3. Recent emergence of five major disease-associated STs

Declaration of work contributions

Timothy Harrison, Julian Parkhill and Carmen Buchrieser initiated this study. Collaborators at PHE (London, UK) and the National Reference Center of *Legionella* (Lyon, France) performed culture and DNA extraction of all newly sequenced isolates. The core sequencing facilities at the WTSI and Institut Pasteur performed library preparation and sequencing. Collaborators at the Institut Pasteur performed the gene content analysis and the visualisation of SNP distributions using SynTView software. I conducted the remaining bioinformatics analyses.

Publication

The following work has been published:

David, S., Rusniok, C., Mentasti, M., Gomez-Valero, L., Harris, S. R., Lechat, P., Lees, J., Ginevra, C., Glaser, P., Ma, L., Bouchier, C., Underwood, A., Jarraud, S., Harrison, T. G., Parkhill, J. & Buchrieser, C. Multiple major disease-associated clones of *Legionella pneumophila* have emerged recently and independently. *Genome Research* **26**, 1555-1564 (2016).

3.1 Introduction

L. pneumophila is an environmental bacterium that survives in natural aquatic and soil habitats as well as modern, man-made water systems (Fields *et al.*, 2002). Humans are primarily infected with *L. pneumophila* via the inhalation of aerosols containing the bacteria (Muder *et al.*, 1986) and most usually from man-made environmental sources. Human infection is thought to be "accidental" and an evolutionary dead-end for the bacteria.

Investigation of clinical *L. pneumophila* isolates by various typing methods has revealed that some types cause human infection far more commonly than others. For example, while there are 16 serogroups (sg) currently described, sg 1 was responsible for 83% culture-confirmed Legionnaires' disease cases attributed to *L. pneumophila* in Europe in 2013 (ECDC, 2015). Furthermore, as of 8 July 2016, 2191 STs have been reported to the ESGLI SBT database but a relatively small proportion has been commonly associated with disease. Indeed, analysis of all clinical isolates submitted to the ESGLI SBT database (*n*=6116) prior to April 2015 found that isolates belonging to just five STs (1, 23, 37, 47, 62) accounted for over 40% clinical isolates submitted to the SBT database from Europe (**Figure 3.1A**). There is no evidence that the high proportion of isolates found in clinical samples belonging to these five STs is a result of laboratory artefacts 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 from patients (*n*=99) confirmed a similar distribution of these STs among culture-proven and culture-negative but NP-SBT positive patients in France.

One of these five STs, ST1, has been described as a leading cause of Legionnaires' disease from numerous countries worldwide including Canada (Tijet *et al.*, 2010), Japan (Amemura-Maekawa *et al.*, 2010), France (Ginevra *et al.*, 2012), Belgium (Vekens *et al.*, 2012) and Israel (Moran-Gilad *et al.*, 2014). ST1 isolates have been reported to the SBT database from all continents that actively report *L. pneumophila* isolates, and comprise 11.0% of the total including 19.1% of isolates from North America (**Figure 3.1A**). However, several studies have found that ST1 isolates are also found commonly in environmental samples (Kozak-Muiznieks *et al.*, 2014; Amemura *et al.*, 2012). A study of 443 isolates (including 167 clinical and 276 environmental isolates) obtained between 2000 and 2008 in England and Wales showed that ST1 isolates are more prevalent in environmental samples than clinical samples (Harrison *et al.*, 2009), a finding that was mirrored in the analysis of clinical (*n*=6116) and environmental (*n*=2826) isolates submitted to the SBT database (**Figure 3.1B**).

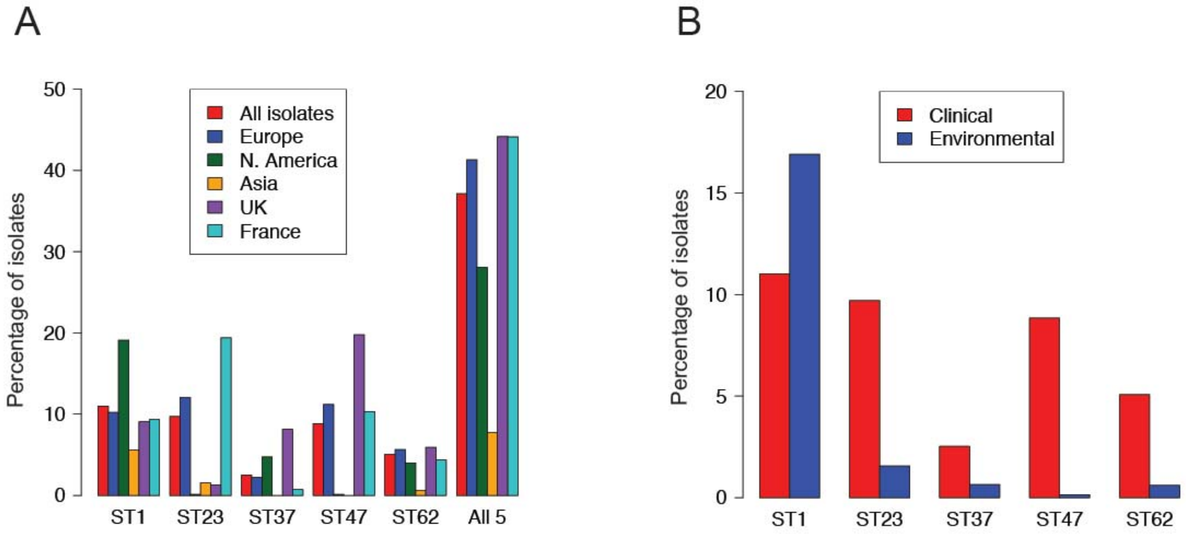


Figure 3.1. Geographical distribution of STs and their prevalence in clinical and environmental samples. A) The percentage of clinical isolates submitted to the ESGLI SBT database from different geographical regions that belong to STs 1, 23, 37, 47 and 62. These data are based on a total of 6116 epidemiologically unrelated clinical isolates (i.e. including only one representative isolate from clusters and outbreaks) submitted to the database prior to April 2015. Of these, 4785 were detected in Europe (including 541 in the UK and 2313 in France), 801 in North America and 323 in Asia. These particular regions were chosen because the numbers that were submitted were deemed sufficient for a comparison. B) The percentage of isolates submitted to the SBT database that belong to one of the five major disease-associated STs that are of clinical or environmental origin. These data are based on a total of 6116 and 2826 epidemiologically unrelated clinical and environmental isolates, respectively, that were submitted to the SBT database prior to April 2015.

Of the remaining four disease-associated STs (23, 37, 47 and 62), none have the global distribution observed for ST1 isolates. Nonetheless, STs 23, 37 and 62 have large distributions and isolates have been reported to the SBT database from Europe, North America and Asia, although most commonly from Europe. By contrast, ST47 isolates have been almost exclusively isolated from Western European countries including England and Wales (Harrison *et al.*, 2009), France (Ginevra *et al.*, 2008), Belgium (Vekens *et al.*, 2012) and the Netherlands (Euser *et al.*, 2013). A small number of non-travel-associated cases of ST47 have also been reported from Canada (Tijet *et al.*, 2010). Furthermore, in contrast to ST1, the study by Harrison *et al.* (2009) revealed that while three of the five major disease-associated STs (37, 47 and 62) accounted for 11.4%, 25.7% and 9.0% of clinical isolates in the collection from England and Wales, respectively, they comprised only 0.7%, 0.4% and 0% of environmental isolates. This highly uneven distribution of ST37, ST47 and ST62 in clinical and environmental isolates was also found in the analysis of isolates submitted to the SBT database, and a similar distribution was also found for ST23 isolates (**Figure 3.1B**).

Despite differences in their geographical and environmental distributions, the five STs (1, 23, 37, 47 and 62) are all linked by their predominance in human infections. The aim of this chapter is to explore the genomic diversity of these five STs in the context of the *L. pneumophila* species diversity. It seeks to understand their emergence as important human pathogens and explore whether there are signals of convergent evolution that could explain their increased disease association.

3.2 Materials & Methods

3.2.1 Bacterial isolates

A total of 364 *L. pneumophila* isolates, of which 35 are previously published and 329 are newly sequenced, were used in this thesis chapter. The previously published isolates include those belonging to 32 STs (**Appendix Table 1**), which were selected as representatives of the known species diversity (Underwood *et al.*, 2013). 337 isolates, including 327 that are newly sequenced, belong to one of the five major disease-associated STs and include 71 ST1 (or "ST1-derived"), 37 ST23, 72 ST37, 122 ST47 and 35 ST62 isolates (**Appendix Tables 1 & 2**). ST1-derived isolates belong to other STs that are nested within, and thus evolved from, ST1 isolates. All newly sequenced isolates are from the culture collections at PHE, UK or the National Reference Center of *Legionella*, France. Culture and DNA extraction of all isolates was performed as described in *Chapter 2 (Materials & Methods)*.

