Research

# Multiple major disease-associated clones of *Legionella pneumophila* have emerged recently and independently

Sophia David,[1,2,9] Christophe Rusniok,[3,4,9] Massimo Mentasti,[2] Laura Gomez-Valero,[3,4] Simon R. Harris,[1] Pierre Lechat,[5] John Lees,[1] Christophe Ginevra,[6] Philippe Glaser,[4,7] Laurence Ma,[8] Christiane Bouchier,[8] Anthony Underwood,[2] Sophie Jarraud,[6] Timothy G. Harrison,[2] Julian Parkhill,[1] and Carmen Buchrieser[3,4]

[1]*Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA Cambridge, United Kingdom;* [2]*Public Health England, NW9 5HT London, United Kingdom;* [3]*Institut Pasteur, Biologie des Bactéries Intracellulaires, 75015, Paris, France;* [4]*CNRS UMR 3525, 75724, Paris, France;* [5]*Hub de Bio-informatique et Biostatistiques, Centre de Bio-informatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, 75724, Paris, France;* [6]*National Reference Centre for Legionella, Hospice Civil de Lyon, International Center for Infection Research, Legionella pathogenesis Team, 69364, Lyon, France;* [7]*Institut Pasteur, Evolution et Ecologie de la Resistance aux Antibiotiques, 75724, Paris, France;* [8]*Institut Pasteur, Plate-Forme Génomique, 75724, Paris, France*

*Legionella pneumophila* is an environmental bacterium and the leading cause of Legionnaires' disease. Just five sequence types (ST), from more than 2000 currently described, cause nearly half of disease cases in northwest Europe. Here, we report the sequence and analyses of 364 *L. pneumophila* genomes, including 337 from the five disease-associated STs and 27 representative of the species diversity. Phylogenetic analyses revealed that the five STs have independent origins within a highly diverse species. The number of de novo mutations is extremely low with maximum pairwise single-nucleotide polymorphisms (SNPs) ranging from 19 (ST47) to 127 (ST1), which suggests emergences within the last century. Isolates sampled geographically far apart differ by only a few SNPs, demonstrating rapid dissemination. These five STs have been recombining recently, leading to a shared pool of allelic variants potentially contributing to their increased disease propensity. The oldest clone, ST1, has spread globally; between 1940 and 2000, four new clones have emerged in Europe, which show long-distance, rapid dispersal. That a large proportion of clinical cases is caused by recently emerged and internationally dispersed clones, linked by convergent evolution, is surprising for an environmental bacterium traditionally considered to be an opportunistic pathogen. To simultaneously explain recent emergence, rapid spread and increased disease association, we hypothesize that these STs have adapted to new man-made environmental niches, which may be linked by human infection and transmission.

[Supplemental material is available for this article.]

A number of environmental bacteria have emerged to become human pathogens, either through accidental infection or adaptation to the human host. One example is *Legionella*, a bacterium that is ubiquitous in natural aquatic environments but also a contaminant of modern, man-made water systems. *Legionella* can infect humans, mainly through inhalation of contaminated aerosols, and can cause a severe, sometimes fatal, pneumonia known as Legionnaires' disease (LD) (Fields et al. 2002; Newton et al. 2010). Since the first cases were reported in Philadelphia, Pennsylvania, in 1976, *Legionella* has increasingly been recognized as an important cause of both community- and hospital-acquired pneumonia worldwide (Phin et al. 2014).

*Legionella* are intracellular bacteria whose survival depends on the ability to replicate in eukaryotic cells such as aquatic protozoa (Rowbotham 1980). It is thought that the conservation of signaling pathways and cellular functions from protozoa to higher eukaryotes allows *Legionella* to also infect human alveolar macrophages. However, although humans have traditionally been considered a dead-end host for *Legionella*, one probable case of person-to-person transmission has recently been reported (Correia et al. 2016).

Interestingly, among the 62 species known in the genus *Legionella*, *Legionella pneumophila* is responsible for >90% of known LD cases worldwide (Yu et al. 2002). The prevalence of different *L. pneumophila* subtypes among clinical isolates is also unevenly distributed (Beauté et al. 2013). For example, of the 15 serogroups (sg) described, just one (sg1) is responsible for >80% of culture-confirmed LD cases (Yu et al. 2002; Beauté et al. 2013). Furthermore, data from the sequence-based typing (SBT) scheme (analogous to multilocus sequence typing), that allows a subdivision of *L. pneumophila* into sequence types (ST) based on the sequence of seven genetic loci (http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php), revealed that a small subset of STs accounts for a disproportionately high number of clinical cases. In particular, five STs (1, 23, 37, 47, and 62) have accounted for nearly half of all epidemiologically unrelated LD cases in

[9]**These authors are joint first authors and contributed equally to this work.**

**Corresponding authors: cbuch@pasteur.fr, parkhill@sanger.ac.uk**

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.209536.116. Freely available online through the *Genome Research* Open Access option.

northwest Europe reported to the SBT database (Fig. 1A; Supplemental Methods). There is no evidence that the high proportion of isolates found in clinical samples belonging to these five STs is a result of laboratory artifacts such as an increased growth of these STs in culture compared with other STs. Data from 2009 to 2014 obtained by SBT on clinical isolates ($n = 1762$) and nested-PCR-based SBT (NP-SBT) performed directly from respiratory samples of patients ($n = 99$), confirmed that STs 1, 23, 47, and 62 are the major STs in France, representing 41.6% of all STs from SBT and 46.5% of all STs from NP-SBT. The distribution of STs among culture-proven and culture-negative but NP-SBT positive patients appears to be the same for ST1 (9.8%/9.1% for SBT and NP-SBT, respectively); ST23 (18.5%/26.3%), and ST62 (4.8%/8.1%) appear a little more represented in NP-SBT, and ST47 appears a little less represented in NP-SBT (8.5%/3.0%). Furthermore, *L. pneumophila* strains detected by culture (~65% of LD cases) or PCR (~20% of LD cases) in another study show a similar distribution of STs (Mentasti et al. 2012).

Although ST47 has been nearly exclusively isolated in northwest Europe (a small number of cases have been reported from Canada), the other STs have been detected worldwide (Fig. 1A). In particular, ST1 (also known as the "Paris-like strain") shows a global distribution (Cazalet et al. 2008). The mechanism of global spread is unknown. Interestingly, except for ST1, these disease-associated STs are rarely isolated from either natural or man-made environmental sources (Fig. 1B).

