the Lp strains associated with the PA healthcare facility, we sequenced the complete genomes of the following isolates: i) one clinical isolate and one environmental isolate of ST1395 (D-7119 and F-4185, respectively) (Kozak-Muiznieks et al., 2016), ii) one environmental isolate of ST154 (F-4198), and iii) U8W and MICU-B environmental isolates that belonged to ST1335 (Table 1 and (Kozak-Muiznieks et al., 2016)). In addition, we sequenced the complete genome of an Lp5 type strain Dallas 1E (England et al., 1980; Raphael et al., 2017). These genomes were compared to the following reference genomes available from NCBI: Lp1 (Corby, Lens, Lorraine and Philadelphia-1), Lp6 (Thunder Bay) and Lp12 (ATCC43290). A core SNP tree showed that the ST1395 isolates (D-7119 and F-4185) were more closely related to the ST1335 strains (U8W and MICU-B) than to any other *L. pneumophila* strain used in this comparison (Fig. 1). The results of the core genome SNP pairwise comparison showed that there were 7240 core genome SNP differences between the D-7119 (ST1395) and U8W (ST1335) strains, whereas there were 144,936 core genome SNP differences between the Dallas 1E and U8W strains that belonged to the same serogroup 5. Similar to the Dallas 1E and U8W comparison, there were 146,401 core genome SNP differences between the D-7119 and Dallas 1E strains. Interestingly, the core genome SNP tree also showed that the ST154 environmental isolate F-4198 obtained

at the same time and place as the ST1395 environmental isolate F-4198 appeared to cluster together with the Dallas 1E strain (Fig. 1). Dallas 1E was previously identified as subsp. *fraseri* (Brenner et al., 1988; Selander et al., 1985), suggesting that F-4198 could be a member of this subspecies as well.

3.2. Consensus SBT patterns for subspecies fraseri and pascullei

In order to determine if SBT allelic profiles could help classify Lp isolates according to their subspecies, we examined four Lp strains that were previously identified as subsp. *fraseri* (Table 2A). SBT profiles were already listed in the ESGLI SBT database for three of these isolates. For Detroit 1, the SBT profile was obtained for the first time in this study and a novel ST (ST2206) was assigned. A comparison of SBT profiles for all four subsp. *fraseri* strains revealed that all strains shared the same *flaA* 11, *asd* 16 and *proA* 13 alleles. Therefore, we used an allele pattern **11**-x-**16**-x-x-**13**-x (***flaA***-*pilE*-***asd***-*mip*-*mompS*-***proA***-*neuA*) to search the ESGLI SBT database for more Lp STs that shared this pattern. This search yielded 46 STs (Table 2B).

In addition to sharing the same *flaA, asd* and *proA* alleles, the majority of the 46 STs identified shared the same *pilE* and *mompS* alleles as well: *pilE* 14 (40/46 STs) and *mompS* 15 (36/46 STs). Notably, the 10 STs that did not contain *mompS* 15 had the *mompS* 7 allele, which differed by only a single nucleotide (Supplemental data, Fig. S2A). There were five STs with *pilE* 4 and one ST with *pilE* 6 (Table 2B). In contrast to the *mompS* locus, the predominant *pilE* allele, *pilE* 14, shared only 92% sequence identity with the *pilE* 4 and *pilE* 6 alleles differing from them by 25 and 26 bp, respectively (Supplemental data; Fig. S2B). However, *pilE* 4 and *pilE* 6 were 99% identical, with only 4 bp differences between them. The Lp isolates of these 46 STs belonged to serogroups 1–8, 13 and 15 (Table 2B). We also searched the ESGLI SBT database with more inclusive patterns incorporating various combinations of just two or one of the *flaA* 11, *asd* 16 or *proA* 13 alleles (Supplemental data; Table S2). The search revealed an additional 65 STs containing at least one of the alleles used for the original search. Interestingly, 47/65 of these STs contained *pilE* 14 and 43/65 STs had *mompS* 15; both of these alleles were most frequently observed in previously identified *fraseri* strains.

The subspecies *pascullei* was the least represented among the three Lp subspecies and included only two known STs, ST1335 and ST1395 (Brenner et al., 1988; Kozak-Muiznieks et al., 2016; Selander et al., 1985). A search of the ESGLI database showed that no other STs shared the allele pattern **14-18-8**-x-**28-19**-x (***flaA*-*pilE*-*asd*-*mip*-*mompS*-*proA*-*neuA***). Searches of the ESGLI SBT database with each of the alleles included in the consensus pattern individually resulted in eight additional STs. Four of these STs contained *pilE* 18, one each contained *asd* 8 and *mompS* 28, and, finally, two had *proA* 19 (Supplemental data; Table S3).

To investigate how the STs identified by the subsps. *fraseri* and *pascullei* consensus patterns relate to each other, we performed eBURST analysis on all STs available in the ESGLI database as of December 14, 2017 (containing 2500 STs of 11,842 isolates). The eBURST analysis indicated that a majority of the 46 putative *fraseri* STs belonged to the third largest clonal complex shown in Fig. 2A as clonal complex A (CC A); whereas 4/46 STs did

not relate to any other STs in the database and thus were identified as “singletons” (Table 2B). The primary founder of CC A was ST154, which was the most frequent ST in the complex and had 16 single locus variants (Fig. 2B). The CC A consisted of 82 STs, Half of which (i.e. 41/82 STs) had the complete subsp. *fraseri* consensus pattern **11**-x-**16**-x-x-**13**-x (***flaA***-*pilE*-***asd***-*mip*-*mompS*-***proA***-*neuA*) and the other half was found with partial consensus pattern searches (Table 2B and Supplemental data; Table S3). For the subsp. *pascullei* STs, neither ST1335 nor ST1395 related to any other ST in the database. Hence, even assuming that subsp. *fraseri* contains not only all STs with the complete subsp. *fraseri* consensus pattern but also those with the partial pattern included in the CC A, the proportion of subsp. *fraseri* STs in the entire Lp population may be < 4%. Subsp. *pascullei* may even be less represented (0.08% or 2 out of 2500 STs).

3.3. Identification of additional putative subspecies

In a previous study (Raphael et al., 2016) of Lp1 isolates selected from various LD investigations occurring in New York State, core SNP analysis revealed that the genomes of a clinical isolate (NY23) and an environmental isolate (NY24) from the patient's residence were highly divergent from the rest of the genomes examined, suggesting that these isolates may not belong to subsp. *pneumophila*. However, the SBT profile of these isolates (34-27-56-57-72-29-44/*flaA*-*pilE*-*asd*-*mip*-*mompS*-*proA*-*neuA*) did not share a single allele with the subsps. *fraseri* or *pascullei* SBT consensus patterns (Table 1).

In a separate study, it was shown that Lp8 and Lp17 genomes were related more closely to each other than to genomes of any other Lp serogroups (Joseph et al., 2016). The Lp8 strain (D-5744) used in that study belonged to ST2379 and had an SBT profile in complete agreement with the subsp. *fraseri* consensus pattern (Table 1). However, the SBT profile of the Lp17 strain D-4954 (21-27-28-83-15-29-x/*flaA*-*pilE*-*asd*-*mip*-*mompS*-*proA*-*neuA*) did not match the subsp. *fraseri* consensus pattern (Tables 1 and 3A).

Comparison of SBT profiles of the NY strains and D-4954 indicated that they shared *pilE* 27 and *proA* 29 alleles (Table 3A). The *flaA* alleles of the NY strains and D-4954 (34 and 21, respectively) differed by only one base pair and were 99% identical (Supplemental data; Fig. S3A). Similarly, the *asd* alleles (56 and 28) had just 3 bp difference and shared 99% identity (Supplemental data; Fig. S3B). Thus, the ESGLI SBT database was searched with the x-**27**-x-x-x-**29**-x (*flaA*-***pilE***-*asd*-*mip*-*mompS*-***proA***-*neuA*) SBT pattern to identify additional SBT profiles that shared *pilE* and *proA* alleles with the NY and D-4954 strains. The search revealed 14 STs (Table 3B). Interestingly, 12/14 STs identified in this search shared *mompS* 15, 11/14 STs shared *flaA* 21 and 11/14 STs shared *asd* 28 alleles. We also searched the ESGLI SBT database for additional STs that contained *pilE* 27 or *proA* 29 alleles. The results revealed five additional STs containing *pilE* 27, two of which also had *flaA* 21, and three of which had *mompS* 15 (Supplemental data; Table S4A). There were also 30 STs sharing *proA* 29, 28 of which had *flaA* 21, 26 had *pilE* 14, 22 had *asd* 29 and all of them shared *mompS* 15 (Supplemental data; Table S4B).

The eBURST analysis indicated that the majority of STs identified with the x-**27**-x-x-x-**29**-x (*flaA*-***pilE***-*asd*-*mip*-*mompS*-***proA***-*neuA*) pattern belonged to two small clonal complexes, consisting of five STs each. Clonal complex B (CC B) had the primary founder ST259 and

one of its STs, ST1766, contained *pilE* 23 instead *pilE* 27. Clonal complex C (CC C) had the primary founder ST1789 and all of its STs shared *pilE* 27 and *proA* 29 (Fig. 2A and Table 3B). Out of 30 additional STs that were found with the x-x-x-x-x-**29**-x search, 25 belonged to clonal complex D (CC D), with the primary founder ST819 (Supplemental data, Table S4B; Fig. 2A and Supplemental data, Fig. S4).

Given these data, we predicted that Lp strains sharing *pilE* 27 and *proA* 29 alleles may represent a new Lp subspecies. The putative subspecies was designated *raphaeli* in recognition of the original description of the founding members (NY23 and NY24) of this subspecies which were reported by Raphael et al., 2016). We selected NY23 (D-7705) as the representative of this putative new subspecies in further analyses.