3.2.2 Whole genome sequencing

Paired-end sequencing was performed on 248 isolates at the WTSI using the Illumina HiSeq platform and a read length of 100 bases, as described in *Chapter 2 (Materials & Methods)*. Paired-end sequencing was also performed at the WTSI on one ST1 isolate, OLDA1, using the Illumina MiSeq platform and a read length of 150 bases. A further 80 isolates were sequenced using the Illumina HiSeq platform at the Institut Pasteur. All sequence reads were deposited in the European Nucleotide Archive (ENA) under the study accession numbers ERP002503, ERP003631 and ERP010118. Individual accession numbers for each sample are provided in **Appendix Table 2.**

3.2.3 Mapping of sequence reads and phylogenetic analysis

Sequence reads from the 32 isolates representing the species diversity were mapped to the Corby reference genome (Gloeckner *et al.*, 2008) to analyse the species-wide population structure. Isolates belonging to each of the five STs were also mapped to a reference genome of the same ST to analyse the population structure of each ST at a higher resolution. Complete reference genomes, known as Paris and Lorraine, were available for STs 1 (Cazalet *et al.*, 2004) and 47 (Gomez-Valero *et al.*, 2011), and *de novo* assemblies were used for STs 23 (EUL 11), 37 (EUL 132) and 62 (H043540106). All mapping was performed using SMALT v0.7.4 (available at: http://www.sanger.ac.uk/science/tools/smalt-0). An in-house pipeline at the WTSI was used to call bases and identify SNPs as described in *Chapter 2 (Materials & Methods)*. Recombination detection was performed using Gubbins (Croucher *et al.*, 2015) and BRATNextGen (Marttinen *et al.*, 2012). Phylogenetic analyses were performed as described in *Chapter 2 (Materials & Methods).*

3.2.4 Time-dependent phylogenetic reconstruction

TempEst software (formerly known as Path-O-Gen) (Rambaut *et al.*, 2016) was used to perform linear regression analysis of the root-to-tip distances against the sampling date in each phylogenetic tree belonging to the five major disease-associated STs. Time-dependent phylogenetic reconstruction of the ST37 lineage was also attempted using BEAST v1.7 (Drummond *et al.*, 2012). After identifying and removing any SNPs that were imported *via* recombination, a SNP alignment together with the isolation dates of all ST37 samples, were used as input. A variety of population size models were tested including constant, exponential and Bayesian skyline (variable) together with a variety of clock models including strict, lognormal relaxed, exponential relaxed and random. Path sampling and stepping stone sampling were used to calculate Bayes factors, allowing comparison of different models and selection of the most appropriate (Baele *et al.*, 2012). Each model was tested using three independent chains of 100 million steps, sampling every 10,000 steps and discarding the first 10 million steps as burn-in. The convergence of the runs and effective sample sizes were verified using Tracer v1.5 (available at: http://tree.bio.ed.ac.uk/software/tracer). The results from the three independent runs were combined using LogCombiner and a maximum clade credibility (MCC) tree was produced using TreeAnnotator. Both programmes are available in the BEAST package (Drummond *et al.*, 2012).

3.2.5 Estimation of the age of the ST1, ST23, ST47 and ST62 lineages

The roots of the ST1, ST23, ST47 and ST62 maximum likelihood trees (constructed after recombination removal) were established using outgroup isolates. Each tree was subsequently constructed without the outgroup and rooted appropriately. Using the evolutionary rates estimated for the ST37 lineage and the previously published ST578 lineage (Sanchez-Buso *et al.*, 2014) with BEAST, the approximate length of time that it would have taken for the diversity to be acquired in each of the four lineages was estimated. Firstly, accelerated transformation parsimony was used to scale the branches of each phylogenetic tree by the number of SNPs that had occurred (see *Chapter 2 (Materials & Methods)*). The number of SNPs on each branch was then scaled up by the proportion of the genome that had been removed due to recombination, in order to account for any additional SNPs that may have occurred by *de novo* mutation on top of the recombined regions. Using the two estimated substitution rates and the known sampling dates of all isolates, each root-to-tip distance in the phylogenetic tree was used to calculate the length of time it would have taken for each isolate to have evolved from the common ancestor of the lineage. An estimated age of the tree root was inferred from the mean of these values.

3.2.6 Gene content analysis

De novo assemblies were generated for all isolates as described in *Chapter 2 (Materials & Methods)*. Genes were identified in the assemblies using the Prodigal gene finder software and clustered into orthologous groups using BLAST+ (BLASTp) and the micropan R package (Snipen & Liland, 2015). Genes that are present in the five major disease-associated STs, but not in other STs, were identified using custom Python scripts. This analysis was performed by Christophe Rusniok (Institut Pasteur).

3.2.7 Searching for evidence of positive selection using CodeML

Genes that were present in all 364 isolates were determined using Roary (Page *et al.*, 2015). For each core gene, a nucleotide alignment comprising sequences from all 364 isolates was used to generate a maximum likelihood tree using RAxML (Stamatakis, 2006). Each core gene was tested individually using the branch-site model in CodeML (Yang, 2007) to determine whether any specific regions had been subjected to positive selection on the branches of the phylogenetic tree leading to each of five disease-associated STs. Each gene was tested five times, each time specifying one of the five branches, and each test involved comparison of a null model (specifying that no difference in dN/dS exists between the selected branch and the remaining branches in the tree) and an alternative model (specifying that the gene contains regions that underwent positive selection on the selected branch). The log likelihood values derived from the two models were compared to determine the best-fitting model.

3.2.8 Identification of genes with high nucleotide similarity in the five STs

Genes that were present in all 364 isolates excluding the distantly related STs (ST336, ST154 and ST707) were determined using Roary (Page *et al.*, 2015). A nucleotide alignment was generated for each core gene using one representative isolate from STs 1, 23, 37, 47 and 62, which were Paris, EUL 11, EUL 132, Lorraine, H043540106, respectively. All other isolates were excluded from this alignment. An R package, “pegas”, and custom Python scripts were used to determine the nucleotide diversity (pi) value (Nei & Li, 1979) for each of these core gene alignments. To test whether the nucleotide diversity between the five major disease-associated STs was significantly lower in any of the core genes than expected, given the overall phylogenetic relatedness of the five STs and the overall conservation of each gene across the species, nucleotide diversity values were calculated for all possible combinations of any five STs within the set of species representatives. The distantly related STs (ST336, ST154 and ST707) were excluded from these calculations as well as ST5 and ST152, which are nested within the ST1 lineage in the phylogenetic tree, and Philadelphia/ST36, Alcoy/ST578 and ST42, which belong to strains commonly associated with disease. The total number of combinations using the remaining 24 STs (including the five major disease-associated STs) was 42,504. For each combination of five STs, the median nucleotide diversity across all core genes was calculated. The nucleotide diversity values of individual genes were then divided by the median values, thereby adjusting for the phylogenetic distance between the particular combinations of five isolates. For each core gene, these adjusted nucleotide diversity values (*n*=42,504) were used together with the nucleotide diversity value of the five major-disease associated STs to derive a p-value. The Benjamini-Hochberg method, implemented in R, was used to correct for multiple testing.

3.2.9 Identification of recombination donors

Predicted recombination regions were used as query sequences in BLASTn to determine possible matches amongst the *de novo* assemblies of 364 isolates, which include the 32 species representatives. Matches with a p-value of <1e-05 and >75% length of the query sequence were recorded.

3.3 Results

3.3.1 Independent emergence of the five STs

A phylogenetic tree was constructed comprising representative isolates belonging to each of the five disease-associated STs (1, 23, 37, 47 and 62) and a further 27 isolates belonging to different STs of *L. pneumophila* (**Appendix Table 1 & Figure 3.2**). This was generated by mapping sequence reads to the Corby reference genome (Gloeckner *et al.*, 2008) and identifying SNPs. Together, these 32 STs represented the most distantly related STs in the SBT database when they were selected for sequencing in a previous study (Underwood *et al.*, 2013). While many of the isolates from the additional 27 STs are derived from clinical samples, the STs to which they belong have mostly been implicated in human disease far less frequently than STs 1, 23, 37, 47 and 62, and thus they provide a comparative set that is used in this study. **Figure 3.2** shows that the five major disease-associated STs all belong to the *L. pneumophila pneumophila* subspecies although, with the exception of ST23 and ST62, they belong to separate major clades of the tree. This indicates that these major disease-associated STs have evolved independently from different genomic backgrounds. Nucleotide identities were also calculated between all pairs of isolates representing the five STs using the core genome. Pairwise similarities range from 97.5% to 98.7%, except between ST23 and ST62, which share 99.25% nucleotide similarity.