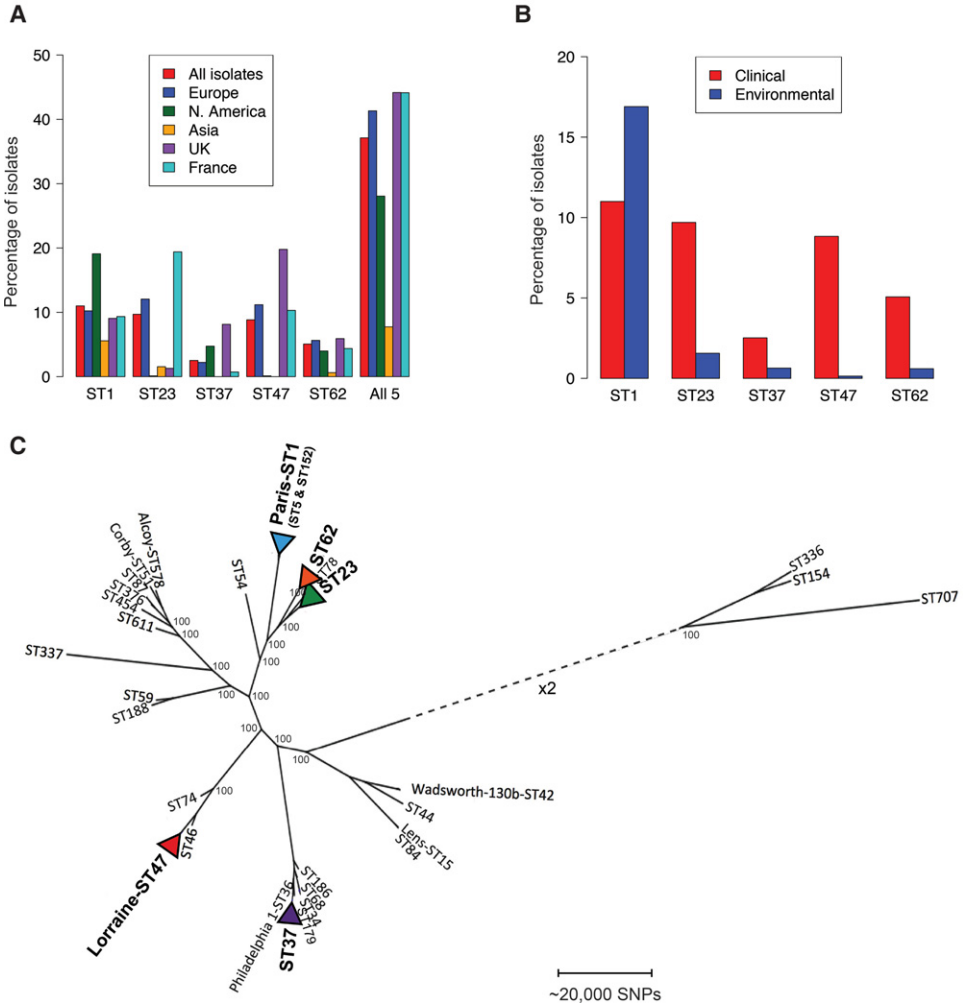


**Figure 1.** Geographical and environmental distribution of the five major disease-associated STs and their phylogenetic context within the species. (*A*) The percentage of clinical isolates submitted to the ESGLI sequence-based typing (SBT) database for *L. pneumophila* that are sequence types (STs) 1, 23, 37, 47, or 62, or from all five STs combined, are shown. Data are shown from different regions and countries where the numbers of isolates submitted were deemed sufficient for comparison. Data are based on a total of 6116 epidemiologically unrelated clinical isolates (i.e., including only one isolate from clusters/outbreaks) submitted to the database, 4785 of which were detected in Europe, including 541 in the United Kingdom and 2313 in France, as well as 801 in North America and 323 in Asia. (*B*) The percentage of clinical and environmental isolates from each of the five STs from the total number of clinical and environmental isolates submitted to the SBT database. Data are based on a total of 6116 epidemiologically unrelated clinical and 2826 unrelated environmental isolates (i.e., including only one isolate from clusters/outbreaks) submitted to the database. (*C*) Maximum likelihood phylogeny, generated by mapping sequence reads to the Corby reference genome, of representatives from each of the five dominant disease-associated STs together with an additional 27 isolates that are representative of the species diversity. The size of the alignment was 3,576,470 bp. Generally, the five STs are found in separate major clades of the species except for ST23 and ST62 that share a closer phylogenetic relationship. The scale represents the number of SNPs. Bootstrap values, derived from 1000 resamples, are shown for the major nodes of the tree.

Whole-genome comparisons of *L. pneumophila* isolates have indicated that *L. pneumophila* is a genetically diverse and ancient species (Gomez-Valero et al. 2011; Underwood et al. 2013). Thus, key questions include how and when did these disease-associated STs evolve, and how have they been able to spread globally? Here, we undertook genomic analysis of 337 isolates belonging to five dominant disease-associated STs together with 27 additional *L. pneumophila* isolates representative of the species diversity to investigate their emergence as important human pathogens.

## Results

### The major disease-associated STs have emerged independently

We first analyzed the position of five representative isolates belonging to the major disease-associated STs within a phylogenetic tree containing a total of 32 previously sequenced *L. pneumophila* genomes. These represented the most distantly related STs in the database at the time of their selection (Supplemental Table S1; Underwood et al. 2013). This analysis showed that the five major disease-associated STs are found in separate major clades of the species tree with the exception of ST23 and ST62 that share a closer phylogenetic relationship (Fig. 1C). This suggests that the dominant disease-associated STs have evolved independently from within a genetically diverse species.

To investigate the evolution and diversity of each of the five STs, we separately analyzed 71 ST1 (59 ST1 and 12 "ST1-like isolates"), 37 ST23, 72 ST37, 122 ST47, and 35 ST62 isolates (Supplemental Tables S1, S2) by mapping sequence data to a selected reference genome of the same ST (Supplemental Table S3).

### ST47 isolates show no recombination and are highly clonal

The 122 ST47 isolates were recovered between 1994 and 2013 from the United Kingdom and France, but also include some travel-associated isolates for which the origin is uncertain. In recent years, ST47 has become the most common cause of Legionnaires' disease in northwest Europe, accounting for more than one-quarter of cases in England and Wales, the Netherlands, France, and Belgium (Harrison et al. 2009; Vekens et al. 2012; Euser et al. 2013; Cassier et al. 2015), yet rarely isolated outside of Europe. ST47 isolates are also extremely rarely isolated from the environment and sources of infection usually remain unknown (Fig. 1B; Edelstein and Metlay 2009; Gomez-Valero et al. 2011).

Surprisingly, the maximum number of SNPs between any pair of the 122 ST47 isolates is just 19. Furthermore, 21 isolates recovered between 2003 and 2012 from distant regions of the United Kingdom have no SNPs at all, and a further 17 isolates are just one SNP different from these 21 isolates. No SNPs are homoplasic, and no recombination events were detected using Gubbins, which uses high SNP density as a marker for recombined regions. This is an expected result given the sparse distribution of SNPs, as visualized using the SynTView program (Fig. 2A; Lechat et al. 2011).

### Recombination is the major driving force of diversity within STs 1, 23, 37, and 62

In contrast, the ST1 isolates were recovered between 1981 and 2011 from 14 countries over four continents (Europe, Asia, North America, and Africa). We included isolates within this data set, here termed "ST1-like isolates," that are nested within, and thus evolved from, ST1 isolates. We have also sequenced and analyzed the oldest known isolate of *L. pneumophila* (OLDA1/ST1_31), which is an ST1 isolated in 1947, almost 30 years prior to the description of the species. Of the five STs analyzed, the ST1 lineage exhibits the greatest diversity with a maximum of 15,227 SNPs between the two most distant isolates, a sharp contrast to the low number of SNPs observed in the ST47 lineage (Supplemental Table S3).

Isolates belonging to STs 23, 37, and 62 are commonly isolated across Europe and occasionally elsewhere (Fig. 1A); thus, the extent of their distributions appears to be between those of ST1 and ST47. Like ST47 isolates, they are only rarely isolated from commonly expected environmental sources of *Legionella* (Fig. 1B). The ST23 and ST37 isolates analyzed in this study were recovered between 1987 and 2012, and the ST62 isolates were recovered between 1994 and 2012. The SNP analyses showed that the maximum pairwise SNP differences between isolates are 12,964, 13,776, and 12,842 within the ST23, ST37, and ST62 lineages, respectively (Supplemental Table S3), slightly lower than that observed in the ST1 lineage.