In 2016, we identified an Lp4 clinical isolate, D-7708, from a possible healthcare-associated LD case in Georgia (USA). This isolate had a novel *mip* allele (*mip* 77), and thus a novel SBT profile, which was assigned ST2186 (Table 1). Because of the novelty of the SBT profile, the genome of D-7708 was sequenced and compared with other genomes in this study. The isolate appeared to be more closely related to genomes of subsps. *fraseri, raphaeli* and putative subsp. *pascullei* rather than to subsp. *pneumophila* (data not shown). However, the D-7708 SBT profile did not match either subsps. *fraseri, raphaeli*, or putative subsp. *pascullei* SBT consensus patterns, with the exception that it had the same *proA* 13 allele as subsp. *fraseri*. Comparison of D-7708 *flaA* 30 sequence with *flaA* 11 (associated with subsp. *fraseri*) indicated that they shared 99% sequence identity, with only one base pair difference (Supplemental data; Fig. S5A). Similarly, comparison of the D-7708 *asd* 44 sequence with *asd* 16 (associated with subsp. *fraseri*) showed that these alleles also shared 99% sequence identity, with seven base pair differences (Supplemental data; Fig. S5B). The ESGLI SBT database contained seven STs that shared the *flaA* 30 allele (Supplemental data; Table S5). All but one, ST2160, were also identified during searches with partial SBT patterns of subsp. *fraseri* (Supplemental data; Table S2) since 6/7 of them shared *proA* 13 associated with subsp. *fraseri*. The analysis of sequences of D-7708 SBT loci suggested that D-7708 was either an atypical member of subsp. *fraseri* or a representative of a novel subspecies.

Because members of subspecies *fraseri, pascullei* and putative subsp. *raphaeli* appeared to have very distinct alleles of some SBT loci like *flaA* or *proA*, we were interested to see whether alleles of any SBT loci group into subspecies-specific clusters. For each of the seven SBT loci, the nucleotide sequences of their alleles extracted from the ESGLI SBT database were aligned and the alignments were analyzed in RAxML. The resulting trees for *asd, flaA, mompS, pilE* and *proA* showed that the alleles identified in the majority of SBT profiles of subsps. *fraseri*, D-7708, subsp. *pascullei* and putative subsp. *raphaeli* were grouped in distinct clades separate from the alleles common to subsp. *pneumophila* (Supplemental data; Fig. S6A-E). Interestingly, these clades contained additional alleles that could, potentially, belong to several new subspecies distinct from subsp. *pneumophila*. No separate clades that would include or exclude any of the subspecies were identified in the trees for *mip, neuA* or *neuAhH* alleles (Supplemental data; Fig. S6F-H).

3.4. Genomic characterization of Lp subspecies

Complete genomes of Lp isolates, representing four Lp subspecies and D-7708, together with the Lp subsp. *pneumophila* reference sequences available from NCBI, were used for comparative analysis of Lp subspecies (Table 1). Core SNP analysis using kSNP (Fig. 3A) showed that the strains identified as subsp. *pneumophila* clustered separately from the strains identified as subsps. *fraseri, pascullei* or putative subsp. *raphaeli*. Additionally, the subsp. *pascullei* strains formed a distinct cluster, whereas subsps. *fraseri* and putative subs. *raphaeli* were more closely related to each other than to the other two subspecies. The D-7708 strain appeared to be closer to the clade of subsp. *fraseri*.

To further characterize the phylogenetic relationships among the subspecies, fastGear was used to identify recent recombination events (i.e. recombinations that affected only a subset of strains in a subspecies (Mostowy et al., 2017)). As a result, a total of 2307 recent recombination events were detected with an average size of 2868.3 base pairs (Supplementary information; fastGear_Analyses.xlsx; Recent_Recombination_Events worksheet). The average number of events per genome for each subspecies was 123.8, 55.5, 10.5, and 86 for subsps. *pneumophila, fraseri, pascullei*, and putative subs. *raphaeli*, respectively. Fig. 3B depicts the phylogenetic relationships of isolates representing Lp subspecies with putative recombination sites masked (which represents ~6% of genome on average; Supplementary information; fastGear_Analyses.xlsx; Percentage_of_Masked_Genome worksheet). All four subspecies, including the putative subsp. *raphaeli*, represented distinct lineages. The strain D-7708, that may represent another subspecies, clustered more closely with subsp. *fraseri* but had the largest number of recombination events identified in the analysis (n = 396). Similar to the topology of the tree generated with kSNP, analysis of a core SNP tree with the putative recombination sites masked (Fig. 3B) indicated that the most common ancestor of these lineages diverged into a lineage containing subsp. *pneumophila* and a separate lineage associated with all of the remaining Lp subspecies. Furthermore, subsp. *fraseri* and putative subsp. *raphaeli* appear to have diverged from a common ancestor most recently. In addition, fastGear analysis results showed that ancestral recombination events (i.e. recombinations that were present in all strains of a subspecies (Mostowy et al., 2017)) took place among all four subspecies with the recombinations subsp. *fraseri* - putative subsp. *raphaeli* and subsps. *fraseri* - *pascullei* being the most frequent (Supplementary information; fastGear Analysies.xlsx; Ancestral_Recombination_Events worksheet).

A 16S tree (Fig. 4A) showed that the 16S sequences of all strains belonging to subsps. *fraseri, pascullei* and putative subsp. *raphaeli* as well as D-7708 were identical, and were clearly separated from the 16S sequences of subsp. *pneumophila* strains. The 16S sequences of subsp. *pneumophila* contained some differences. In contrast, the relationship of *mip* sequences did not reflect the subspecies grouping (Fig. 4B). Here, the *mip* sequences of such strains as Philadelphia-1 (subsp. *pneumophila*) and D-6026 (subsp. *fraseri*) appeared to be identical. Interestingly, Lp5 strains Dallas 1E (subsp. *fraseri*), U8W and MICU-B (both subsp. *pascullei*) were indistinguishable. The *mip* sequences of each subspecies type strain was used to calculate the percent identity in pairwise comparison with *mip* sequences of other strains. Within subsp. *pneumophila, mip* did not differ by > 1.14% (Supplemental

data; Table S6). In contrast, a *mip* sequence comparison among other subspecies showed maximum differences of 3.99%, 4.42% and 3.85% for subsps. *fraseri, pascullei* and putative subsp. *raphaeli*, respectively (Supplemental data; Table S6). On the other hand, a *gyrB* based tree showed a clear separation of all four subspecies (Fig. 4C). The subsp. *fraseri* type strain, Los Angeles 1, appeared to be more distant from other subsp. *fraseri* strains and more closely related to D-7708. A pairwise comparison of *gyrB* sequences of the subspecies type strains with other strains was conducted. With the exception of subsp. *pneumophila*, the *gyrB* sequences of strains belonging to the same subspecies appeared to be highly similar (Supplemental data; Table S7). In subsp. *pneumophila* the *gyrB* gene appeared to be more diverse, with the maximum percent difference reaching 2.15%.

The average nucleotide identity (ANI) method is often used to estimate genetic relationships between strains and assign them to the same or different species (Goris et al., 2007). ANI was calculated for a set of 38 *L. pneumophila* strains representing different subspecies (Tables 1 and 4A). The intra-subspecies ANI values were high, with the range 97.54–99.88%. In contrast, the inter-subspecies ANI values for subsp. *pneumophila* and subsp. *pascullei* were markedly below 95%, which is considered to be the threshold for assigning members to the same species. The inter-subspecies ANI values for subsp. *fraseri*, putative subsp. *raphaeli* and D-7708 were above 95%. ANI between Lp strains and four other *Legionella* species for which complete genomes were available from NCBI were markedly below 90%, with the highest value for the pairwise comparison of subsp. *pneumophila* and *L. hackeliae* (84.63%) and the lowest value between D-7708 and *L. fallonii* (77.14%) (Table 4B).

The results of genomic characterization of Lp subspecies supported our hypothesis that Lp strains sharing *pilE* 27 and *proA* 29 alleles represent a distinct subspecies *raphaeli*.

3.5. Identifying unique genetic features of Lp subspecies

A total of 5416 genes were predicted in all 38 genomes listed in Table 1, 43% of which (2326/5416) were shared by all strains and encoded such essential bacterial structures as ribosomal subunits, DNA polymerase and sigma factors (Supplementary information; Gene_Presense_Absense.xlsx). Unique genes were identified for each subspecies as well as a distinct combination of genes specific for subsp. *fraseri* and D-7780. Not surprisingly, the number of unique genes appeared to be inversely related to the number of strains used in each set. For example, 71 unique genes were identified among the small number of subsp. *pascullei* examined and 58 unique genes were identified in the genome of D-7708. However, only 9 unique genes were found among the large set of subsp. *pneumophila* strains (Supplemental data; Table S8 and Gene_Presense_Absense.xlsx).

We used 212 Lp genomes representing all four subspecies to verify that these unique genes were present in all Lp isolates of a select subspecies, but absent in all isolates of other subspecies (see Materials and methods). A final subset of five unique genes for each subspecies were identified (Table 5 and Supplementary information; Subspecies_Unique_Genes.txt). Almost half of these genes (9/20) encodes for proteins of unknown function. However, there are several potentially interesting genes such as subsp.

pneumophila-specific *ddrA* gene that encodes for a Dot/Icm effector, or subsp. *raphaeli*-specific genes encoding for signaling proteins (Table 5).