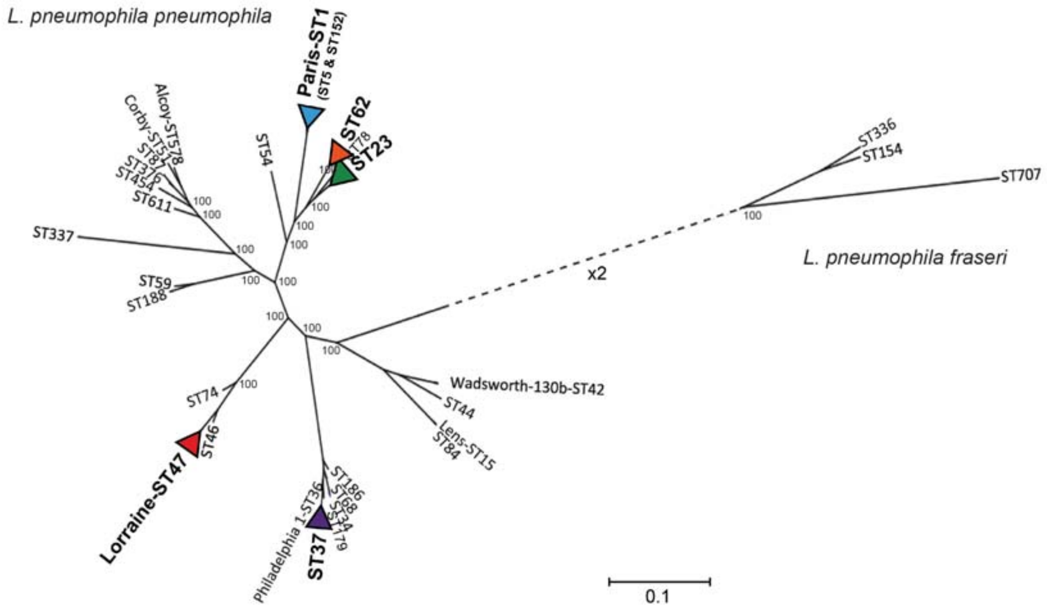


Figure 3.2. Population structure of *L. pneumophila* highlighting five major disease-associated STs of interest (previous page). A maximum likelihood tree of 32 *L. pneumophila* isolates that represent the known species diversity. The five major disease-associated STs, which are highlighted by coloured triangles, are generally found in separate major clades with the exception of ST23 and ST62 that share a more recent common ancestor. Bootstrap values, derived from 1000 re-samples, are shown for major nodes of the tree. The scale represents the number of SNPs per variable site in the genome alignment.

3.3.2 Investigation of the diversity within the five STs

To investigate the genomic diversity and evolution of each of the five major disease-associated STs, we analysed 71 ST1 (including 12 ST1-derived), 37 ST23, 72 ST37, 122 ST47 and 35 ST62 isolates (**Appendix Tables 1 & 2**). ST1-derived isolates belong to other STs that are nested within, and thus evolved from, the ST1 lineage (ST5, ST6, ST7, ST8, ST10, ST72, ST152). ST1 is a globally dispersed lineage and the 71 isolates included in this study were isolated from 14 countries over four continents (Europe, Asia, North America and Africa) between 1981 and 2011. The oldest known isolate of *L. pneumophila* (OLDA1) recovered in 1947, thirty years prior to the description of the species, was also sequenced and analysed with the ST1 collection. STs 23, 37 and 62 are most usually isolated in Europe although have also been isolated elsewhere. All sequenced isolates of these three STs were recovered in Europe between 1987 and 2012, with the exception of a small number of travel-associated isolates for which the origin is uncertain. Finally, almost all ST47 isolates have been detected in the UK, France, the Netherlands and Belgium, although a small number have also been detected in other European countries and, notably, also in Canada. The 122 ST47 isolates included in this study were recovered from the UK and France between 1994 and 2013, although some travel-associated isolates for which the origin is uncertain are also included. Furthermore, some of the sequenced isolates belonging to the five major disease-associated STs are epidemiologically related (i.e. recovered from the same cluster or outbreak) (see **Appendix Tables 1 & 2**).

Sequence reads from these isolates were mapped to a reference genome of the same ST and the total number of SNPs in each of the lineages was determined (**Table 3.1**).

Remarkably, just 186 SNPs were found between the 122 ST47 isolates and the maximum difference between any pair of ST47 isolates is only 19 SNPs. 21 isolates recovered from geographically distinct regions of the UK between 2003 and 2012 possess no detectable SNPs and another 17 isolates, which were recovered either in the UK or from travel-associated cases (i.e. with an unknown origin), possess just one difference from these. No SNPs are homoplasic and visualisation of the SNPs using SynTView (Lechat *et al.*, 2011), performed by Pierre Lechat (Institut Pasteur), also shows that they are evenly spread across the genome (**Figure 3.3A**). Gubbins detected no recombination in the ST47 lineage, an observation that is concordant with the low number of SNPs detected and their even distribution.

Table 3.1. Reference genomes used for mapping isolates belonging to each of the five STs and the number of SNPs detected within each lineage.

ST	Number of isolates	Mapping reference	Total number of SNPs	Maximum number of pairwise SNP differences
ST1 (and ST1-derived)	71	Paris (complete genome)	48,655	15,227
ST23	37	EUL 11 (*de novo* assembly)	26,945	12,964
ST37	72	EUL 132 (*de novo* assembly)	14,829	13,776
ST47	122	Lorraine (complete genome)	186	19
ST62	35	H043540106 (*de novo* assembly)	33,200	12,842

In contrast to ST47, the total numbers of SNPs detected within STs 1, 23, 37 and 62 were substantially higher (**Table 3.1**). The highest number of SNPs was observed in the globally dispersed ST1 lineage, which has 48,655 SNPs between 71 isolates, and a maximum difference between any pair of 15,227. The second highest number of SNPs was observed in the ST62 lineage (*n*=33,200), followed by ST23 (*n*=26,945) and ST37 (*n*=14,829). However, visualisation of the SNP distributions in these four lineages using SynTView (**Figure 3.3B-E**) showed that many of the SNPs are found in very close proximity to others, suggesting the occurrence of recombination events.

Indeed, Gubbins (Croucher *et al.*, 2015) predicted that between 96.3% and 99.0% of the total SNPs detected in STs 1, 23, 37 and 62 were imported *via* recombination (**Table 3.2**). These results were confirmed by an alternative recombination detection programme, BRATNextGen (Marttinen *et al.*, 2012), which predicted over 90% of SNPs identified as recombined by Gubbins to be within horizontally exchanged regions. The mean length of each genome predicted by Gubbins to have been affected by recombination varied between 3.4% (ST23 lineage) and 12.9% (ST62 lineage). Once predicted recombined regions were removed from the alignments, the number of remaining vertically inherited SNPs in each of the four lineages was more similar to that observed between ST47 isolates, ranging from 182 (ST23) to 867 (ST1) (**Table 3.2**). Therefore, all five disease-associated lineages are characterised by a very low number of *de novo* mutations, which is in contrast to the high diversity observed across the species.

Table 3.2. Mean length of genome affected by recombination in each lineage and the percentage of total SNPs that are predicted to be within recombined regions. The remaining numbers of vertically inherited SNPs in each lineage are also shown as well as the maximum number found between any two isolates.

ST	Mean length (and %) of genome affected by recombination (bp)	% SNPs in recombined regions	Number of vertically-inherited SNPs in lineage	Maximum number of vertically-inherited SNPs between two isolates
ST1	335,382 (9.6%)	98.2	867	127
ST23	118,597 (3.4%)	99.3	182	59
ST37	144,953 (4.2%)	96.3	546	75
ST47	0 (0%)	0	186	19
ST62	447,320 (12.9%)	99.0	335	110