However, when we analyzed the origin of these nucleotide variants using Gubbins, we found that 96.3%–99.0% of SNPs in STs 1, 23, 37, and 62 had been acquired by recombination (Table 1). There was a high level of concordance between these results and those obtained using an alternative recombination detection software, BRATNextGen (Marttinen et al. 2012), which uses a hidden Markov model (HMM) to detect SNP patterns in an isolate that are more similar to those from another phylogenetic clade than the isolate's own clade. Overall, >90% of SNPs identified as recombined by Gubbins were confirmed by BRATNextGen to be within horizontally exchanged regions. The importance of recombination within these STs becomes very apparent when the SNP distributions are visualized using SynTView (Fig. 2B–E). The locations and content of the recombined regions are provided in Supplemental Tables S4–S8, and the distribution of recombination fragment lengths is shown in Supplemental Figure S1. Further details on their composition and predicted origin are also given in the Supplemental Results.

After removal of recombined regions (representing a mean of 3.4% [ST23] to 12.9% [ST62] of the 3.2-Mb genomes) from the analysis of STs 1, 23, 37, and 62 to leave only those SNPs generated by point mutation, the maximum pairwise SNP differences ranged from just 59 (ST23) to 127 (ST1), similar to the 19 SNPs observed in the ST47 lineage (Table 1). There might be additional de novo mutations that have occurred within the recombined regions and removed from this analysis; however, these are unlikely to constitute more than another 12.9%, in proportion with the length of genome removed. Thus, all five STs are characterized by a very low number of de novo mutations, in sharp contrast to the high species diversity.

### The major disease-associated clones emerged very recently

The small number of de novo mutations within each of the five major disease-associated STs suggests that these are very recently emerged lineages. To estimate their emergence date, we attempted to date the most recent common ancestor (MRCA) of each lineage using linear regression of root-to-tip distances against time, as well as with a Bayesian coalescent model as implemented in the BEAST software (Drummond et al. 2012). Only the ST37 lineage showed some temporal signal in terms of SNP accumulation using Path-O-Gen (Supplemental Fig. S2). Using a relaxed molecular clock model in BEAST, we estimated that the ST37 clone emerged in about 1979 (95% highest posterior density [HPD] intervals:

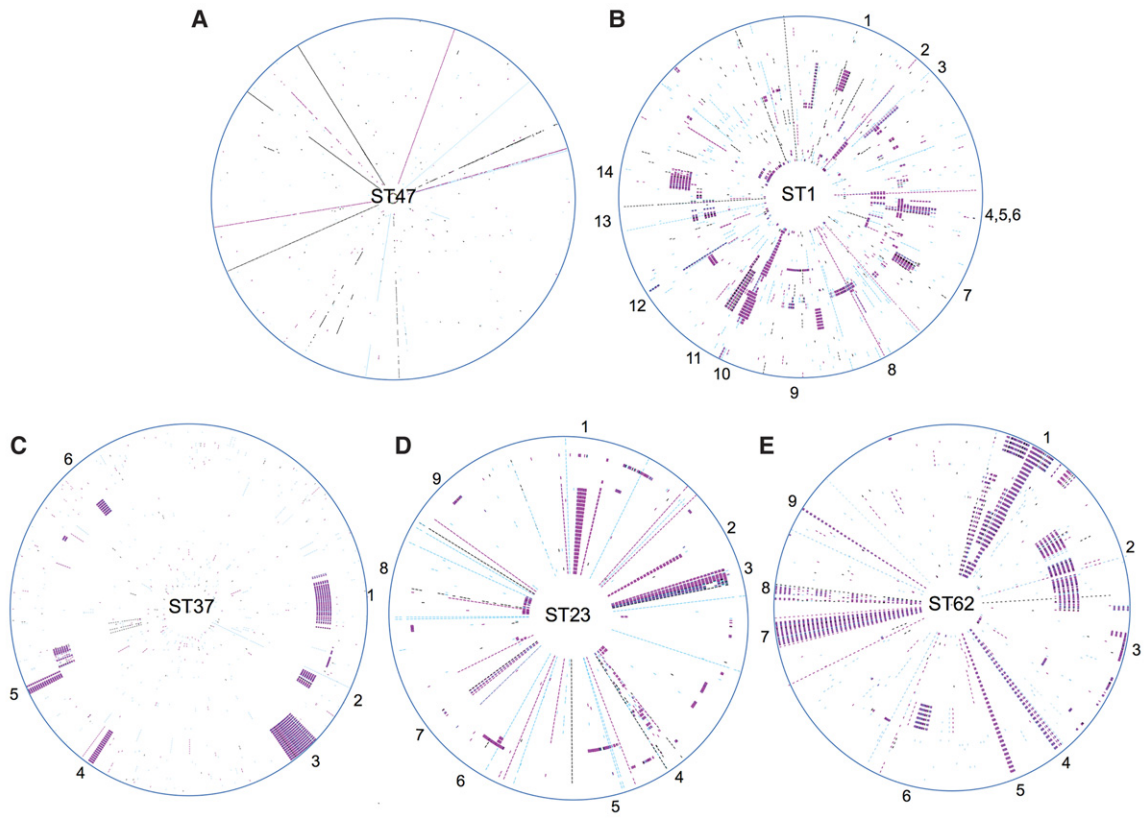


**Figure 2.** Distribution of SNPs across *L. pneumophila* ST47 (*A*), ST1 (*B*), ST37 (*C*), ST23 (*D*), and ST62 (*E*) lineages. Circular representation of the five major disease-associated STs with each genome shown as a concentric circle. The maps were generated with the SynTView software. SNPs with respect to the reference genome are indicated by short lines in the concentric circle. Recombined regions can be seen as regions with a higher density of SNPs. SNPs are colored according to the type of mutation: (black) intergenic; (pink) synonymous; (blue) nonsynonymous. Around the *outside* circle, selected recombined loci are indicated with numbers, the content of which is provided in Supplemental Tables S5–S8. Access to the interactive SNP server is available at: http://genopole.pasteur.fr/SynTView/flash/Legionella_st/SynWebST1.html, http://genopole.pasteur.fr/SynTView/flash/Legionella_st/SynWebST47.html, http://genopole.pasteur.fr/SynTView/flash/Legionella_st/SynWebST37.html, http://genopole.pasteur.fr/SynTView/flash/Legionella_st/SynWebST23.html, http://genopole.pasteur.fr/SynTView/flash/Legionella_st/SynWebST62.html. The circular representation can be visualised using the "Circular" tab. An alternative linear representation can be displayed using the "Local View" tab (then by selecting the horizontal line between the arrows in the panel on the left hand side, then "snp" and "show/hide snps").

1968–1985) (Fig. 3), which is 3 yr prior to the earliest ST37 isolate recorded in the SBT database. The evolutionary rate estimated by BEAST is $2.07 \times 10^{-7}$ substitutions per site per year (95% HPD interval: $1.69 \times 10^{-7}$ to $2.44 \times 10^{-7}$), very similar to that previously calculated for the ST578 lineage ($1.39 \times 10^{-7}$) (Table 2; Sánchez-Busó et al. 2014). We further used the estimated evolutionary rates of the ST37 and ST578 lineages to provide approximations of the length of time it would have taken for the observed diversity in STs 1, 23, 47, and 62 to arise. This analysis suggested emergence dates of 1851/1899 for ST1, 1972/1983 for ST23, 1943/1964 for ST62, and 1998/2002 for ST47, with the two dates provided corresponding to the application of the ST578 and ST37 mean evolutionary rates, respectively. Further details on all dating analyses are provided in the Supplemental Methods and Results.