4. Discussion

As traditional bacterial characterization methods are being replaced by WGS, we now have the tools to ask whether previously determined phylogenetic relationships among bacterial strains remain stable. For Lp, WGS showed that, while some of the earlier typing methods, such as serogrouping, may not represent the true phylogenetic relatedness among lineages, other established methods continue to be helpful and may be used for a quick taxonomic placement of Lp strains. In this study, we employed the WGS data first to identify a new Lp subspecies (subsp. *raphaeli*) and then to characterize the relationships between four Lp subspecies.

WGS data supported the hypothesis made in the 1980s, based on enzyme electrophoretic mobility and DNA-DNA hybridization studies, that the Lp species was not homogeneous, but consisted of at least three groups. Upon analysis of 38 complete Lp genomes, we observed four distinct groups, subsps. *pneumophila, fraseri, raphaeli* and *pascullei* (Fig. 3). Strain D-7708 appeared to be closely related to subsp. *fraseri*, but it may represent a separate group. Each subspecies included serogroup 1 strains, but other serogroups were also distributed among the subspecies.

The observed phylogenetic relationships among the 38 Lp isolates suggested that subsp. *pneumophila* have diverged from other subspecies into a separate lineage. Based on the international SBT database, the subsp. *pneumophila* appears to be the most numerous as well as heterogeneous group, perhaps due to its successful adaptation to the manmade environment and/or occupation of numerous ecological niches. The most recent common ancestor of the non-subsp *pneumophila* lineages have diverged into one lineage containing subsp. *pascullei* and another lineage containing subsps. *fraseri, raphaeli* and-D-7708.

Subspecies *pascullei* is almost exclusively represented by strains isolated from the same healthcare facility in PA. The only subsp. *pascullei* strain found outside of the PA was isolated from a cooling tower in Switzerland (personal communication, Valeria Gaia). This Lp5 strain was assigned to subsp. *pascullei* based on its ST (ST1335) and ANI value of 99.93% between this strain and subsp. *pascullei* type strain U8W. Interestingly, during the PA12 outbreak investigation, none of the ST1335 strains were found at the PA location. We hypothesize that recombination occurred at the LPS biosynthesis locus, together with limited vertically acquired mutations resulted into the emergence of a 2012 Lp1 ST1395 outbreak strain from the Lp5 ST1335 strain (Kozak-Muiznieks et al., 2016). Hopefully, more subsp. *pascullei* members will be discovered in the future as more Lp genome sequences become publically available. It is possible that subsp. *pascullei* is adapting to a unique ecological niche that restricts this subspecies representation in the environment.

Subspecies *fraseri* and *raphaeli* are more related to each other and to D-7708 than to either subsps. *pascullei* or *pneumophila*. Based on the SBT database, both subsps. *fraseri* and *raphaeli* appeared to be quite numerous, relatively heterogeneous, represented by several

serogroups and found throughout the world. In these terms, both subspecies are more similar to subsp. *pneumophila* rather than subsp. *pascullei*. Yet, the total number of subsps. *fraseri* and *raphaeli* STs currently deposited in the international database represents < 5% of the Lp population. Perhaps the subsp. *pneumophila* is more evolutionary successful and thus more numerous compared to other Lp subspecies. Alternatively, the underrepresentation of these subspecies in the database could be due to them occupying different ecological niches besides the built environment or having geographical distribution that does not coincide with the location of the heaviest contributors to the database.

The phylogenetic relationship of D-7708 to other subspecies is not clear. The core phylogenetic trees with either unmasked or masked recombination sites (Fig. 3) indicated that D-7708 was more closely related to subsp. *fraseri* than to any other subspecies, yet it appeared distinct from the other subsp. *fraseri* strains. The very large number of recombination events (396) identified in D-7708 could suggest that this is a subsp. *fraseri* strain representing a population that is on the way to becoming a distinct subspecies.

After verifying with currently available computational methods and WGS data that the 38 Lp strains were consistently grouped into the same distinct subspecies, we evaluated the ability of several established typing methods to separate Lp strains into subspecies. As previously shown, serogroup determination was not particularly useful in detecting subspecies. However, some sequence-based methods were more informative. The 16S rRNA sequences of 22 strains that did not belong to subsp. *pneumophila* were identical to each other and markedly different from more heterogeneous 16S sequences of subsp. *pneumophila*. This difference supports the hypothesis of the early split of subsp. *pneumophila* from the rest of the Lp population. This observation also highlights the potential for incorrect species determination based on the 16S sequence. For example, if a PCR for Lp detection relies on a 16S region that is not conserved among all Lp subspecies, a portion of Lp population would be misidentified as non-Lp *Legionella*.

Whereas the *mip* sequence based method works well for species determination, it was not useful for discriminating subspecies. The *mip* typing scheme is a gold standard method for identifying *Legionella* species since the *mip* gene is present in all legionellae with the exception of *L. geestiana* (Ratcliff et al., 1998). The sequences of the *mip* gene demonstrate sufficient heterogeneity to be different in each species, yet able to be amplified by a single pair of primers. However, *mip* sequences of 38 Lp strains were not separated along their assigned subspecies groups. The percent identity of full length *mip* sequences also showed that, within the same subspecies, the *mip* could be as much as 4.42% different. Hence, *mip* sequencing is not reliable for identification of Lp subspecies.

A recent publication demonstrated that the *gyrB* gene that encodes the subunit B protein of DNA gyrase could successfully distinguish several *Legionella* species as well as Lp subsps. *pneumophila, fraseri* and *pascullei* (Xi et al., 2017). The *gyrB* gene is over 2 kb in length, universally distributed, has higher base substitution frequency than 16S rDNA and thus has been used as a phylogenetic marker for many bacteria (Kakinuma et al., 2003; Yamamoto and Harayama, 1995). A phylogenetic tree based on full length *gyrB* of the 38 strains representing Lp subspecies separated all four subspecies into distinct clades in

perfect agreement with WGS data. These results support our designation of a new subsp. *raphaeli* and highlight the *gyrB* gene as a helpful phylogenetic marker for *Legionella* at both the species and subspecies levels.

We also observed that the 7-gene based SBT scheme was quite helpful for subspecies typing. The SBT consensus patterns derived from comparison of a few subspecies representatives were supported by the whole genome sequencing data from 38 strains. Therefore, the following patterns could be used: **11**-x-**16**-x-x-**13**-x (***flaA***-*pilE*-***asd***-*mip*-*mompS*-***proA***-*neuA*) for subsp. *fraseri*, **14-18-8**-x-**28-19**-x (***flaA***-***pilE***-***asd***-*mip*-***mompS***-***proA***-*neuA*) for subsp. *pascullei* and x-**27**-x-x-x-**29**-x (*flaA*-***pilE***-*asd*-*mip*-*mompS*-***proA***-*neuA*) for subsp. *raphaeli*. It is assumed that Lp strains with SBT profiles that do not match the above consensus patterns belong to subsp. *pneumophila*, but there may be additional, undiscovered subspecies (see below). On the other hand, the results of the eBURST analysis showed that for both subsps. *fraseri* and *raphaeli* there may be additional Lp strains with SBT profiles that do not exactly match the consensus pattern, yet are closely related to the “classical” subsps. *fraseri* and *raphaeli* (Supplemental data; Table S2 and S4). The availability of WGS data for strains with partial consensus SBT patterns is needed to test this prediction. The finding that in the phylogenetic trees of *flaA, pilE, asd, mompS* and *proA* alleles of subsp. *fraseri*, D-7708, subsp. *raphaeli* and subsp. *pascullei* appeared to group separately, supports the use of SBT patterns for determining these subspecies.

Comparison of ANI values with DNA-DNA hybridization (DDH) data for several Gram-positive and Gram-negative bacteria has led to the conclusion that species delineation with a cut-off point of 70% DDH corresponds to 95% ANI (Goris et al., 2007). However, for *Legionella* the ANI species demarcation appears to be much lower. Whereas within each Lp subspecies the ANI values were above 95%, the ANI calculated for Lp strains from different subspecies was as low as 90.3% (data not shown). In fact, the low ANI values obtained in the inter-subspecies comparisons as well as the 16S sequencing data that distinctly separated subsp. *pneumophila* from other subspecies could be suggestive of considering these groups as independent species. In their recent work on the role of homologous recombination (HR) in Lp evolution, David et al. observed that subsp. *fraseri* has rarely been an HR donor to subsp. *pneumophila* (David et al., 2017). The authors concluded that the lack of genomic exchange between these subspecies may indicate divergence to the point of species differentiation (David et al., 2017). However, fastGear analysis of our Lp strain dataset, representing different subspecies, indicated that HR has occurred among all subspecies, including subsp. *fraseri*. Due to the high heterogeneity of the Lp population, selection of particular strains for analysis can bias the results and explain discrepancies in the detection of HR events by us and others. In addition, whereas inter-subspecies ANI values were below the 95% cut-off, the ANI values between Lp and other *Legionella* spp. were much lower (≤81%). This indicates a high degree of genomic diversity within the *Legionella* genus compared to other genera (Goris et al., 2007). Therefore, *Legionella* species (Lp in particular) are separated by larger differences in genetic content than species in other genera. It is important to keep in mind such a high degree of diversity within Lp species when designing pan-Lp detection or typing methods that rely on conserved sequence regions. There may be far less of such regions when all Lp subspecies are included in the analyses.

We propose that the ANI species cut-off should be lowered to 90% for Lp and to 96% for the Lp subspecies. To assign an Lp strain with available WGS data to a subspecies, the ANI should be calculated between this strain and the type strains of each of the four subspecies. It would be expected that the highest ANI value > 96% would place this strain in the correct subspecies. In the case where ANI for all subspecies is < 96% yet > 90%, we would suggest that the strain belongs to a novel Lp subspecies.