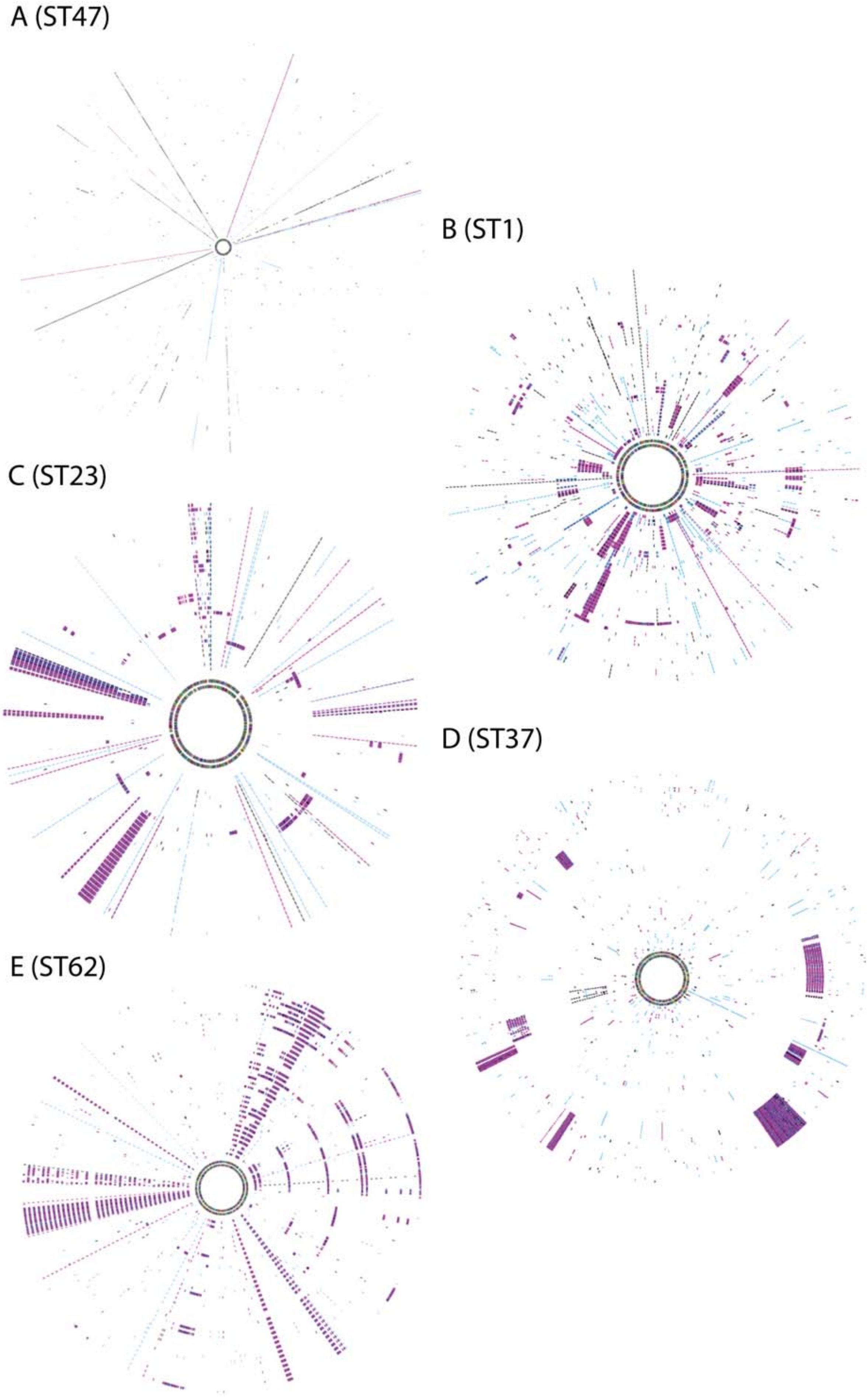

A (ST47)
B (ST1)
C (ST23)
D (ST37)
E (ST62)

Figure 3.3. Distribution of SNPs in isolates belonging to STs 47 (A), 1 (B), 23 (C), 37 (D) and 62 (E) (previous page). Each genome is shown as a concentric circle and each short line represents a SNP with respect to the reference genome. SNPs are coloured according to the type of mutation: black – intergenic; pink – synonymous; blue – non-synonymous. Recombined regions are evident in STs 1, 23, 37 and 62 as regions with a higher density of SNPs. The figures were generated using SynTView software by Pierre Lechat (Institut Pasteur).

3.3.3 Dating the emergence of the five STs

The small number of vertically inherited SNPs detected within each of the five disease-associated lineages strongly suggests that all emerged recently. We attempted to date the most recent common ancestor (MRCA) of STs 1, 23, 37, 47 and 62 by generating phylogenetic trees of each lineage and performing a linear regression analysis of all root-to-tip distances in each tree against the time of sampling using TempEst (Rambaut *et al.*, 2016). A strong positive correlation between these two variables indicates the presence of a strict molecular clock (i.e. SNPs occurring at fixed intervals), and extrapolation of the trend allows the dating of the MRCA. This analysis was performed after the removal of recombinant SNPs, a process that should improve the correlation. However, in each of the five STs, there was a poor, sometimes even negative, correlation with the exception of the ST37 lineage in which the correlation was slightly higher (Pearson’s correlation coefficient = 0.23) (**Figure 3.4**). An emergence date of 1884 was estimated for the ST37 lineage, albeit under the assumption of a strict molecular clock. The lack of temporal signal in STs 1, 23, 47 and 62 prohibited us from estimating the date of the MRCA with this method.

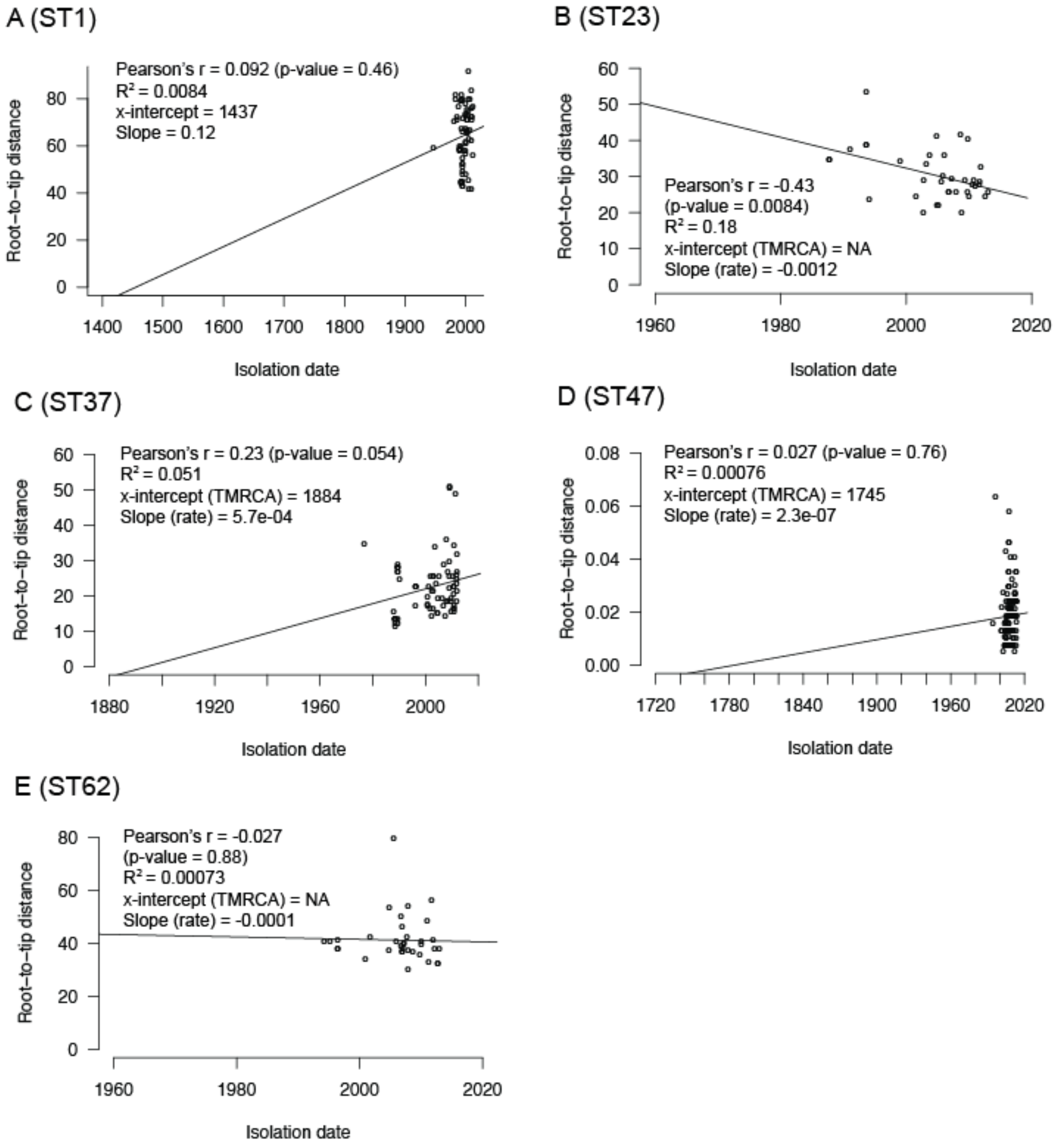


Figure 3.4. Linear regression analyses of root-to-tip distances against sampling date in each of the five STs. Each regression was performed after the removal of SNPs in recombined regions. The correlation coefficient (Pearson's r) is weak in all lineages although slightly higher in the ST37 lineage. TMRCA – time to most recent common ancestor.

Given the results of the linear regression analyses, we next attempted to date only the ST37 lineage using an alternative Bayesian coalescent method implemented with BEAST software (Drummond *et al.*, 2012). This allows a relaxed molecular clock (i.e. a substitution rate that varies between tree branches) to be incorporated into the model, which was predicted to provide a better fit to the ST37 data. Indeed, after testing and comparing a range of model parameters (see *Materials & Methods*), a model that uses an exponential relaxed substitution rate and a Bayesian skyline (variable) population size was found to converge and have the best fit to the data. The model predicted the median age of emergence of the ST37 lineage to be 1979 (95% highest posterior density (HPD) intervals: 1968 to 1985) (**Figure 3.5**).