## The disease-associated clones have spread rapidly and internationally

Phylogenetic analyses of the five STs show that isolates from the same country do not always cluster together, whereas isolates

**Table 1. Percentages of SNPs that have arisen via recombination and the number of vertically inherited mutations outside recombined regions in each of the five major disease-associated STs**

| Lineage (number of isolates) | Mean length of the genome (bp) involving recombination | Number of vertically inherited SNPs | Maximum number of vertically inherited SNPs between two isolates | Percentage of SNPs in recombined regions | Percentage of vertically inherited homoplasic SNPs |
|---|---|---|---|---|---|
| ST1 (71) | 335,382 (9.6%) | 867 | 127 | 98.2 | 5.6 |
| ST23 (37) | 118,597 (3.4%) | 182 | 59 | 99.3 | 2.2 |
| ST37 (72) | 144,953 (4.2%) | 546 | 75 | 96.3 | 1.1 |
| ST47 (122) | 0 (0%) | 186 | 19 | 0 | 0 |
| ST62 (35) | 447,320 (12.9%) | 335 | 110 | 99.0 | 4.9 |

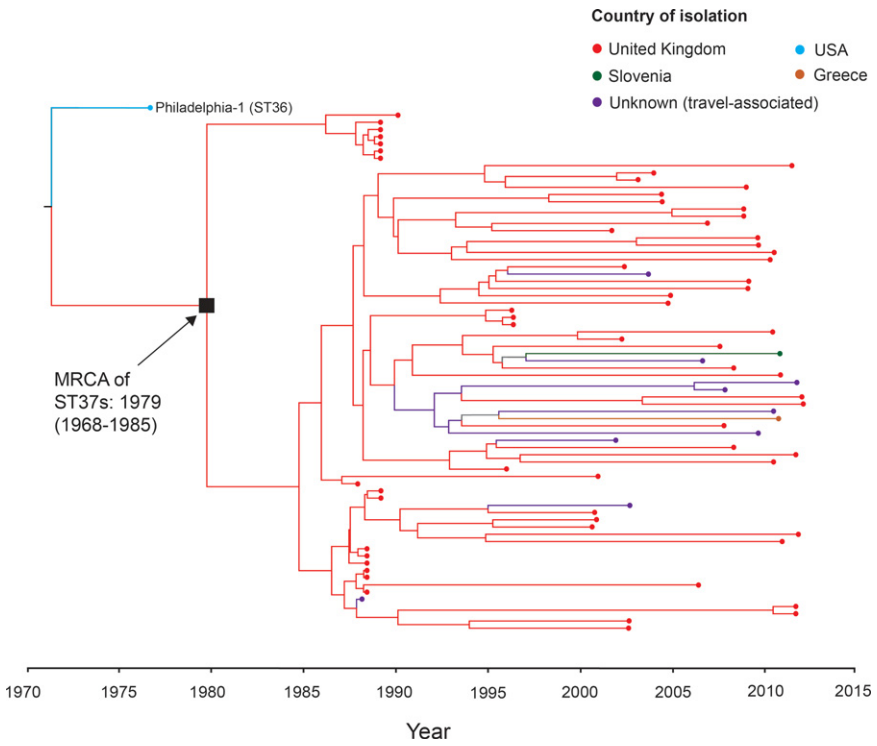


**Figure 3.** Maximum clade credibility tree of the ST37 lineage showing the estimated age of the MRCA. A time-dependent phylogenetic reconstruction of the ST37 lineage, inferred by Bayesian inference using BEAST, is shown. The Philadelphia isolate (a single locus variant of ST37) was also included in the analysis as an out-group. The node representing the MRCA of the ST37 lineage is labeled with the median estimate for the inferred date and the 95% highest posterior probability (HPD) intervals. Isolates are colored according to the country of isolation, and branches are similarly colored to indicate the origin of descendant nodes.

from distant geographical regions frequently cluster very closely (Figs. 3, 4; Supplemental Figs. S3–S7). This is most apparent in the globally dispersed ST1 lineage, but true in all lineages, including the more geographically restricted ST47 lineage, whereby isolates from the United Kingdom are nested within a cluster of predominantly French isolates (Fig. 4B). These results demonstrate the occurrence of multiple, recent, long-distance spreading events. It is also notable that the geographical distribution of these STs correlates with their predicted ages. For example, ST1 is estimated to be the oldest lineage and is distributed globally, whereas ST47 is the youngest predicted lineage and is mostly restricted to northwest Europe.

## The disease-associated clones show evidence of convergent evolution via recent recombination

Next, we investigated whether specific signatures of convergent evolution exist between the five STs that could explain possible adaptation to a common niche or increased propensity to cause disease compared with other STs. Although many of the specific isolates from the other STs were from LD patients, the STs to which they belong are far less associated with disease than isolates belonging to STs 1, 23, 37, 47, and 62. Analysis of the gene content using de novo assemblies of all isolates did not identify any genes specifically present in the five STs but absent from the other genomes from our collection. Analyses of the pan-genome of each of the five STs individually showed that the pan-genome content of each ST is beginning to plateau, suggesting that this gene analysis is representative of each ST (Fig. 5A). For example, the 306 known substrates of the Dot/Icm secretion system, key virulence factors of *L. pneumophila*, were all highly conserved across the five STs (Supplemental Table S9). Further details of this analysis are provided in the Supplemental Results. This finding led us to focus our attention on core genes (i.e., genes that are shared among all isolates).

Analysis of all core genes using CodeML identified none that had been subjected to positive selection on more than one of the five branches in the species tree leading to each of the disease-associated STs, a result which could have indicated common adaptation to a particular niche. Further details are provided in the Supplemental Methods and Results. However, we did identify seven SNPs that are convergent on four of these branches, and 38 on three branches (Supplemental Table S10). One of the SNPs that occurred on four of the five branches (leading to STs 1, 37, 47, and 62) causes an amino acid change and was also found on two other branches of the species tree. This SNP is in *lpp0942/lpg0879*, a gene that encodes a diguanylate kinase with a GGDEF domain,

**Table 2.** Evolutionary rates of different bacterial pathogens

| Species | Evolutionary rate estimates (SNPs/site/year, unless specified)[a] | Reference |
|---|---|---|
| *Yersinia pestis* | $2.9 \times 10^{-9}$–$2.3 \times 10^{-8}$ | Morelli et al. 2010 |
| *Mycobacterium tuberculosis* | 0.3–0.5 SNPs/genome/year | Ford et al. 2013 |
| *Mycobacterium abscessus* (two subspecies) | 0.47–1.8 SNPs/genome/year | Bryant et al. 2013 |
| *Legionella pneumophila* (ST578) | $1.39 \times 10^{-7}$ (0.49 SNPs/genome/year) | Sánchez-Busó et al. 2014 |
| *Legionella pneumophila* (ST37) | $2.07 \times 10^{-7}$ (0.71 SNPs/genome/year) | This study |
| *Pseudomonas aeruginosa* | $4 \times 10^{-7}$ | Marvig et al. 2013 |
| *Shigella sonnei* (global collection) | $6.0 \times 10^{-7}$ (2.2 SNPs/genome/year) | Holt et al. 2012 |
| *Vibrio cholerae* | $8.3 \times 10^{-7}$ | Mutreja et al. 2011 |
| *Klebsiella pneumoniae* (ST258) | $1.03 \times 10^{-6}$ (3.9 SNPs/genome/year) | Bowers et al. 2015 |
| *Staphylococcus aureus* (ST8) | $1.22 \times 10^{-6}$ (~3 SNPs/genome/year) | Uhlemann et al. 2014 |
| *Streptococcus pneumoniae* (13 different clades) | $1.45 \times 10^{-6}$–$4.81 \times 10^{-6}$ | Chewapreecha et al. 2014 |

[a]The units provided are those given in the original publications.