It is tempting to speculate that genetic differences between subspecies might also translate into distinguishable phenotypes. Qin et al. observed that subsp. *fraseri* did not grow as well as subsp. *pneumophila* in J774 cells and hence it was hypothesized that subsp. *fraseri* was less invasive and less virulent to humans (Qin et al., 2016). The same authors also found 19 genes, including the *katA* gene that encodes for catalase-peroxidase, to contain nonsense mutations in subsp. *fraseri* but not in subsp. *pneumophila*. The authors suggested that the *katA* mutation could decrease subsp. *fraseri* virulence. In contrast, we found that the *katG1* and *katG2* genes that encode for catalase-peroxidase were present in all 38 genomes, including all 11 of subsp. *fraseri* (Supplementary information; Gene_Presense_Absense.xlsx) without any mutations leading to the premature termination of the encoded protein. The observation that subsps. *fraseri, pascullei* and *raphaeli* strains were responsible for multiple LD outbreaks in the past (Table 1) and that some of the subsps. *fraseri* and *raphaeli* strains (ST154 and ST259, respectively) are among the most frequent sporadic STs in the US (Kozak-Muiznieks et al., 2014), instead indicate that these subspecies are not less pathogenic than subsp. *pneumophila*.

Our search for a unique gene set for each subspecies revealed a number of genes with potentially interesting functions, such as biocide resistance and pathogenicity. A number of subspecies-unique genes identified based on the set of 38 strains appeared not to be unique to a single subspecies or not common in all subspecies strains after the use of the 212 Lp strain test set, and thus had to be eliminated. Nevertheless, this testing process helped select a panel of five unique genes for each subspecies. This panel can be used for the development of rapid testing and classification methods.

One of the most interesting genes that ended up in the final panel is a subsp. *pneumophila* unique gene that encodes for the Dot/Icm effector DrrA, also called SidM. It is a multifunctional type IV secretion effector that helps *Legionella* to recruit GTPase Rab1 on the *Legionella*-containing vacuole (LCV) (So et al., 2015). The study describing the discovery of DrrA/SidM indicated that *drrA* mutants were severely impaired in Rab1 recruitment to the LCV, which assumingly affects pathogenicity of these mutants (Murata et al., 2006). This would suggest that the three other subspecies are less virulent compared to subsp. *pneumophila*, but, as previously mentioned, is contradicted by their association with many LD outbreaks and sporadic cases. Interestingly the DrrA/SidM antagonist SidD that deAMPylates Rab1 is also absent in all non-subsp. *pneumophila* strains, and neither is it found in 8/16 subsp. *pneumophila* strains (Supplementary information; Subspecies_Unique_Genes.txt). Because of the high level of redundancy among the Dot/Icm effectors (O'Connor et al., 2011) there are likely other effector(s) in these strains that duplicate the DrrA/SidM and SidD functions.

In this study, we identified new members of Lp subspecies *fraseri* and *pascullei* as well as described a novel subspecies *raphaeli*. We demonstrated that the population structure of Lp is highly complex containing multiple subspecies that appear to be capable of recombination among each other. Moreover, we identified several additional SBT alleles that appear to cluster with those detected in subsp. *fraseri, raphaeli* and *pascullei*, and that were distinct from those associated with subsp. *pneumophila*, suggesting that additional Lp subspecies may be discovered in the future. Identification of these distinct taxonomic groups within the Lp species may enable the development of more detailed and robust strain databases to examine trends such as geographic distribution, ecological niches, and virulence factors as well as antibiotic and biocide resistance that may be specific for each subspecies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The findings and conclusions in this presentation are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. This study was supported in part by funds made available through the CDC Office of Advance Molecular Detection.

We thank the curators of and contributors to the ESGLI SBT database, which provided invaluable data necessary for studying the Lp population. In addition, we thank Valeria Gaia for sharing ST1335 environmental isolate from a cooling tower.

References

Benitez AJ, Winchell JM, 2013. Clinical application of a multiplex real-time PCR assay for simultaneous detection of *Legionella* species, *Legionella pneumophila*, and *Legionella pneumophila* serogroup 1. J. Clin. Microbiol 51, 348–351. [PubMed: 23135949]

Borges V, Nunes A, Sampaio DA, Vieira L, Machado J, Simoes MJ, Goncalves P, Gomes JP, 2016. *Legionella pneumophila* strain associated with the first evidence of person-to-person transmission of Legionnaires' disease: a unique mosaic genetic backbone. Sci. Rep 6, 26261. [PubMed: 27196677]

Brenner DJ, Steigerwalt AG, Epple P, Bibb WF, McKinney RM, Starnes RW, Colville JM, Selander RK, Edelstein PH, Moss CW, 1988. *Legionella pneumophila* serogroup Lansing 3 isolated from a patient with fatal pneumonia, and descriptions of *L. pneumophila* subsp. *pneumophila* subsp. nov., *L. pneumophila* subsp. *fraseri* subsp. nov., and *L. pneumophila* subsp. *pascullei* subsp. nov. J. Clin. Microbiol 26, 1695–1703. [PubMed: 3053773]

Cherry WB, Pittman B, Harris PP, Hebert GA, Thomason BM, Thacker L, Weaver RE, 1978. Detection of Legionnaires disease bacteria by direct immunofluorescent staining. J. Clin. Microbiol 8, 329–338. [PubMed: 359594]

Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, Pessia A, Aanensen DM, Mather AE, Page AJ, Salter SJ, Harris D, Nosten F, Goldblatt D, Corander J, Parkhill J, Turner P, Bentley SD, 2014. Dense genomic sampling identifies highways of pneumococcal recombination. Nat. Genet 46, 305–309. [PubMed: 24509479]

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J, 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods 10, 563–569. [PubMed: 23644548]

Ciesielski CA, Blaser MJ, Wang WL, 1986. Serogroup specificity of *Legionella pneumophila* is related to lipopolysaccharide characteristics. Infect. Immun 51, 397–404. [PubMed: 2417953]

David S, Sanchez-Buso L, Harris SR, Marttinen P, Rusniok C, Buchrieser C, Harrison TG, Parkhill J, 2017. Dynamics and impact of homologous recombination on the evolution of *Legionella pneumophila*. PLoS Genet. 13, e1006855. [PubMed: 28650958]

Demirjian A, Lucas CE, Garrison LE, Kozak-Muiznieks NA, States S, Brown EW, Wortham JM, Beaudoin A, Casey ML, Marriott C, Ludwig AM, Sonel AF, Muder RR, Hicks LA, 2015. The importance of clinical surveillance in detecting Legionnaires' disease outbreaks: a large outbreak in a hospital with a legionella disinfection system-Pennsylvania, 2011–2012. Clin. Infect. Dis 60, 1596–1602. [PubMed: 25722201]

Edwards MT, Fry NK, Harrison TG, 2008. Clonal population structure of *Legionella pneumophila* inferred from allelic profiling. Microbiology 154, 852–864. [PubMed: 18310031]

England AC 3rd, McKinney RM, Skaliy P, Gorman GW, 1980. A fifth serogroup of *Legionella pneumophila*. Ann. Intern. Med 93, 58–59. [PubMed: 7396318]

Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG, 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol 186, 1518–1530. [PubMed: 14973027]

Gaia V, Fry NK, Afshar B, Luck PC, Meugnier H, Etienne J, Peduzzi R, Harrison TG, 2005. Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila*. J. Clin. Microbiol 43, 2047–2052. [PubMed: 15872220]

Gardner SN, Slezak T, Hall BG, 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics 31, 2877–2878. [PubMed: 25913206]

Garrison E, Marth G, 2012. Haplotype-based Variant Detection From Short-read Sequencing. Cornell University Library.

Garrity GM, Elder EM, Davis B, Vickers RM, Brown A, 1982. Serological and genotypic diversity among serogroup 5-reacting environmental *Legionella* isolates. J. Clin. Microbiol 15, 646–653. [PubMed: 6175657]

Ginevra C, Lopez M, Forey F, Reyrolle M, Meugnier H, Vandenesch F, Etienne J, Jarraud S, Molmeret M, 2009. Evaluation of a nested-PCR-derived sequence-based typing method applied directly to respiratory samples from patients with Legionnaires' disease. J. Clin. Microbiol 47, 981–987. [PubMed: 19225096]

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM, 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol 57, 81–91. [PubMed: 17220447]

Helbig JH, Benson RF, Pelaz C, Jacobs E, Luck PC, 2007. Identification and serotyping of atypical *Legionella pneumophila* strains isolated from human and environmental sources. J. Appl. Microbiol 102, 100–105. [PubMed: 17184324]

Joseph SJ, Cox D, Wolff B, Morrison SS, Kozak-Muiznieks NA, Frace M, Didelot X, Castillo-Ramirez S, Winchell J, Read TD, Dean D, 2016. Dynamics of genome change among *Legionella* species. Sci. Rep 6, 33442. [PubMed: 27633769]

Kakinuma K, Fukushima M, Kawaguchi R, 2003. Detection and identification of *Escherichia coli, Shigella*, and *Salmonella* by microarrays using the gyrB gene. Biotechnol. Bioeng 83, 721–728. [PubMed: 12889036]

Ko KS, Lee HK, Park MY, Park MS, Lee KH, Woo SY, Yun YJ, Kook YH, 2002. Population genetic structure of *Legionella pneumophila* inferred from RNA polymerase gene (rpoB) and DotA gene (dotA) sequences. J. Bacteriol 184, 2123–2130. [PubMed: 11914343]

Kozak-Muiznieks NA, Lucas CE, Brown E, Pondo T, Taylor TH Jr., Frace M, Miskowski D, Winchell JM, 2014. Prevalence of sequence types among clinical and environmental isolates of *Legionella pneumophila* serogroup 1 in the United States from 1982 to 2012. J. Clin. Microbiol 52, 201–211. [PubMed: 24197883]

Kozak-Muiznieks NA, Morrison SS, Sammons S, Rowe LA, Sheth M, Frace M, Lucas CE, Loparev VN, Raphael BH, Winchell JM, 2016. Three genome sequences of *Legionella pneumophila* subsp. *pascullei* associated with colonization of a health care facility. Genome Announc 4.