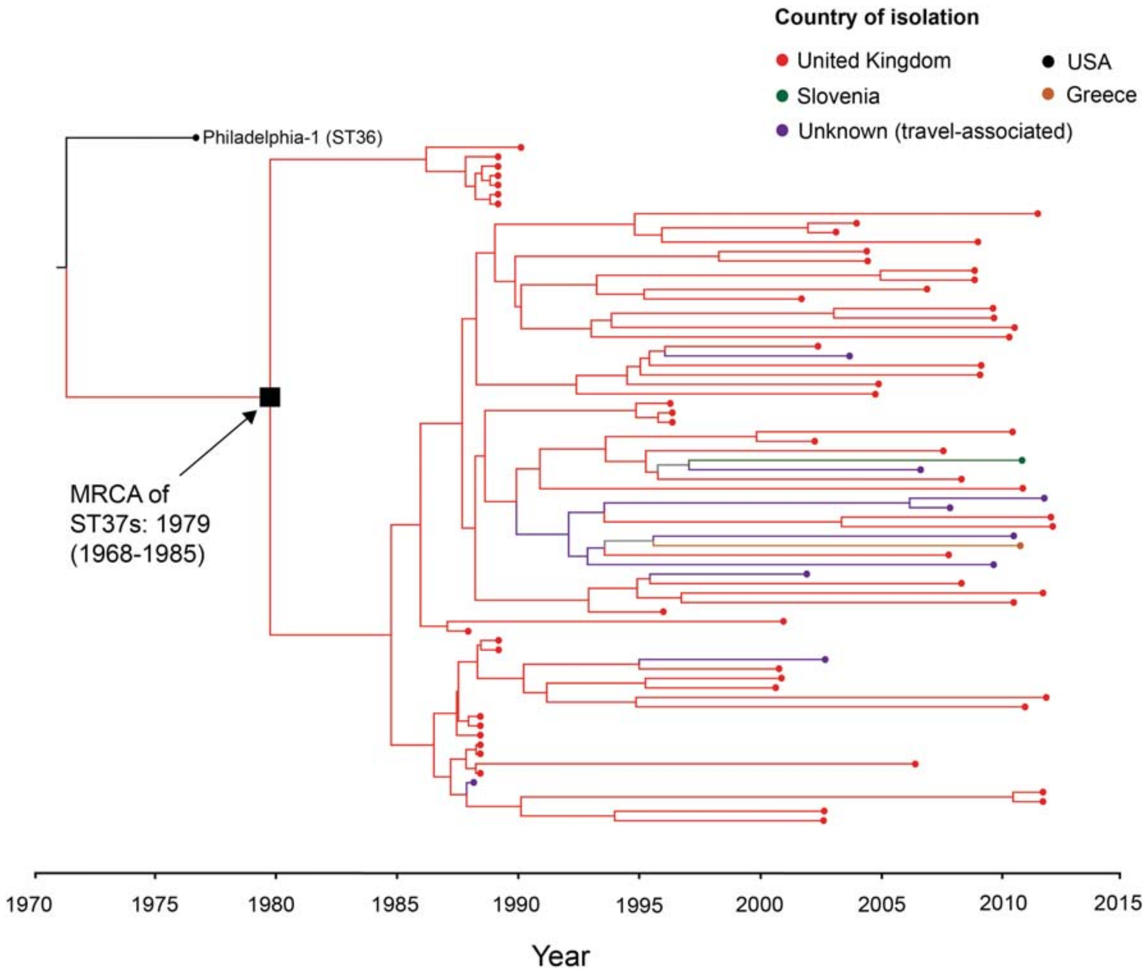


Figure 3.5. Time-dependent phylogenetic reconstruction of the ST37 lineage inferred using a Bayesian coalescent model in BEAST. The Philadelphia-1 isolate (ST36) was included in the analysis as an outgroup. The MRCA of the 72 ST37 isolates is labelled with the median estimated date and the 95% HPD intervals. Isolates are represented by circles and coloured according to the country in which they were recovered. Branches are also coloured to indicate the origin of descendant nodes.

The oldest known isolate of ST37 is from 1982 and within the time interval predicted by BEAST, which makes the dating estimations seem plausible. An evolutionary rate of 2.07x10^{-7} SNPs/site/year (95% HPD interval: 1.69x10^{-7}-2.44x10^{-7}) was estimated, which is slightly higher than the rate predicted for the *L. pneumophila* ST578 lineage (1.39x10^{-7}) (Sanchez-Buso *et al.*, 2014).

The predicted substitution rates of the ST578 and ST37 lineages were used to provide rough estimates for the length of time it would have taken for the diversity observed in STs 1, 23, 47 and 62 to have arisen. Emergence dates of 1851/1899 for ST1, 1972/1983 for ST23, 1943/1964 for ST62 and 1998/2002 for ST47 were predicted, with the two dates for each corresponding to the use of the mean rates of the ST578 and ST37 lineages, respectively. While the earliest recovered ST47 isolate was from 1994, slightly before the mean predicted emergence date, the isolation date does fall within the range estimated by the 95% HPD intervals on the substitution rates. Furthermore, even if large variations in the substitution rate exist between the five disease-associated lineages, these results clearly suggest that all five STs have emerged recently, and four within the last century.

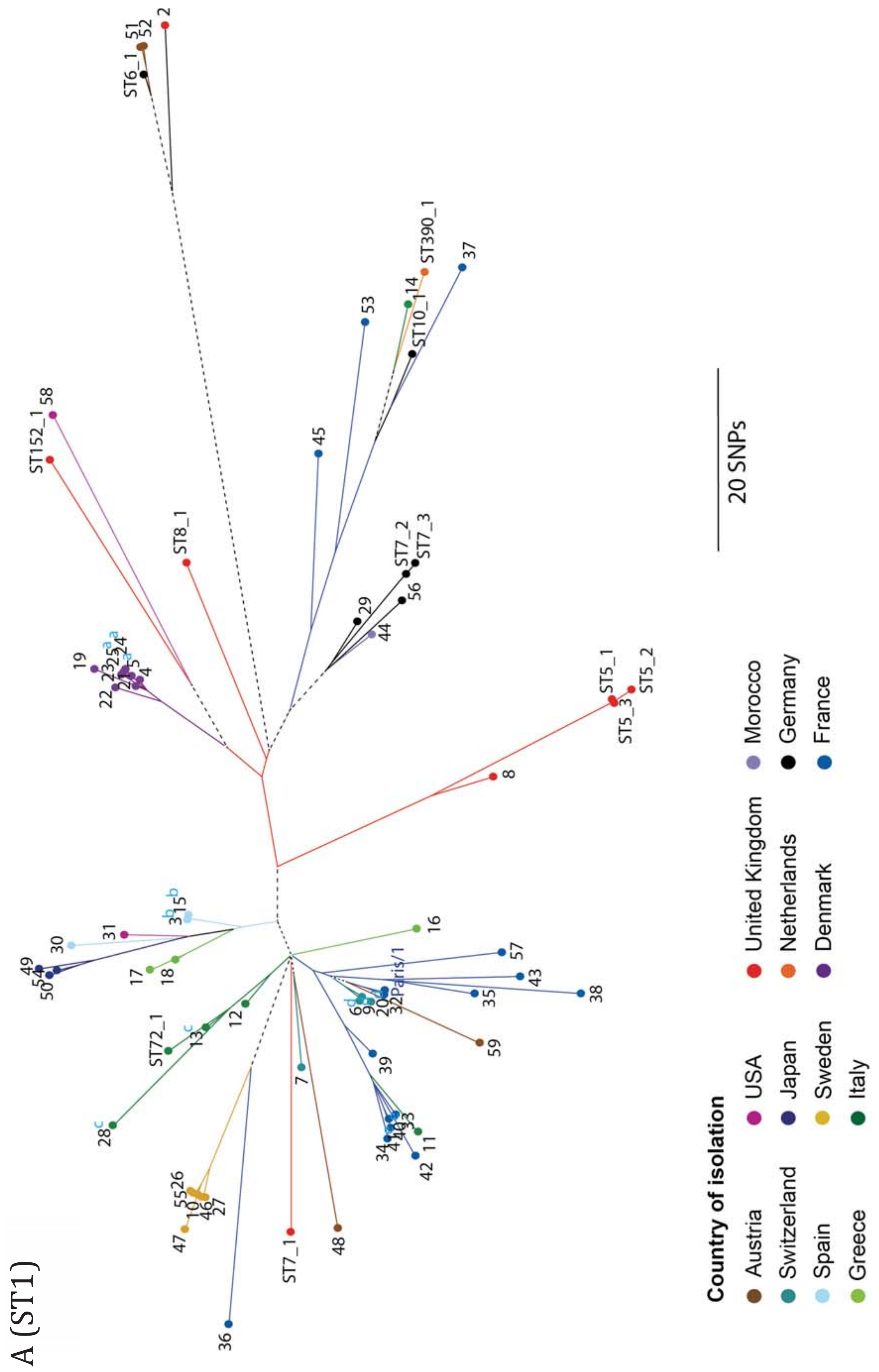
Another interesting observation is that the geographical distribution of the five major disease-associated STs correlates with the estimated ages of their emergence. ST1 has a worldwide distribution and is estimated to have emerged first, STs 23, 37 and 62 are mostly found in Europe but occasionally seen elsewhere and have emergence dates more recent than ST1, and ST47 has mostly been recovered in just a few countries in North West Europe, and is predicted to have emerged most recently.