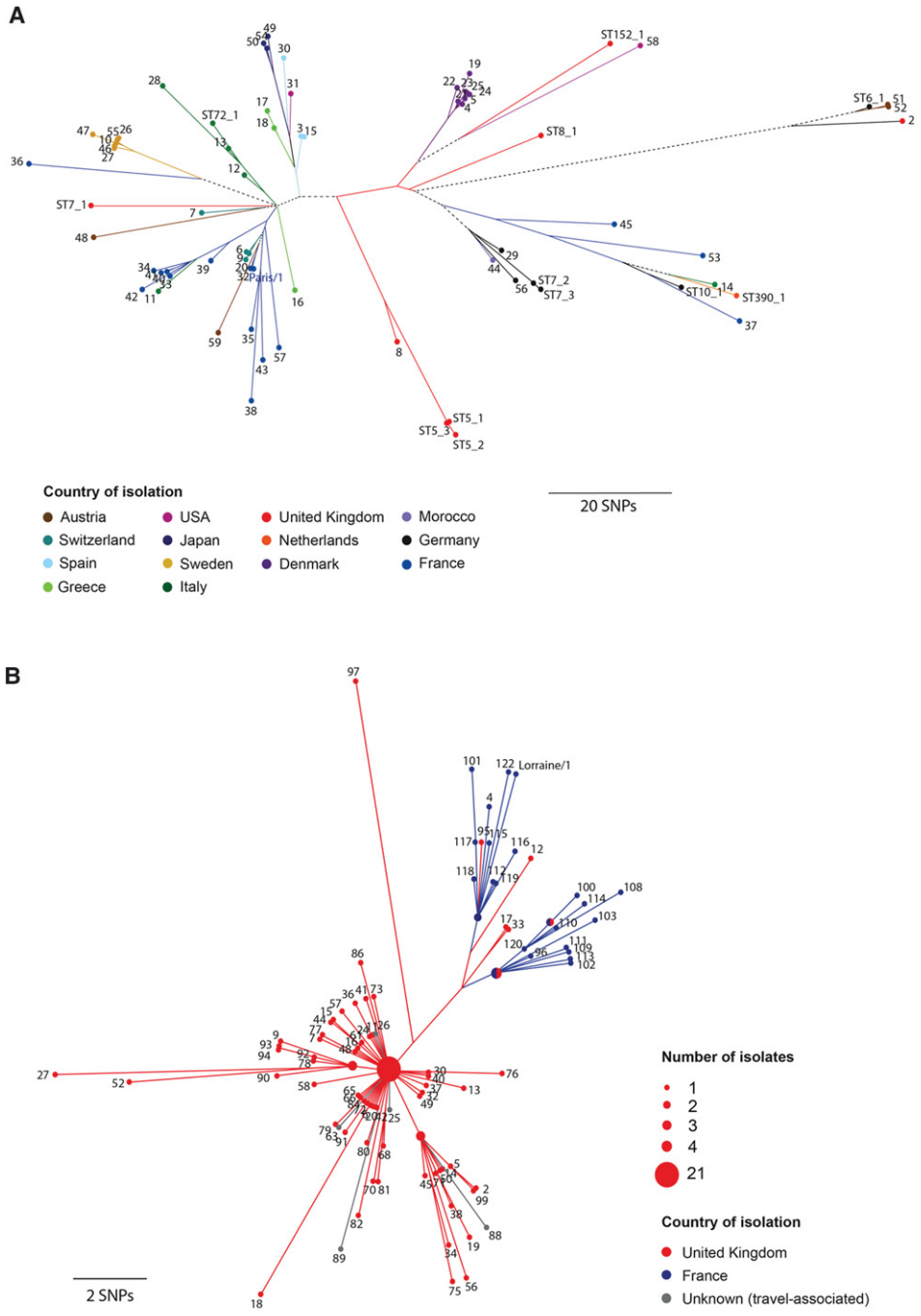


**Figure 4.** Maximum likelihood trees of 71 ST1 and 122 ST47 isolates. (*A*) A phylogeny of ST1 isolates, constructed using 867 SNP differences remaining after recombined regions were removed. (*B*) A phylogeny of ST47 isolates, constructed using 186 SNP differences. Isolates are colored according to the country of isolation, and branches are similarly colored to indicate the origin of descendant nodes. A black dotted line is used where there are descendant nodes from multiple countries. The scales indicate the number of SNPs that have occurred for a given branch length.

which is strongly induced in the virulent, transmissive phase of infection and belongs to the transmissive phase core genes (Brüggemann et al. 2006; Weissenmayer et al. 2011). However, further studies are required to test if this SNP, or any of the others, affect disease propensity.

Finally, we searched for core genes with a higher than expected nucleotide similarity in the five STs with respect to the rest of the species. This approach bypasses a limitation of the previous approaches by taking into account all evolution that has occurred en route to the formation of the five disease-associated STs, rather than searching for evidence of convergent evolution on individual, sometimes short, branches leading to each lineage. To perform this analysis, we first identified core genes present in all 32 STs with the exclusion of ST154, ST336, and ST707 that are distantly

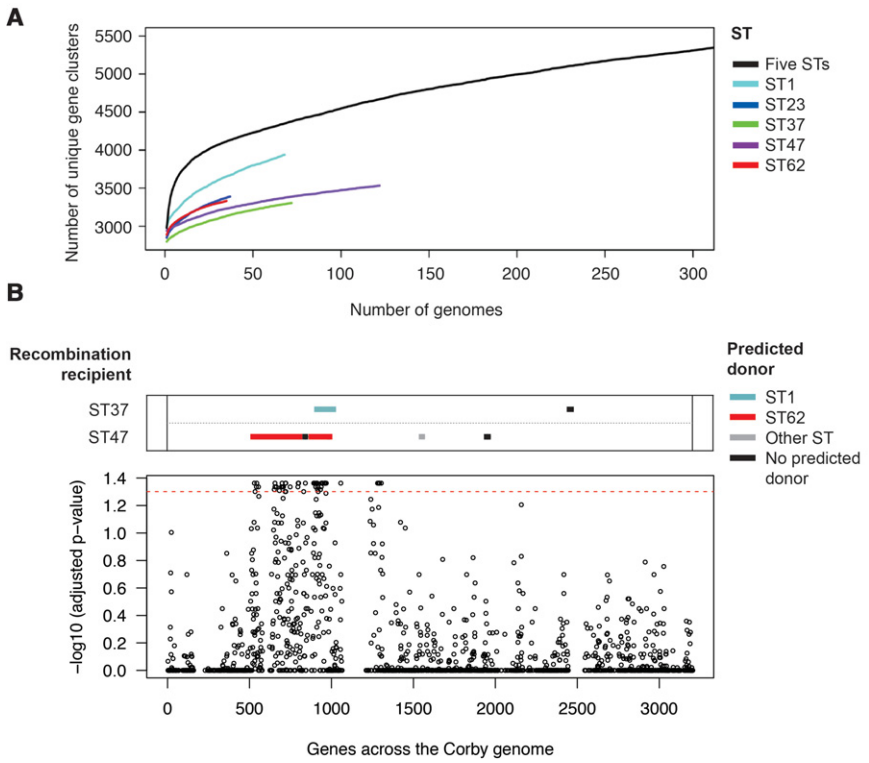


**Figure 5.** Gene content analysis and the nucleotide diversity of core genes within the five disease-associated STs. (*A*) Rarefaction curves applied to the strains of *L. pneumophila* ST1 (71 isolates), ST23 (37 isolates), ST37 (72 isolates), ST47 (122 isolates), ST62 (35 isolates), and all five STs together (337 isolates), showing that differences in gene content exist among the five STs, but that the number of novel genes in the overall pan-genome is beginning to plateau. (*B*) Log-transformed *P*-values derived from testing whether the five disease-associated STs have lower than expected nucleotide diversity values in individual core genes given their nucleotide diversity across all 1888 core genes, and with respect to the gene conservation across the species (excluding isolates from the distant subspecies, strains ST5 and ST152, which are nested within ST1, and strains ST36 [Philadelphia], ST42, and ST578 [Alcoy], which are also disease-associated strains). The core genes are ordered as in the Corby genome. Any noncore genes (genes in <100% isolates) are omitted. The horizontal dotted red line indicates the significance threshold when the Benjamini-Hochberg method is applied to correct for multiple testing. The box at the *top* shows the location and predicted origins of recombined regions that were detected on the branches leading to the ST37 and ST47 lineages. Recombined regions that were found in the ST37 and ST47 accessory genomes are not shown.