Kumar S, Stecher G, Tamura K, 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol 33, 1870–1874. [PubMed: 27004904]

Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW, 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35, 3100–3108. [PubMed: 17452365]

McKinney RM, Thomason BM, Harris PP, Thacker L, Lewallen KR, Wilkinson HW, Hebert GA, Moss CW, 1979. Recognition of a new serogroup of Legionnaires disease bacterium. J. Clin. Microbiol 9, 103–107. [PubMed: 372210]

Mentasti M, Fry NK, 2012. Sequence-Based Typing Protocol for Epidemiological Typing of *Legionella pneumophila*, Version 5.0. ESCMID Study Group for Legionella Infections, European Society of Clinical Microbiology and Infectious Diseases, Basel, Switzerland.

Mentasti M, Underwood A, Luck C, Kozak-Muiznieks NA, Harrison TG, Fry NK, 2014. Extension of the *Legionella pneumophila* sequence-based typing scheme to include strains carrying a variant of the N-acylneuraminate cytidylyltransferase gene. Clin. Microbiol. Infect 20, O435–441. [PubMed: 24245827]

Mercante JW, Winchell JM, 2015. Current and emerging *Legionella* diagnostics for laboratory and outbreak investigations. Clin. Microbiol. Rev 28, 95–133. [PubMed: 25567224]

Mercante JW, Morrison SS, Desai HP, Raphael BH, Winchell JM, 2016. Genomic analysis reveals novel diversity among the 1976 Philadelphia Legionnaires' disease outbreak isolates and additional ST36 strains. PLoS One 11, e0164074. [PubMed: 27684472]

Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P, 2017. Efficient inference of recent and ancestral recombination within bacterial populations. Mol. Biol. Evol 34, 1167–1182. [PubMed: 28199698]

Murata T, Delprato A, Ingmundson A, Toomre DK, Lambright DG, Roy CR, 2006. The *Legionella pneumophila* effector protein DrrA is a Rab1 guanine nucleotide-exchange factor. Nat. Cell Biol 8, 971–977. [PubMed: 16906144]

O'Connor TJ, Adepoju Y, Boyd D, Isberg RR, 2011. Minimization of the *Legionella pneumophila* genome reveals chromosomal regions involved in host range expansion. Proc. Natl. Acad. Sci. U. S. A 108, 14733–14740. [PubMed: 21873199]

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J, 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693. [PubMed: 26198102]

Qin T, Zhang W, Liu W, Zhou H, Ren H, Shao Z, Lan R, Xu J, 2016. Population structure and minimum core genome typing of *Legionella pneumophila*. Sci. Rep 6, 21356. [PubMed: 26888563]

Raphael BH, Baker DJ, Nazarian E, Lapierre P, Bopp D, Kozak-Muiznieks NA, Morrison SS, Lucas CE, Mercante JW, Musser KA, Winchell JM, 2016. Genomic resolution of outbreak-associated *Legionella pneumophila* serogroup 1 isolates from New York State. Appl. Environ. Microbiol 82, 3582–3590. [PubMed: 27060122]

Raphael BH, Kozak-Muiznieks NA, Morrison SS, Mercante JW, Winchell JM, 2017. Complete genome sequences of *Legionella pneumophila* subsp. *fraseri* strains Detroit-1 and Dallas 1E. Genome Announc 5.

Ratcliff RM, Lanser JA, Manning PA, Heuzenroeder MW, 1998. Sequence-based classification scheme for the genus *Legionella* targeting the mip gene. J. Clin. Microbiol 36, 1560–1567. [PubMed: 9620377]

Ratzow S, Gaia V, Helbig JH, Fry NK, Luck PC, 2007. 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, 1965–1968. [PubMed: 17409215]

Seemann T, 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069. [PubMed: 24642063]

Selander RK, McKinney RM, Whittam TS, Bibb WF, Brenner DJ, Nolte FS, Pattison PE, 1985. Genetic structure of populations of *Legionella pneumophila*. J. Bacteriol 163, 1021–1037. [PubMed: 4030689]

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG, 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol 7, 539. [PubMed: 21988835]

So EC, Mattheis C, Tate EW, Frankel G, Schroeder GN, 2015. Creating a customized intracellular niche: subversion of host cell signaling by *Legionella* type IV secretion system effectors. Can. J. Microbiol 61, 617–635. [PubMed: 26059316]

Soda EA, Barskey AE, Shah PP, Schrag S, Whitney CG, Arduino MJ, Reddy SC, Kunz JM, Hunter CM, Raphael BH, Cooley LA, 2017. Vital signs: health care-associated Legionnaires' disease surveillance data from 20 states and a large metropolitan area - United States, 2015. MMWR Morb. Mortal. Wkly Rep 66, 584–589. [PubMed: 28594788]

Stamatakis A, 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. [PubMed: 24451623]

Underwood AP, Jones G, Mentasti M, Fry NK, Harrison TG, 2013. Comparison of the *Legionella pneumophila* population structure as determined by sequence-based typing and whole genome sequencing. BMC Microbiol. 13, 302. [PubMed: 24364868]

Xi D, Dou Y, Ren W, Yang S, Feng L, Cao B, Wang L, 2017. A *gyrB* oligonucleotide microarray for the specific detection of pathogenic *Legionella* and three *Legionella pneumophila* subsp. Antonie Van Leeuwenhoek 110 (12), 1515–1525. [PubMed: 28695408]

Yamada KD, Tomii K, Katoh K, 2016. Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. Bioinformatics 32, 3246–3251. [PubMed: 27378296]

Yamamoto S, Harayama S, 1995. PCR amplification and direct sequencing of *gyrB* genes with universal primers and their application to the detection and taxonomic analysis of *Pseudomonas putida* strains. Appl. Environ. Microbiol 61, 3768. [PubMed: 16535156]

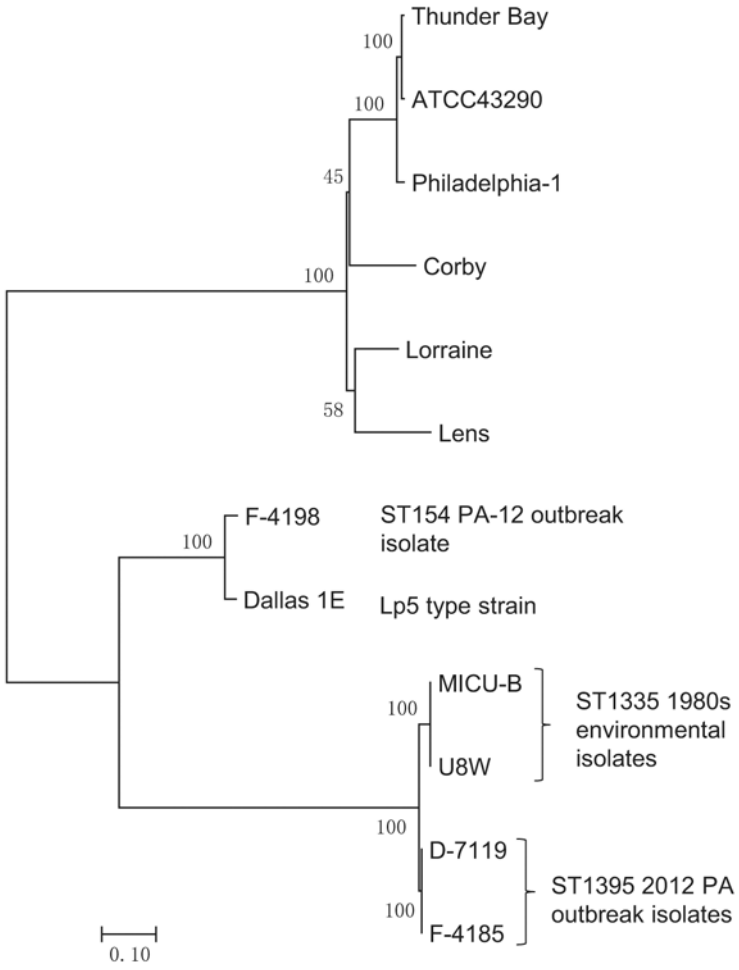

Fig. 1.

PA12 outbreak-associated ST1395 Lp1 strain related more closely to the ST1335 Lp5 strains than to any other Lp reference strain. Complete genomes of Lp isolates obtained during the PA12 outbreak investigation were compared to the genomes of 1981 environmental isolates associated with the same healthcare facility. The genomes of Lp5 type strain, Dallas 1E, as well as genomes of Lp1 strains Corby, Lens, Lorraine and Philadelphia 1, Lp6 Thunder Bay and Lp12 ATCC43290 were used for reference. A total of 26,632 core SNPs were identified using kSNP version 3.