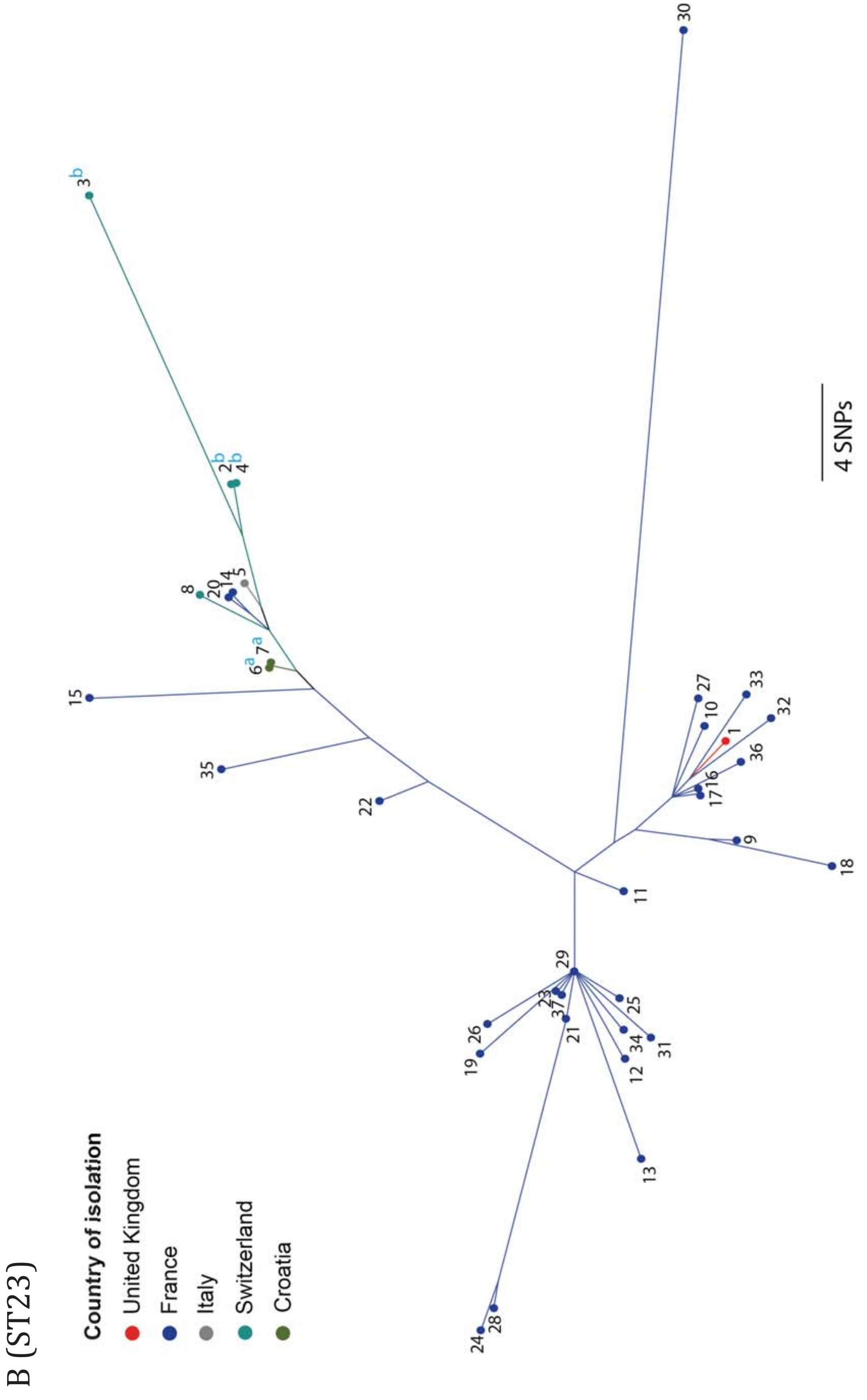
3.3.4 Analysis of the spread of the disease-associated STs

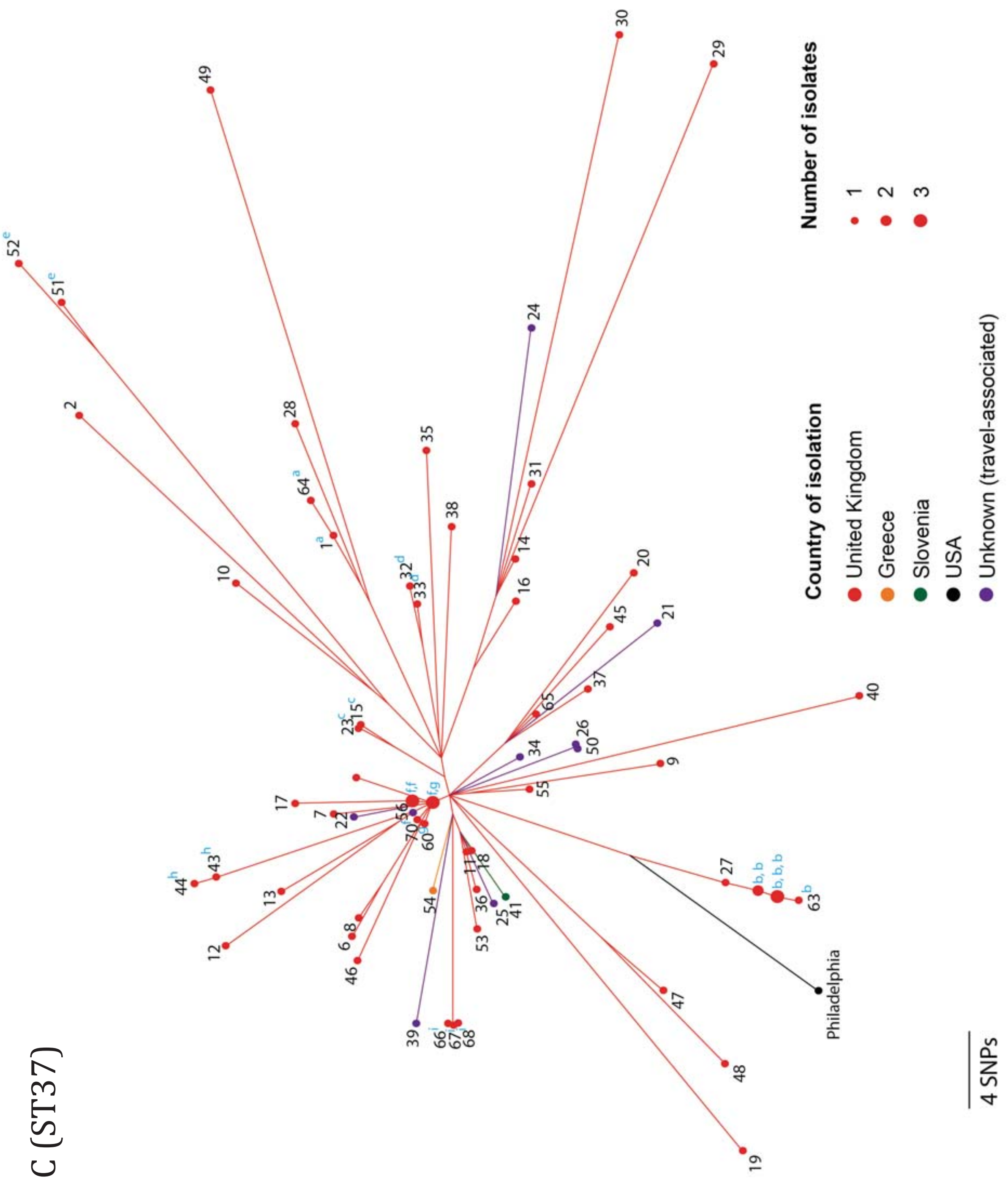
The phylogeographic structure of each of the five disease-associated STs was next analysed using phylogenetic trees constructed using only vertically inherited SNPs. Interestingly, a maximum likelihood tree of the globally dispersed ST1 lineage shows that isolates recovered in the same country do not always cluster together while isolates recovered from different continents sometimes cluster very closely (e.g. ST1_30 from Spain, ST1_31 from USA and ST1_49 from Japan) (**Figure 3.6A**). This observation suggests that ST1 isolates have been spread multiple times between different countries and even across continents.