related. Using one representative isolate from each of STs 1, 23, 37, 47, and 62, we calculated the nucleotide diversity ($\pi$) value, as first described by Nei and Li (1979), between the five isolates for each of the 1888 core genes. Interestingly, a ~700-kb region of the genome was identified that contains several genes with a very low nucleotide diversity ($\pi$) value in the five STs as compared to the rest of the genome. Because this region could simply be more conserved across the species in general, we next tested whether each gene was significantly more similar between the five disease-associated STs than expected given its degree of conservation across the entire species. We also took into account the nucleotide diversity ($\pi$) values observed in the five STs across the whole genome, thus accounting for phylogenetic distance. Further details of these methods are provided in the Supplemental Methods. After correcting for multiple testing, we found that nucleotide diversity ($\pi$) in the five disease-associated STs was statistically lower than expected in 64 genes ($P < 0.05$) (Supplemental Table S11), which are all located in the aforementioned region of 725.1 kb (*lpp0536*/*LPC_2873* to *lpp1176*/*LPC_0640*) (Fig. 5B; Supplemental Fig. S8). Maximum likelihood trees of selected individual gene alignments confirmed that the genes from the five STs cluster together in contrast to their positions within the whole-genome phylogeny (Supplemental Fig. S9). Among the 64 genes, some have been shown to play a role in intracellular infection such as the genes from the cytochrome *c* maturation (*ccm*) locus (Viswanathan et al. 2002; Naylor and Cianciotto 2004), PilR (an important regulator for pilin and flagella synthesis), the phagosomal transporter family Pht (Sauer et al. 2005), or the enhanced entry protein EnhA that was shown to be important for entry in phagocytic cells (Cirillo et al. 2000) and during persistence in water environments (Li et al. 2015).

The detection of shared gene variants within the five STs led us to hypothesize that they have arisen via recombination events prior to the emergence of these major disease-associated STs. Using Gubbins, we were able to detect recombination events on the branches leading to STs 37 and 47, because they contain relatively few SNPs and allow the SNP-dense recombined regions to be detectable above the background level. Indeed, we detected a number of imported recombined regions on both branches that shared 100%, or almost 100%, similarity with other major disease-associated STs (Fig. 5B; Supplemental Table S12) and that lie within the previously detected region of highly similar genes. Of particular note is the large amount of sequence (396,135 bp) imported from the ST62 lineage to ST47, which makes up 11.4% of the ST47 chromosome (Supplemental Fig. S8).

## Discussion

Genomic and phylogenetic analysis of 364 *L. pneumophila* isolates revealed that five major disease-associated STs emerged independently from different genomic backgrounds. In contrast to the high species diversity, they show remarkably little diversity (excluding recombined regions), suggesting recent clonal origins. Further support for the recent emergence of these STs is provided by BEAST analysis of the ST37 lineage, which predicts the most likely emergence date to be between 1968 and 1985. When the estimated evolutionary rate of the ST37 lineage, and that of the previously published Alcoy lineage (Sánchez-Busó et al. 2014), is applied to the four remaining lineages, emergences in the last century are also predicted. Although these observations might be expected for a human-adapted pathogen, the results are surprising for an environmental bacterium that is traditionally thought to "accidentally" infect humans when given the opportunity. The results suggest that these *L. pneumophila* clones have adapted to new niches that presumably are related to modern, man-made water systems, from which the majority of infections are acquired. However, because most of these disease-associated STs are not

more frequently detected in commonly expected environmental sources than other STs, and indeed some are rarely found, it is possible that they are also more efficient at infecting humans.

Because the five disease-associated STs have emerged independently from within the species, we explored whether there are signs of convergent evolution that could explain their common adaptation to specific niches and increased propensity to cause human disease. Indeed, we identified many genes with particular allelic variants in the five STs that are rarely seen in other STs, and which have arisen at least partially via recombination events. This finding, together with observations from this study and others (Gomez-Valero et al. 2011; Sánchez-Busó et al. 2014) that recombination accounts for almost all the observed diversity in some STs, confirms the importance of this process for *L. pneumophila* evolution and the emergence of disease-associated lineages.

In contrast to all other STs studied, no recombination was detected within the ST47 lineage. Although this lack of observed recombination events may simply reflect its very recent emergence leaving no time for recombination to occur, it could also suggest that ST47 inhabits a specific environmental niche where no opportunity for recombination exists. ST47 may also have lost the ability to recombine with other *L. pneumophila* strains, or have lost natural competence. However, we have been able to construct a streptomycin-resistant ST47 isolate by natural competence, and thus the latter possibility can be ruled out.

The evolutionary rate, as estimated by BEAST, is very low with on average $2.07 \times 10^{-7}$ SNPs/site/year (0.71 SNPs/genome/year) for the ST37 lineage (Table 2). This is similar to the rate of 0.49 SNPs/genome/year estimated for the Spanish ST578 lineage (Sánchez-Busó et al. 2014). Further support for a low evolutionary rate is provided by the 21 identical ST47 isolates recovered between 2003 and 2012 as well as the existence of only 20 vertically inherited SNPs between the OLDA1 isolate from 1947 and another ST1 isolate from 1995. The evolutionary rate is comparable to that described for *Mycobacterium tuberculosis*, a notoriously slow-evolving pathogen (Ford et al. 2013). Together with the absence of a strict molecular clock, the low evolutionary rate suggests that *Legionella* may undergo periods of dormancy, either within amoebas, biofilm, or during its free-living phase, and perhaps also due to the water temperature in temperate climates being <15°C for most of the year.

Phylogenetic analyses of each of these disease-associated STs showed that isolates from different countries, and even different continents, differ by just a few SNPs (Figs. 3, 4; Supplemental Figs. S3–S7). This demonstrates the occurrence of long-distance spread, which must have occurred relatively rapidly given the recent emergence of these STs. Possible spreading mechanisms include wind transport, natural water currents, or human-related activities such as the movement of contaminated vehicles, ships, or other objects harboring water. The latter possibility may also explain the recent emergence of these clones, because they may have adapted to the new environments. However, it would be very surprising that those few STs that are highly associated with human disease have also spread widely and rapidly, if these phenomena were unlinked.

The observations gained from our evolutionary, phylogenetic, and comparative genome analyses lead us to hypothesize that *L. pneumophila*–infected humans may indeed contribute to the spread of these highly disease-causing strains by linking modern man-made water systems through human transmission. *L. pneumophila* has been isolated from human feces (Rowbotham 1998) and is regularly isolated from the sputa of Legionnaires' disease patients, suggesting that human infection may not actually be a dead end. Further support for this has also come from a recently reported case of probable person-to-person transmission of Legionnaires' disease (Correia et al. 2016). Scenarios involving human-to-human transmission and/or human-to-environment transmission could simultaneously explain why these specific strains have emerged recently, spread widely, and are primarily associated with human infection. Adaptation to man-made water systems, when coupled with human infection and transmission at least partially via humans, would select strains most fit for human infection. These would then be more frequently transmitted to other man-made water systems. Humans as vectors would also link similar sites, enhancing the ability of these clones to adapt to this niche and promoting long-distance transmission. Such a scenario would effectively create a new evolutionary niche and allow expansion and further adaptation of these clones. The finding that this has happened independently to multiple strains suggests that it is the new niche that has driven the establishment and expansion of these strains, rather than the attributes of a specific strain.