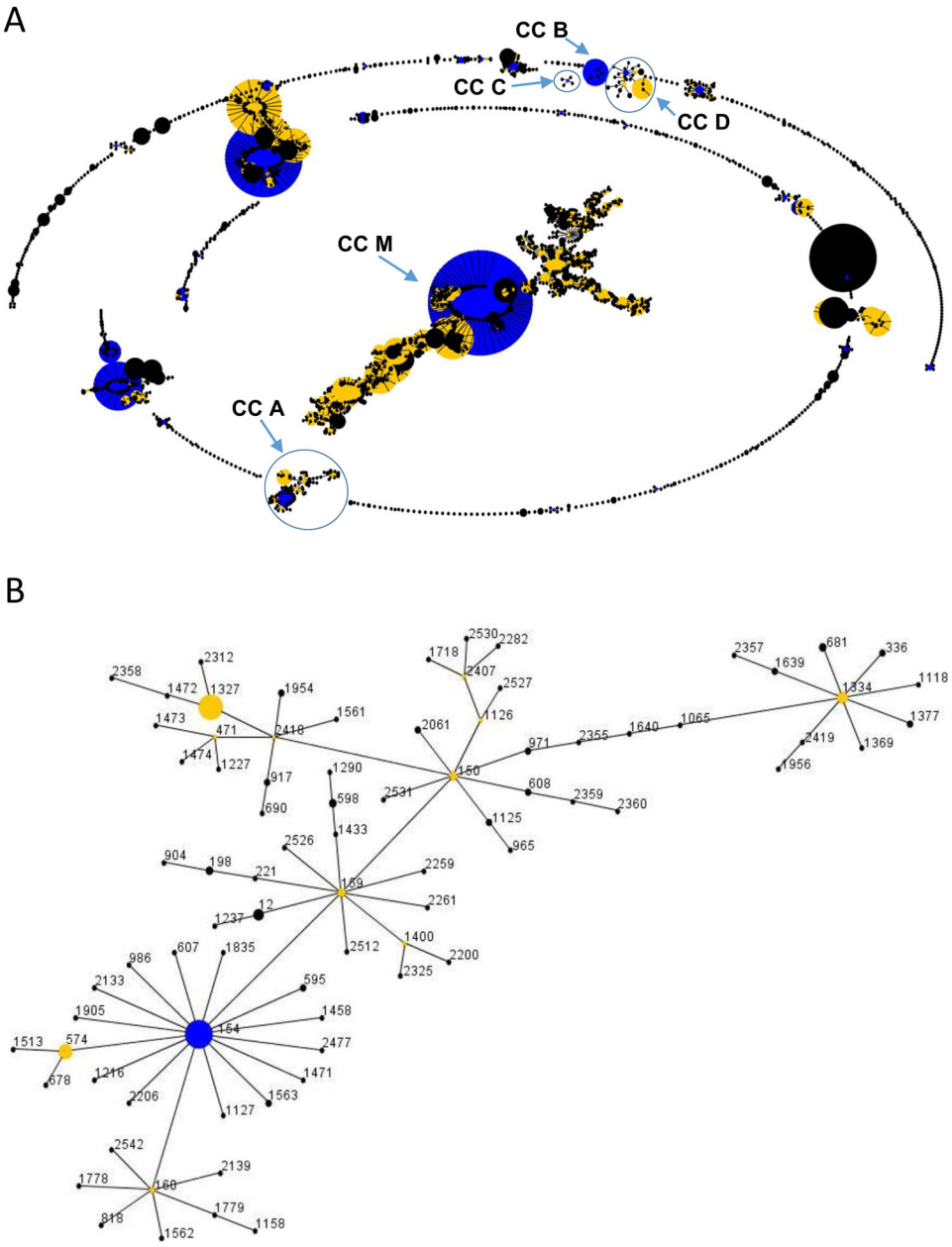

Fig. 2.

Phylogenetic relationship among STs predicted to belong to different Lp subspecies done by eBURST analysis. A. The population snapshot of 2500 STs listed in the ESGLI database contained 91 clonal complexes and 439 singletons. The largest clonal complex with the predicted founder ST1 is designated with letter M for “main”. The clonal complexes formed by STs that are predicted to belong to subspecies *fraseri* (clonal complex A or CC A) and *raphaeli* (CC B, CC C and CC D) are marked by the corresponding letters. B. Detailed view of a clonal complex A is formed by STs predicted to belong to subspecies *fraseri*. ST154 is the clonal complex primary founder and is identified as blue circle. The yellow circles represent subgroup founders (STs that have at least two descendant single-locus variants). The area of each circle represents the prevalence of the ST in the input data.

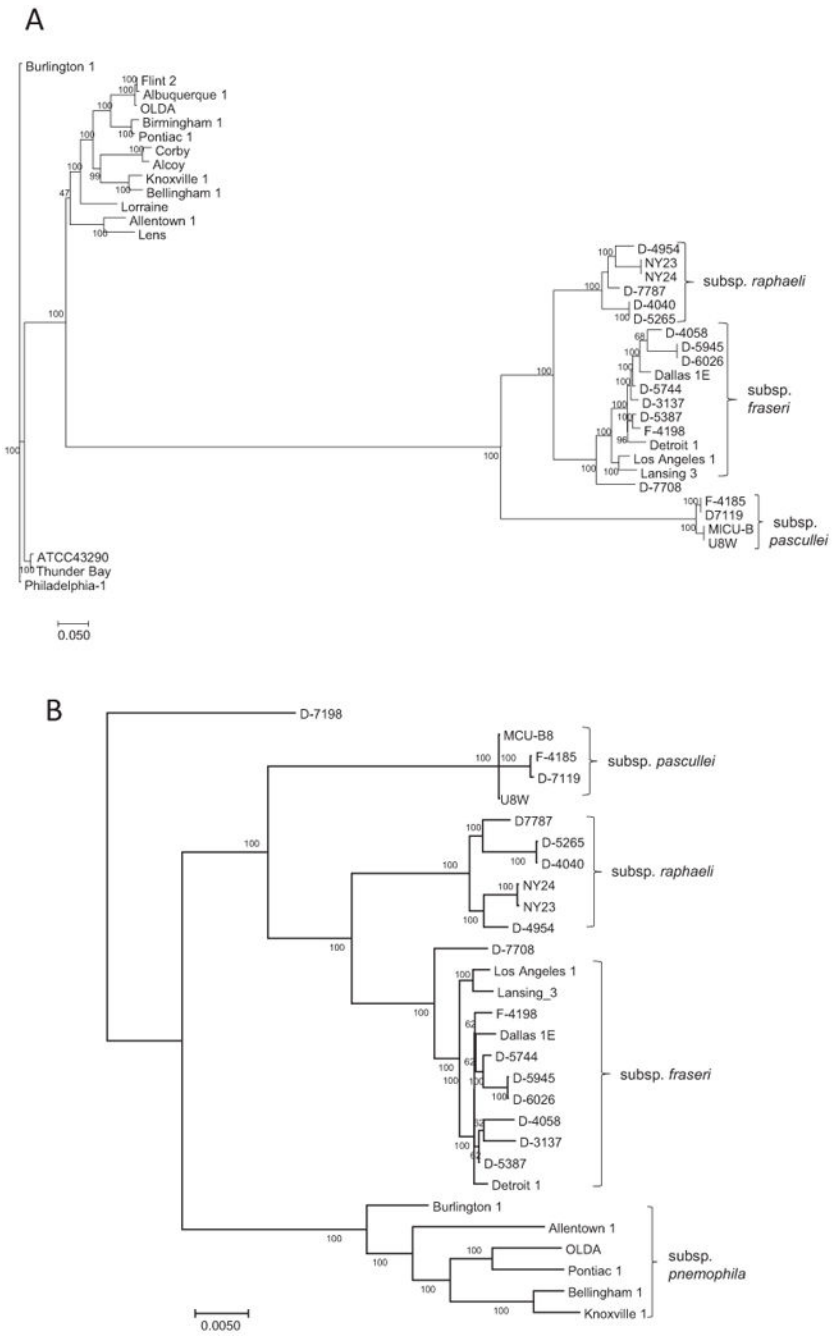

Fig. 3.

Lp strains predicted to belong to different subspecies based on the SBT profiles formed distinct groups according to the predictions. A. A core SNP tree of 38 genome sequences was built using kSNP version 3.0 application. B. A maximum-likelihood phylogenetic tree of 28 Lp strains with masked recombination sites. The Lp1 strain D-7198 that according to the SBT profile and ANI data did not belong to any of the four recognized Lp subspecies was used as an outlier.

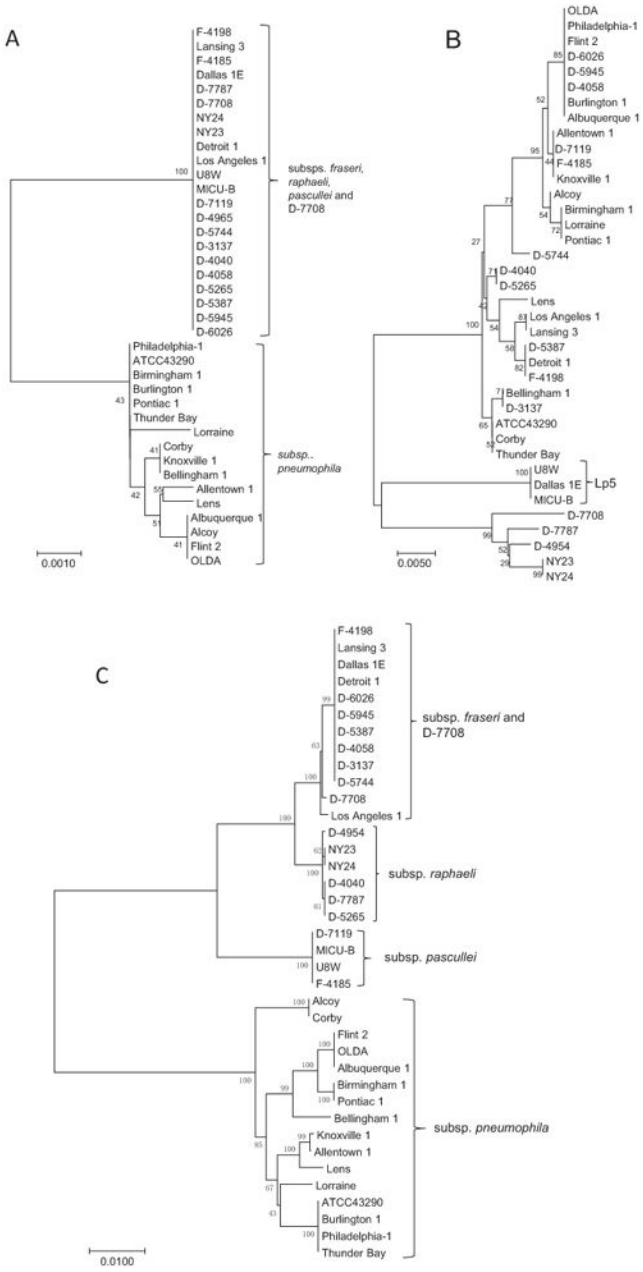

Fig. 4.