While the majority of our ST23 isolates were recovered from France, and the majority of ST37 and ST62 isolates from the United Kingdom, the collections also include small numbers of isolates from other European countries. In concordance with the ST1 tree, phylogenetic trees of each of these lineages also show that isolates from different countries are often very closely related and sometimes more similar than isolates from the same country (**Figures 3.6B-D**).

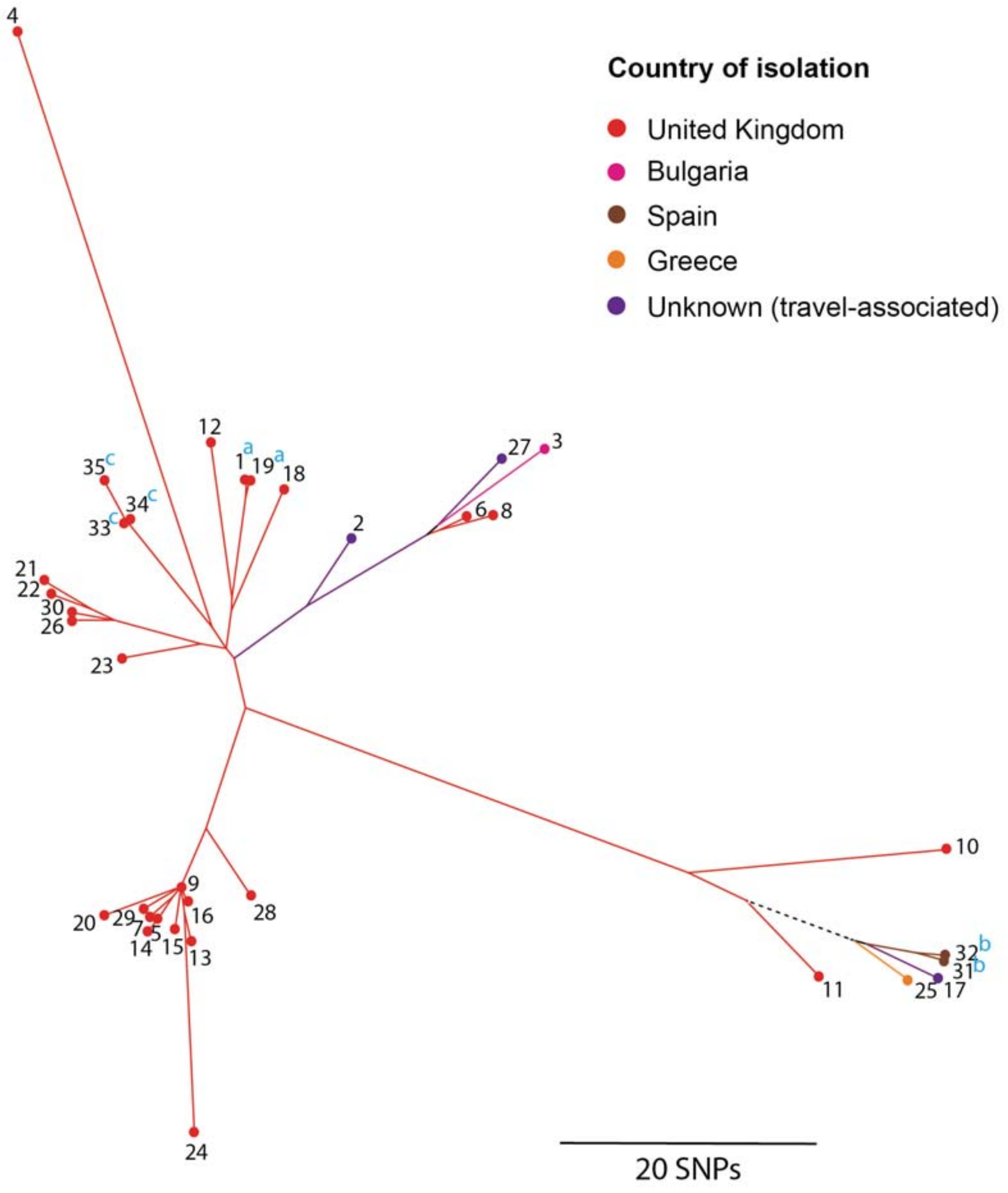
The phylogenetic tree of 122 ST47 isolates shows that isolates mostly cluster by the country of origin (UK or France), although the two clusters are separated by only two SNPs and with low bootstrap support due to the low number of SNPs involved (**Figure 3.6E**). However, there are isolates recovered from the UK nested between French isolates, which suggests that several transmission events between the two countries have occurred. The 21 UK isolates that possess no SNPs, together with the numerous more that are just one or two SNPs different, were also recovered from distant areas of the UK suggesting the occurrence of frequent spreading within the UK.


A (ST1)
Country of isolation
Austria
Switzerland
Spain
Greece
USA
Japan
Sweden
Italy
United Kingdom
Netherlands
Denmark
Morocco
Germany
France
20 SNPs
2
51
52
ST6_1
ST152_1
58
ST8_1
ST390_1
37
14
ST10_1
53
45
ST7_2
ST7_3
29
56
44
19
22
23
25
24
21
5
4
ST5_1
ST5_2
ST5_3
8
31
30
49
54
50
17
18
16
57
43
38
35
Paris/1
32
20
9
6
59
12
13
ST72_1
28
39
34
4
40
33
11
42
7
ST7_1
48
47
55
26
10
46
27
36
a
b
c
d


B (ST23)
Country of isolation
United Kingdom
France
Italy
Switzerland
Croatia
4 SNPs
30
3b
2b
4b
8
20
14
5
6a
7a
15
35
22
27
10
33
32
1
36
16
17
9
18
11
29
23
37
25
21
34
31
12
26
19
13
24
28


C (ST37)
1 a
2
6
7
8
9
10
11
12
13
14
15 c
16
17
18
19
20
21
22
23 c
24
25
26
27
28
29
30
31
32 d
33 d
34
35
36
37
38
39
40
41
43 h
44 h
45
46
47
48
49
50
51 e
52 e
53
54
55
56
60
63 b
64 a
65
66 i
67 i
68
70
b, b
b, b, b
f, f
f, g
Philadelphia
4 SNPs
Country of isolation
United Kingdom
Greece
Slovenia
USA
Unknown (travel-associated)
Number of isolates
1
2
3

D (ST62)


Country of isolation
● United Kingdom
● Bulgaria
● Spain
● Greece
● Unknown (travel-associated)
4
12
1a
19a
18
35c
34c
33c
27
3
2
6
8
21
22
30
26
23
10
9
20
29
7
5
14
15
16
13
28
11
32b
31b
25
17
24
20 SNPs

E (ST47)