The discovery of several independent disease-associated clones that are recently emerged has important implications for the understanding of Legionnaires' disease. In particular, our data support the idea that the majority of clinical cases do not arise due to infection by any *L. pneumophila* strain that happens to be present in a source, but rather are caused by selected clones that may have adapted to a specific niche. Identifying the environmental niche and mechanism of spread of these clones should become a priority if we are to reduce human exposure to *L. pneumophila* and alleviate the disease burden. We also believe that our hypothesis that specific *L. pneumophila* clones may have shifted from being accidental to more human-adapted pathogens is worthy of further investigation.

## Methods

### Bacterial isolates and whole-genome sequencing

Of the 364 *L. pneumophila* isolates used in this study, 35 have been previously sequenced, whereas 329 are newly sequenced (Supplemental Tables S1, S2). All newly sequenced isolates are from the culture collections at Public Health England (PHE), United Kingdom, and the *Legionella* National reference center, Lyon (LRC), France. Details of these isolates and their sequencing are provided in the Supplemental Methods.

### Mapping of sequence reads and phylogenetic analysis

Sequence reads were mapped to a reference genome using SMALT v0.7.4, and SNPs were identified using a standard approach (Harris et al. 2010). Further details are provided in the Supplemental Methods. After removing recombined regions, as defined by Gubbins (Croucher et al. 2014) (except for the species representatives, as the large amount of diversity renders recombination detection very difficult), maximum likelihood trees were constructed based on variable positions using RAxML v7.0.4 (Stamatakis 2006). A general time reversible (GTR) model with gamma correction for among-site rate variation and 1000 bootstrap replicates were used. SNPs were reconstructed onto the individual phylogenies using accelerated transformation parsimony, meaning that SNPs are inferred to have occurred as early as possible.

### Time-dependent phylogenetic reconstruction

Linear regression analysis of the root-to-tip distances against sampling time was performed using Path-O-Gen. Time-dependent phylogenetic reconstructions and calculations of evolutionary rates were undertaken using BEAST v1.7 (Drummond et al. 2012). The evolutionary rates of ST37 (this study) and ST578 (Sánchez-Busó et al. 2014) were used to infer the approximate length of time it would have taken the diversity in the STs 1, 23, 47, and 62 lineages to arise. Further details are provided in the Supplemental Methods.

### Gene content analysis

De novo assemblies were generated using an in-house Sanger Institute pipeline that uses Velvet (Zerbino and Birney 2008), SSPACE (Boetzer et al. 2011), and GapFiller (Boetzer and Pirovano 2012). Prodigal software was used to predict genes in the assemblies, which were then clustered into orthologous groups using BLAST+ (Blastp) and the micropan R package (Snipen and Liland 2015). Custom Python scripts were used to identify "accessory" genes present in STs 1, 23, 37, 47, and 62, but not in other STs.

### Identification of core genes under positive selection in the five STs

The branch-site model in CodeML was used to test whether any core genes (i.e., genes found in every isolate in the collection) had been subjected to positive selection on the branches leading to each of the five disease-associated STs. Further details are provided in the Supplemental Methods.

### Identification of core genes with high nucleotide similarity in the five STs

The core genome of the ST representatives, with the exclusion of the distantly related STs (ST336, ST154, and ST707), was defined using Roary (Page et al. 2015). For each core gene, a nucleotide alignment was generated using one representative from each of the five STs—Paris/ST1 (Cazalet et al. 2004), EUL00011/ST23_3, EUL00132/ST37_69, Lorraine/ST47 (Gomez-Valero et al. 2011), H043540106/ST62_2—and excluding all other ST representatives. The nucleotide diversity ($\pi$) value, a measurement first described by Nei and Li (1979), was calculated for each of these alignments using the R package, "pegas," and custom Python scripts. To test whether each core gene possessed significantly higher nucleotide similarity (or lower diversity) in the five major disease-associated STs than would be expected, these values were compared to those derived from testing all possible combinations of five STs from the set of species representatives, after adjusting for phylogenetic distance. Further details are provided in Supplemental Methods.

### Prediction of recombination donors

Each predicted recombined region was used as a query sequence in BLASTn to identify matches among the de novo assemblies of all 364 isolates used in this study. BLAST hits with a *P*-value of $<1 \times 10^{-5}$ and >75% of the length of the recombined region were recorded.

## Data access

Raw sequence reads from this study have been submitted to the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) under accession numbers ERP002503, ERP003631, and ERP010118. The assembled genomes used as references for STs 23 (EUL00011/ST23_3), 37 (EUL00132/ST37_69), and 62 (H043540106/ST62_2) have also been submitted to ENA under accession numbers FJBI01000001–FJBI01000031, FJFB01000001–FJFB01000024, and FJLN01000001–FJLN01000039, respectively. The scripts used for the diversity analyses are in the Supplemental Material and are available at https://github.com/sophiadavid1/diversity_analysis.

## Acknowledgments

Work in the C.B. laboratory is financed by the Institut Pasteur, the Institut Carnot-Pasteur MI, the French Region Ile de France (DIM Malinf), the Agence Nationale de la Recherche (ANR) Grant No. ANR-10-LABX-62-IBEID, the ANR-10-PATH-004 project, in the frame of ERA-Net PathoGenoMics, and the Fondation pour la Recherche Médicale (FRM) Grant No. DEQ20120323697. We thank M. Tichit and M. Hunt for their technical support. The Plate-forme Génomique is a member of "France Génomique" consortium (ANR10-INBS-09-08). Work at the Sanger Institute is funded by The Wellcome Trust Grant No. 098051.

*Author contributions:* S.D., C.R., and J.P., C.B., contributed equally to this work. M.M., C.G., S.J., and T.G.H. contributed to sample collection and epidemiological analyses; M.M., C.G., L.G.-V., Ch.B., and L.M. to DNA extraction, library preparation, and sequencing; and C.R., S.D., P.G., L.G.-V., A.U., P.L., J.L., and S.R.H. to data analyses and interpretation. The manuscript was written by S.D., C.R., J.P., and C.B. with input from coauthors. The project was conceived, planned, and supervised by T.G.H., J.P., and C.B.

## References

Beauté J, Zucs P, de Jong B, European Legionnaires' Disease Surveillance Network. 2013. Legionnaires' disease in Europe, 2009–2010. *Euro Surveill* **18:** 20417.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* **57:** 289–300.

Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome Biol* **13:** R56.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27:** 578–579.

Bowers JR, Kitchel B, Driebe EM, MacCannell DR, Roe C, Lemmer D, de Man T, Rasheed JK, Engelthaler DM, Keim P, et al. 2015. Genomic analysis of the emergence and rapid global dissemination of the clonal group 258 *Klebsiella pneumoniae* pandemic. *PLoS One* **10:** e0133727.

Brüggemann H, Hagman A, Jules M, Sismeiro O, Dillies MA, Gouyette C, Kunst F, Steinert M, Heuner K, Coppée JY, et al. 2006. Virulence strategies for infecting phagocytes deduced from the *in vivo* transcriptional program of *Legionella pneumophila*. *Cell Microbiol* **8:** 1228–1240.

Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, Reacher M, Haworth CS, Curran MD, Harris SR, et al. 2013. Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. *Lancet* **381:** 1551–1560.