Phylogenetic trees based on 16S, *imp* and *gyrB* sequences of Lp subspecies demonstrate different abilities of these sequence based typing methods to correctly separate Lp strains into subspecies. 16S (A), *mip* (B) and *gyrB* (C) sequences were used to infer phylogenetic relationships between 38 Lp strains representing Lp subspecies employing the Neighbor-Joining method. The percentage of replicate trees in which Lp strains were clustered together in the bootstrap test (500 replicates) are shown next to the branches. The trees are drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed

using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site.

Infect Genet Evol. Author manuscript; available in PMC 2022 April 18.

Table 1

Legionella pneumophila strains which genomes were used for comparative analysis in this study.

Strain	Serogroup	SBT[a] profile	ST[b]	Accession number	SRA[c]	Comments
		flaA-pilE-asd-mip-mompS-proA-neuA				
Subspecies *pneumophila*						
Alcoy	1	6-10-15-13-9-14-6	578	CP001828	N/A	NCBI reference strain
ATCC 43290	12	3-10-1-28-14-9-3	187	CP003192	N/A	NCBI reference strain
Corby	1	6-10-15-28-9-14-6	51	CP000675	N/A	NCBI reference strain
Lens	1	12-9-26-5-26-17-15	15	CR628337 and CR628339[d]	N/A	NCBI reference strain
Lorraine	1	5-10-22-15-6-2-6	47	FQ958210 and FQ958212	N/A	NCBI reference strain
Philadelphia-1	1	3-4-1-1-14-9-1	36	CP013742	N/A	Type strain for subsp. *pneumophila*, NCBI reference strain
Thunder Bay	6	3-10-1-28-14-9-3	187	CP003730	N/A	NCBI reference strain
Albuquerque 1 (D-7474)	1	1-4-3-1-1-1-1	1	CP021286 and CP021287	N/A	CDC historical collection
Allentown 1 (D-7475)	1	4-8-11-10-10-12-2	44	CP021283, CP021284 and CP021285	SRR5831903	CDC historical collection
Bellingham 1 (D-7473)	1	7-6-17-3-24-11-11	61	CP021269	SRR5831904	CDC historical collection
Birmingham 1 (D-7470)	1	8-3-3-15-21-1-6	294	CP021268	N/A	CDC historical collection
Burlington 1 (D-7481)	1	3-4-1-1-14-9-1	36	CP021267	SRR5831902	CDC historical collection
Flint 2 (D-7477)	1	1-4-3-1-1-1-1	1	CP021281 and CP021282	N/A	CDC historical collection
Knoxville 1 (D-7468)	1	7-10-17-10-5-4-13	58	CP021266	SRR5831901	CDC historical collection
OLDA (D-7466)	1	1-4-3-1-1-1-1	1	CP016030 and CP016031	SRR3648078	CDC historical collection
Pontiac 1 (D-7467)	1	8-10-3-15-18-1-6	62	CP016029	SRR3655343	CDC historical collection
Subspecies *fraseri*						
Los Angeles 1 (D-7696)	4	11-14-16-25-7-13-206	1334	CP021265	SRR5831908	Type strain for subsp. *fraseri*, CDC historical collection
Lansing 3	15	11-14-16-25-7-13-24	336	CP021257	SRR5831916	CDC historical collection
F-4198	1	11-14-16-16-15-13-2	154	CP021279 and CP021280	SRR5831907	2012 PA outbreak environmental isolate
Detroit 1 (D-7698)	1	11-6-16-16-15-13-2	2206	CP017457	SRR5832166	CDC historical collection
Dallas 1E	5	11-14-16-18-15-13-201	1300	CP017458	SRR5832165	CDC historical collection
D-6026	1	11-14-16-1-15-13-207	1400	CP017601	N/A	CDC collection, sporadic isolate
D-5945	1	11-14-16-1-7-13-207	2200	CP017602	N/A	CDC collection, sporadic isolate

Strain	Serogroup	SBT[a] profile	ST[b]	Accession number	SRA[c]	Comments
		flaA-pilE-asd-mip-mompS-proA-neuA				
D-5744	8	11-14-16-30-15-13-213	2379	CP021258	SRR5831913	CDC collection, 2008 AZ outbreak
D-5387	1	11-14-16-16-15-13-2	154	CP021264	SRR5831906	CDC collection, 1998 St. Croix, Virgin Islands outbreak clinical isolate
D-4058	1	11-14-16-1-15-13-1	150	CP021277 and CP021278	SRR5831905	CDC collection, 1994 CT outbreak clinical isolate
D-3137	1	11-14-16-3-15-13-9	818	CP021263	SRR5831910	CDC collection, 1991 CA outbreak clinical isolate
Possible new subspecies						
D-7708	4	30-18-44-77-61-13-217	2186	CP021259	SRR5831914	CDC collection, 2016 GA nosocomial cluster isolate
Subspecies *pascullei*						
U8W (D-7160)	5	14-18-8-18-28-19-201	1335	CP021262	SRR5831909	Type strain for subsp. *pascullei*, ATCC purchase, ATCC 33737
MICU-B (D-7158)	5	14-18-8-18-28-19-201	1335	CP014256	SRR3134833	ATCC purchase, ATCC 33735
D-7119	1	14-18-8-10-28-19-2	1395	CP014257	SRR3134832	2012 PA outbreak clinical isolate
F-4185	1	14-18-8-10-28-19-2	1395	CP014255	SRR3134835	2012 PA outbreak environmental isolate
Subspecies *raphaeli*						
NY23 (D-7705)	1	34-27-56-57-72-29-44	1204	CP021261	SRR5831912	Type strain for subsp. *raphaeli*, NYS study
NY24 (D-7706)	1	34-27-56-57-72-29-44	1204	CP021260	SRR5831911	NYS study
D-4954	17[e]	21-27-28-83-15-29-x	N/A	CP021256	SRR5831915	CDC collection, sporadic isolate
D-4040	1	3-27-28-2-15-29-6	884	CP021274, CP021275 and CP021276	SRR5831918	CDC collection, 1994 DE outbreak
D-5265	1	21-27-28-2-15-19-6	259	CP021272 and CP021273	SRR5831917	CDC collection, 2002 PA outbreak
D-7787	5	21-27-29-80-15-29-230	2258	CP021270 and CP021271	SRR5831919	CDC collection; sporadic isolate

The complete genome sequences for all strains listed below the Thunder Bay strains were generated at CDC.

[a] SBT stands for sequence based typing.

[b] ST stand for sequence type.

[c] SRA stand for sequence read archive.

[d] When more than one accession number is listed for genome sequence, the first accession number is for the chromosomal DNA and the following numbers are for the plasmids.

[e] This serogroup designation is used by the CDC *Legionella* laboratory after the Helbig et al. (2007) verification that D-4954 does not belong to serogroups 1–15.

Table 2A

Sequence based typing (SBT) profiles of subspecies *fraseri* strains. Historic subsp. *fraseri* strains.

Historic strain	SBT profile *flaA-pilE-asd-mip-mompS-proA-neuA*	ST[a]	Serogroup
Los Angeles 1	**11**[b]-14-**16**-25-7-**13**-206	1334	4
Dallas 1E	**11**-14-**16**-18-15-**13**-201	1300	5
Detroit 1	**11**-6-**16**-16-15-**13**-2	2206	1
Lansing 3	**11**-14-**16**-25-7-**13**-24	336	15

[a] ST stands for sequence type.

[b] Bold font indicates alleles that are identical among all four strains.

Table 2B

Sequence based typing (SBT) profiles of subspecies *fraseri* strains. SBT profiles identified with the **11**-x-**16**-x-x-**13**-x consensus pattern.

SBT profile: *flaA-pilE-asd-mip-mompS-proA-neuA*	ST[a]	Serogroup	Comments
11[b]-14-**16**-1-15-**13**-6	12	1	CC A[c]
11-14-**16**-1-15-**13**-1	150	1	CC A, 1994 CT outbreak
11-14-**16**-16-15-**13**-2	154	1	CC A founder, 2003 St Croix outbreak
11-14-**16**-1-15-**13**-2	159	1	CC A
11-14-**16**-16-15-**13**-9	160	1 and NK[d]	CC A
11-14-**16**-1-15-**13**-9	221	1	CCA
11-14-**16**-25-7[e]-**13**-24	336	3 and 15	CC A, Lansing 3
11-14-**16**-31-15-**13**-3	471	5	CC A
11-14-**16**-16-15-**13**-11	574	1	CC A
11-14-**16**-10-15-**13**-11	598	1	CC A
11-14-**16**-16-15-**13**-1	607	1	CC A
11-14-**16**-25-7-**13**-6	681	4	CC A
11-14-**16**-3-15-**13**-9	818	1	CC A, 1991 CA outbreak
11-4-**16**-5-15-15-**13**-2	823	1	Singleton
11-14-**16**-31-15-**13**-6	917	5 and 13	CC A
11-14-**16**-25-7-**13**-1	1065	4	CC A
11-14-**16**-25-7-**13**-13	1118	1	CC A
11-4-**16**-1-15-**13**-1	1125	1	CC A
11-14-**16**-12-15-**13**-1	1126	7	CC A
11-14-**16**-25-15-**13**-2	1127	NK	CC A
11-14-**16**-19-15-**13**-3	1227	7	CC A
11-14-**16**-10-15-**13**-6	1237	1	CC A
11-14-**16**-18-15-**13**-201	1300	5	ST1300 and ST2365[f], Dallas 1E
11-14-**16**-31-15-**13**-210	1327	5, 13, 2-14[g] and NK	CC A
11-14-**16**-25-7-**13**-206	1334	4	CC A, Los Angeles 1
11-14-**16**-1-15-**13**-207	1400	1	CC A
11-14-**16**-1-15-**13**-11	1433	1	CC A
11-14-**16**-28-15-**13**-3	1473	6	CC A
11-14-**16**-7-15-**13**-3	1474	2	CC A
11-14-**16**-19-15-**13**-215	1718	7	CC A
11-14-**16**-65-7-**13**-217	1719	4	Singleton
11-14-**16**-10-15-**13**-2	1905	1	CC A, person-to-person transmission[h]
11-14-**16**-31-15-**13**-207	1954	5	CC A
11-4-**16**-25-15-**13**-206	1956	4	CC A
11-14-**16**-71-15-**13**-1	2061	1	CC A

SBT profile: *flaA-pilE-asd-mip-mompS-proA-neuA*	ST[a]	Serogroup	Comments
11-14-**16**-76-15-**13**-2	2133	1	CC A
11-14-**16**-10-15-**13**-9	2139	1	CC A
11-14-**16**-1-7-**13**-207	2200	1	CC A
11-6-**16**-16-15-**13**-2	2206	1	CC A, Detroit 1
11-14-**16**-12-7-**13**-3	2329	1	Singleton
11-14-**16**-30-15-**13**-213	2379	8	Singleton, 2008 AZ outbreak
11-14-**16**-12-15-**13**-215	2407	NK	CC A
11-14-**16**-31-15-**13**-1	2418	2–14	CC A
11-4-**16**-25-7-**13**-206	2419	4	CC A
11-14-**16**-1-7-**13**-2	2512	1	CC A
11-4-**16**-16-15-**13**-9	2542	1	CC A

[a] ST stands for sequence type.

[b] Bold font indicates alleles that are identical among all STs.

[c] "CC A" indicates that the ST belongs to a clonal complex A (Fig. 2A).

[d] NK means that the serogroup is not known.

[e] The underlined numbers indicate alleles that deviate from the predominant *fraseri* alleles.

[f] The ST belongs to a clonal complex consisting of two STs (ST1300 and ST2365).

[g] "2–14" indicates that the serogroup is one of the 2–14 serogroups of Lp.

[h] Described in Borges et al. (2016).

Table 3A

Sequence based typing (SBT) profiles of a putative subspecies *raphaeli*. SBT profiles of the unusual Lp strains identified in two collaboration studies.

Strains	SBT profile *flaA-pilE-asd-mip-mompS-proA-neuA*	ST[a]	Serogroup
NY23 (D-7705)	34-**27**[b]-56-57-72-**29**-44	1204	1
NY24 (D-7706)	34-**27**-56-57-72-**29**-44	1204	1
D-4954	21-**27**-28-83-15-**29**-DEL[c]	N/A[d]	17

[a] ST stands for sequence type.

[b] Bold font indicates alleles that are identical among all three strains.

[c] "DEL" indicates that the *neuA* gene contains partial deletion and hence the *neuA* allele could not be determined.

[d] N/A – the ST could not be assigned due to the lack of the *neuA* allele number.

Table 3B

Sequence based typing (SBT) profiles of a putative subspecies *raphaeli*. SBT profiles identified with the x-**27**-x-x-x-**29**-x consensus pattern.

SBT profile *flaA-pilE-asd-mip-mompS-proA-neuA*	ST[a]	Serogroup	Comments
21-**27**[b]-28-2-15-**29**-6	259	1	CC B[c], founder of CC B
3-**27**-28-2-15-**29**-6	884	1	CC B
21-**27**-28-12-15-**29**-6	1023	1	CC B
21-**27**-28-5-15-**29**-15	1173	1	Singleton
34-**27**-56-57-72-**29**-44	1204	1	Singleton, NY23 and NY24 isolates
21-**27**-28-13-15-**29**-6	1402	1	CC B
21-**27**-28-54-15-**29**-206	1541	2–14	CC C[d]
21-**27**-28-54-15-**29**-9	1789	1	CC C, founder of CC C
21-**27**-28-28-15-**29**-9	1845	1	CC C
21-**27**-28-28-15-**29**-9	2131	1	ST2131 and ST1096[e]
21-**27**-29-80-15-**29**-230	2258	5	Singleton
21-**27**-28-21-15-**29**-9	2302	1	CC C
21-**27**-28-82-15-**29**-9	2374	1	CC C
34-**27**-56-15-72-**29**-6	2417	1	Singleton

[a] ST stands for sequence type.

[b] Bold font indicates alleles that are identical among all STs.

[c] The ST belongs to a clonal complex B (Fig. 2A).

[d] The ST belongs to a clonal complex C (Fig. 2A).

[e] The ST belongs to a clonal complex consisting of two STs (ST2131 and ST1096).

Table 4A

Average nucleotide identity (ANI) values (in %). Average values calculated for ANIs for each *L. pneumophila* subspecies represented by 38 Lp strains.

Subspecies	***pneumophila***	***fraseri***	***raphaeli***	***pascullei***
pneumophila	97.54	91.72	91.18	90.55
fraseri	91.72	99.16	96.4	93.49
D-7708	91.48	97.91	96.31	93.87
raphaeli	91.18	96.4	99.02	93.84
pascullei	90.55	93.49	93.84	99.88

Table 4B

Average nucleotide identity (ANI) values (in %). Average values calculated for ANI for 38 Lp strains and four other *Legionella* species which complete genomes were available on NCBI.

	L. falloni	***L. hackeliae***	***L. longbeachae***	***L. oakridgensis***
	LN614827.1[a] LN614828.1 LN614829.1	**LN681225.1 LN681226.1**	**FN650140.1 FN650141.1**	**CP004006.1 CP004007.1**
pneumophila	78.82	84.63	78.71	81.00
fraseri	80.21	79.93	78.11	81.17
D-7708	77.14	78.75	77.56	78.1
raphaeli	77.16	80.42	79.24	81.46
pascullei	77.29	79.66	78.5	78.6

[a]Accession numbers for the assembly of each complete genome sequence. The second and third accession numbers for each species indicate plasmid sequences.

Table 5

Genes that are unique for *L. pneumophila* subspecies.

Subspecies	Protein name	Length (bp)	*pneumophila*	*fraseri*	*raphaeli*	*pascullei*	Domains
pneumophila	Hypothetical protein group_1109	291	Yes	No	No	No	Two transmembrane domains at 47–66 and 70–89 aa[a] out of 96 aa
	Hypothetical protein group_1136	192	Yes	No	No	No	No confidently predicted domains or other features
	Defect in RAb1 recruitment protein A DrrA	969	Yes	No	No	No	DrrA_P4M domain at C-terminus
	Hypothetical protein group_1891	1170	Yes	No	No	No	Two low complexity regions about 10 aa each
	Hypothetical protein group_2416	1488	Yes	No	No	No	A transmembrane domain at C-terminus
fraseri	ATP-dependent DNA helicase PcrA	1812	No	Yes	No	No	UvrD-helicase superfamily
	Hypothetical protein group_3297	750	No	Yes	No	No	A coiled coil region at the C-terminus
	Hypothetical protein group_3302	1353	No	Yes	No	No	No confidently predicted domains or other features
	Putative ATPase	1842	No	Yes	No	No	No confidently predicted domains or other features
	Hypothetical protein group_4243	255	No	Yes	No	No	No confidently predicted domains or other features
raphaeli	Cyclic di-GMP phosphodiesterase Gmr	915	No	No	Yes	No	Contains PAS and GGDEF domains
	Hypothetical protein group_2608	942	No	No	Yes	No	No confidently predicted domains or other features
	Hypothetical protein group_3340	339	No	No	Yes	No	No confidently predicted domains or other features
	PAS domain S-box protein	363	No	No	Yes	No	PAS domain
	Putative chromate transport protein SrpC	1311	No	No	Yes	No	Eleven transmembrane domains
pascullei	ABC-type sugar transport system, periplasmic component	792	No	No	No	Yes	TIR-like superfamily domain at N-terminus
	BNR/Asp-box repeat protein	1461	No	No	No	Yes	Transmembrane domain at N-terminus and a low complexity domain at C-terminus
	Diguanylate cyclase (GGDEF) domain protein	243	No	No	No	Yes	No confidently predicted domains or other features
	Immunogenic protein MPT70 precursor	471	No	No	No	Yes	Transmembrane domain and FAS1 domain
	Bifunctional AAC/APH	585	No	No	No	Yes	No confidently predicted domains or other features

[a] aa stands for amino acids.