Cassier P, Campese C, Le Strat Y, Che D, Ginevra C, Etienne J, Jarraud S. 2015. Epidemiologic characteristics associated with ST23 clones compared to ST1 and ST47 clones of Legionnaires disease cases in France. *New Microbes New Infect* **3:** 29–33.

Cazalet C, Rusniok C, Brüggemann H, Zidane N, Magnier A, Ma L, Tichit M, Jarraud S, Bouchier C, Vandenesch F, et al. 2004. Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. *Nat Genet* **36:** 1165–1173.

Cazalet C, Jarraud S, Ghavi-Helm Y, Kunst F, Glaser P, Etienne J, Buchrieser C. 2008. Multigenome analysis identifies a worldwide distributed epidemic *Legionella pneumophila* clone that emerged within a highly diverse species. *Genome Res* **18:** 431–441.

Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, Pessia A, Aanensen DM, Mather AE, Page AJ, et al. 2014. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet* **46:** 305–309.

Cirillo SL, Lum J, Cirillo JD. 2000. Identification of novel loci involved in entry by *Legionella pneumophila*. *Microbiology* **146:** 1345–1359.

Correia AM, Ferreira JS, Borges V, Nunes A, Gomes B, Capucho R, Gonçalves J, Antunes DM, Almeida S, Mendes A, et al. 2016. Probable person-to-

person transmission of Legionnaires' disease. *N Engl J Med* **374:** 497–498.
Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2014. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43:** e15.
Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29:** 1969–1973.
Edelstein PH, Metlay JP. 2009. *Legionella pneumophila* goes clonal—Paris and Lorraine strain-specific risk factors. *Clin Infect Dis* **49:** 192–194.
Euser SM, Bruin JP, Brandsema P, Reijnen L, Boers SA, Den Boer JW. 2013. *Legionella* prevention in the Netherlands: an evaluation using genotype distribution. *Eur J Clin Microbiol Infect Dis* **32:** 1017–1022.
Fields BS, Benson RF, Besser RE. 2002. *Legionella* and Legionnaires' disease: 25 years of investigation. *Clin Microbiol Rev* **15:** 506–526.
Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J, Lipsitch M, Fortune SM. 2013. *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat Genet* **45:** 784–790.
Gomez-Valero L, Rusniok C, Jarraud S, Vacherie B, Rouy Z, Barbe V, Medigue C, Etienne J, Buchrieser C. 2011. Extensive recombination events and horizontal gene transfer shaped the *Legionella pneumophila* genomes. *BMC Genomics* **12:** 536.
Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, et al. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327:** 469–474.
Harrison TG, Afshar B, Doshi N, Fry NK, Lee JV. 2009. 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* **28:** 781–791.
Holt KE, Baker S, Weill FX, Holmes EC, Kitchen A, Yu J, Sangal V, Brown DJ, Coia JE, Kim DW, et al. 2012. *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet* **44:** 1056–1059.
Lechat P, Souche E, Mszer I. 2011. SynTView: a dynamic genome browser for microbial synteny and polymorphism information developed at the Institut Pasteur. In *Proceedings JOBIM 2011*, pp. 135–136, Paris, France.
Li L, Mendis N, Trigui H, Faucher SP. 2015. Transcriptomic changes of *Legionella pneumophila* in water. *BMC Genomics* **16:** 637.
Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J. 2012. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res* **40:** e6.
Marvig RL, Johansen HK, Molin S, Jelsbak L. 2013. Genome analysis of a transmissible lineage of *Pseudomonas aeruginosa* reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. *PLoS Genet* **9:** e1003741.
Mentasti M, Fry NK, Afshar B, Palepou-Foxley C, Naik FC, Harrison TG. 2012. Application of *Legionella pneumophila*-specific quantitative real-time PCR combined with direct amplification and sequence-based typing in the diagnosis and epidemiological investigation of Legionnaires' disease. *Eur J Clin Microbiol Infect Dis* **31:** 2017–2028.
Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Feldkamp M, Kusecek B, Vogler AJ, Li Y, et al. 2010. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet* **42:** 1140–1143.
Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, et al. 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477:** 462–465.
Naylor J, Cianciotto NP. 2004. Cytochrome *c* maturation proteins are critical for in vivo growth of *Legionella pneumophila*. *FEMS Microbiol Lett* **241:** 249–256.
Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci* **76:** 5269–5273.
Newton HJ, Ang DK, van Driel IR, Hartland EL. 2010. Molecular pathogenesis of infections caused by *Legionella pneumophila*. *Clin Microbiol Rev* **23:** 274–298.
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.
Phin N, Parry-Ford F, Harrison T, Stagg HR, Zhang N, Kumar K, Lortholary O, Zumla A, Abubakar I. 2014. Epidemiology and clinical management of Legionnaires' disease. *Lancet Infect Dis* **14:** 1011–1021.
Rowbotham TJ. 1980. Preliminary report on the pathogenicity of *Legionella pneumophila* for freshwater and soil amoebae. *J Clin Pathol* **33:** 1179–1183.
Rowbotham TJ. 1998. Isolation of *Legionella pneumophila* serogroup 1 from human feces with use of amebic cocultures. *Clin Infect Dis* **26:** 502–503.
Sánchez-Busó L, Comas I, Jorques G, González-Candelas F. 2014. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. *Nat Genet* **46:** 1205–1211.
Sauer JD, Bachman MA, Swanson MS. 2005. The phagosomal transporter A couples threonine acquisition to differentiation and replication of *Legionella pneumophila* in macrophages. *Proc Natl Acad Sci* **102:** 9924–9929.
Snipen L, Liland KH. 2015. micropan: an R-package for microbial pan-genomics. *BMC Bioinformatics* **16:** 79.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688–2690.
Uhlemann AC, Dordel J, Knox JR, Raven KE, Parkhill J, Holden MT, Peacock SJ, Lowy FD. 2014. Molecular tracing of the emergence, diversification, and transmission of *S. aureus* sequence type 8 in a New York community. *Proc Natl Acad Sci* **111:** 6738–6743.
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.
Vekens E, Soetens O, De Mendonça R, Echahidi F, Roisin S, Deplano A, Eeckhout L, Achtergael W, Piérard D, Denis O, et al. 2012. Sequence-based typing of *Legionella pneumophila* serogroup 1 clinical isolates from Belgium between 2000 and 2010. *Euro Surveill* **17:** 20302.
Viswanathan VK, Kurtz S, Pedersen LL, Abu-Kwaik Y, Krcmarik K, Mody S, Cianciotto NP. 2002. The cytochrome *c* maturation locus of *Legionella pneumophila* promotes iron assimilation and intracellular infection and contains a strain-specific insertion sequence element. *Infect Immun* **70:** 1842–1852.
Weissenmayer BA, Prendergast JG, Lohan AJ, Loftus BJ. 2011. Sequencing illustrates the transcriptional response of *Legionella pneumophila* during infection and identifies seventy novel small non-coding RNAs. *PLoS One* **6:** e17570.
Yu VL, Plouffe JF, Pastoris MC, Stout JE, Schousboe M, Widmer A, Summersgill J, File T, Heath CM, Paterson DL, et al. 2002. Distribution of *Legionella* species and serogroups isolated by culture in patients with sporadic community-acquired legionellosis: an international collaborative survey. *J Infect Dis* **186:** 127–128.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18:** 821–829.

*Received May 8, 2016; accepted in revised form September 16, 2016.*

# Multiple major disease-associated clones of *Legionella pneumophila* have emerged recently and independently

Sophia David, Christophe Rusniok, Massimo Mentasti, et al.

*Genome Res.* 2016 26: 1555-1564 originally published online September 23, 2016
Access the most recent version at doi:10.1101/gr.209536.116