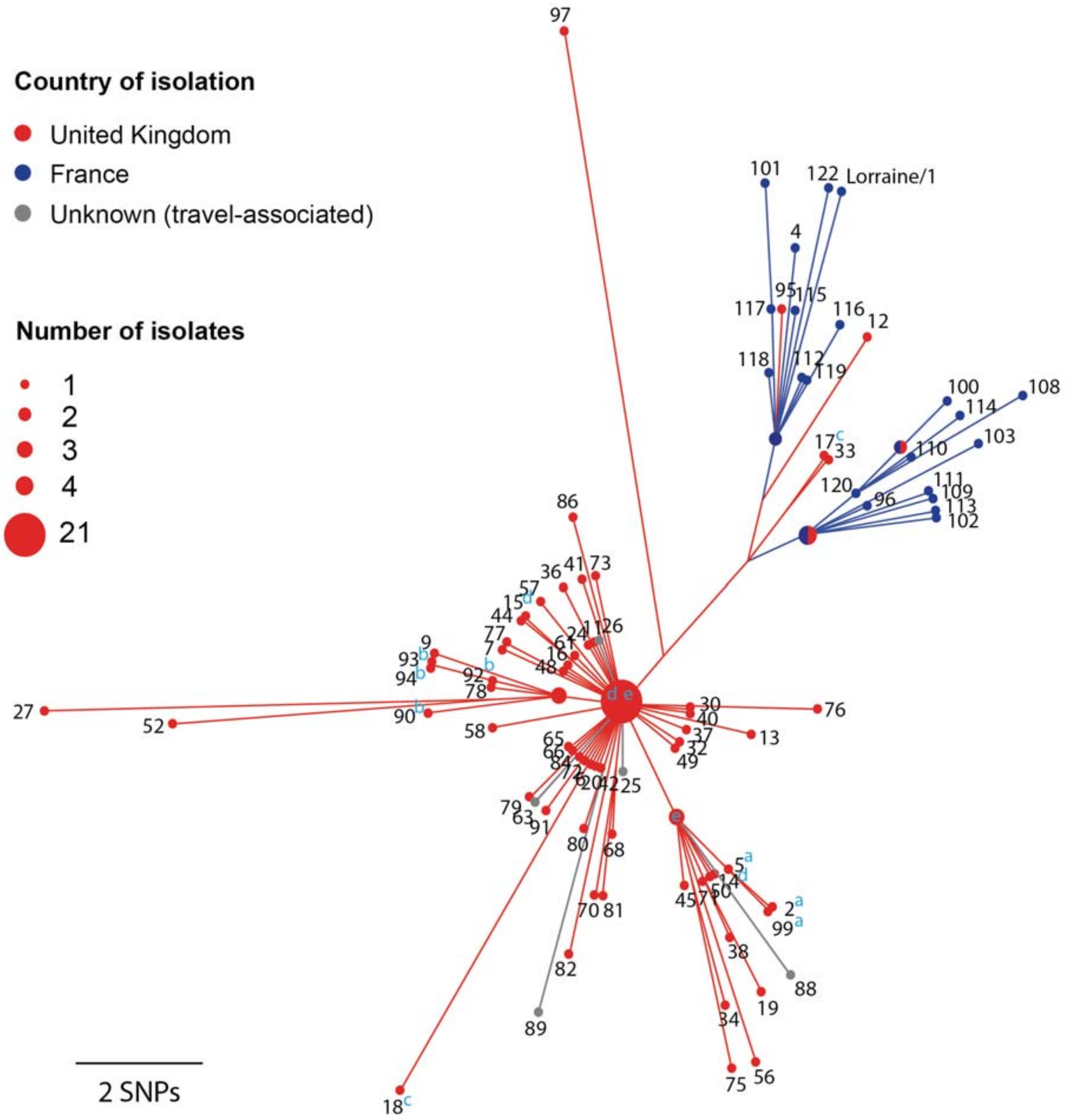


Figure 3.6. Maximum likelihood trees of the ST1 (A), ST23 (B), ST37 (C), ST62 (D) and ST47 (E) lineages. The branch lengths are scaled by the number of SNPs. Isolates are coloured according to the country in which they were recovered and branches are similarly coloured to indicate the origin of descendant nodes. If descendant isolates were recovered from multiple countries, a black dotted line is used instead. Superscripted letters indicate epidemiologically related isolates recovered from the same cluster, outbreak or patient.

3.3.5 Evidence of convergent evolution

Various approaches were used to explore whether the five disease-associated STs show evidence of convergent evolution that could explain their increased propensity to cause disease. Analysis of the gene content using *de novo* assemblies, performed by Christophe Rusniok, showed no association of the five STs with particular "accessory" genes, including effectors of the Dot/Icm secretion system. No genes were identified that were present in the five disease-associated STs and absent in the remaining species-wide collection. A small number of genes that are specific to each of the five STs were identified including 6, 17, 5, 24 and 23 in STs 1, 23, 37, 47 and 62, respectively, but most of these encode transposases, phage-related proteins and hypothetical proteins. Thus the attention was switched to core genes (i.e. those that are shared amongst all isolates).

It was hypothesised that the five disease-associated STs may have adapted to a common niche to which humans are exposed, which could explain their increased propensity to cause disease in humans. This idea was explored by searching for core genes that have undergone positive selection on the branches leading to STs 1, 23, 37, 47 and 62. The branch-site model in CodeML (Yang, 2007) was used to determine if any of the 1538 core genes found in all 364 isolates possessed a significantly higher dN/dS ratio on the branches leading to each of the five STs, in comparison to the rest of the species tree. However, while some genes had undergone positive selection on individual branches, none were common to more than one of the five branches leading to the disease-associated STs. It is possible that this result indicates a true absence of shared positive selection within core genes. However, we also acknowledge various limitations that may have hindered detection of positive selection. The first is that whilst comparing one branch leading to a disease-associated ST to the rest of the tree, it is not possible using CodeML to disregard branches leading to other disease-associated STs, which likely decreases the sensitivity of the method. Secondly, this method is usually used for detecting the occurrence of positive selection on far longer branches (e.g. between species) and there may not have been enough SNPs in the individual gene alignments of *L. pneumophila* to detect a signal.

Next, homoplasic SNPs that have occurred independently on the branches of the species tree that lead to STs 1, 23, 37, 47 and 62 were searched for. No SNPs were found to have occurred independently on all five branches although seven occurred on four of the branches, one of which is a non-synonymous change (**Table 3.3**). This SNP is in *LPC_2413* (Corby)/*lpp0942* (Paris), which encodes a diguanylate kinase with a GGDEF domain. This gene is reported to be strongly induced during the transmissive phase of infection (Bruggemann *et al.*, 2006; Weissenmayer *et al.*, 2011). A further 38 SNPs also occurred on three of the five branches, 12 of which are non-synonymous changes (**Table 3.3**). Future studies will be required to determine whether any of these SNPs affect the propensity of *L. pneumophila* to cause disease.

Table 3.3. Homoplasic SNPs on three or four of the branches leading to STs 1, 23, 37, 47 and 62. synon - synonymous; nonsynon - nonsynonymous

SNP position	Type of SNP	Base change	Gene (Corby/ Paris)	Gene product	Branches leading to
Homoplasic on four branches					
1,025,688	synon	T->C	*LPC_2453 /lpp0904*	toluene tolerance protein Ttg2B	ST1, ST37, ST47 and ST62
1,035,006	synon	T->C	*LPC_2442 /lpp0915*	transcriptional regulator FleQ	ST1, ST37, ST47 and ST62
1,035,015	synon	G->A	*LPC_2442 /lpp0915*	transcriptional regulator FleQ	ST1, ST37, ST47 and ST62
1,035,033	synon	G->A	*LPC_2442 /lpp0915*	transcriptional regulator FleQ	ST1, ST37, ST47 and ST62
1,061,079	nonsynon	T->C	*LPC_2413 /lpp0942*	diguanylate kinase (GGDEF domain)	ST1, ST37, ST47 and ST62
1,061,164	synon	A->G	*LPC_2413 /lpp0942*	diguanylate kinase (GGDEF domain)	ST1, ST37, ST47 and ST62
1,081,231	synon	T->C	*LPC_2394 / lpp0960*	A/G specific adenine glycosylase	ST1, ST37, ST47 and ST62
Homoplasic on three branches					
578,994	synon	C->T	*LPC_2858 /lpp0550*	adenylosuccinate synthetase, (PurA)	ST1, ST23, and ST47
694,526	synon	T->C	*LPC_2735 /lpp0624*	hypothetical protein	ST1, ST47 and ST62
695,083	nonsynon	T->C	*LPC_2735 /lpp0624*	hypothetical protein	ST1, ST47 and ST62
695,464	synon	A->T	*LPC_2734*	spore maturation protein A	ST1, ST47

			/lpp0625		and ST62
798,230	synon	T->A	*LPC_2649 / lpp0699*	conserved C-terminal part of RTX protein	ST1, ST47 and ST62
798,242	synon	G->A	*LPC_2649 / lpp0699*	conserved C-terminal part of RTX protein	ST1, ST47 and ST62
798,245	synon	T->A	*LPC_2649 / lpp0699*	conserved C-terminal part of RTX protein	ST1, ST47 and ST62
798,260	synon	T->A	*LPC_2649 / lpp0699*	conserved C-terminal part of RTX protein	ST1, ST47 and ST62
798,261	nonsynon	G->C	*LPC_2649 / lpp0699*	conserved C-terminal part of RTX protein	ST1, ST47 and ST62
857,435	nonsynon	A->G	*LPC_2602 /lpp0747*	ABC type dipeptide/oligopeptide/nickel transport,	ST1, ST47 and ST62
885,272	nonsynon	A->G	*LPC_2582 /lpp0766*	imidazolonepropionase, (HutI)	ST1, ST47 and ST62
973,922	synon	G->A	*LPC_2502 /lpp0854*	L-serine dehydratase, (Sdh)	ST1, ST47 and ST62
988,058	nonsynon	C->A	*LPC_2491 /lpp0866*	choloylglycine hydrolase/Peptidase C59 family protein	ST1, ST47 and ST62
988,285	synon	A->G	*LPC_2491 /lpp0866*	choloylglycine hydrolase/Peptidase C59 family protein	ST1, ST47 and ST62
988,339	nonsynon	T->G	*LPC_2491 /lpp0866*	choloylglycine hydrolase/Peptidase C59 family protein	ST1, ST47 and ST62
988,801	nonsynon	G->A	*LPC_2490 /lpp0867*	phosphoenolpyruvate synthase, (PpsA)	ST1, ST47 and ST62
989,107	nonsynon	T->A	*LPC_2490 /lpp0867*	phosphoenolpyruvate synthase, (PpsA)	ST1, ST47 and ST62
993,326	synon	G->A	*LPC_2488 /lpp0869*	nicotinate-nucleotide pyrophosphorylase, (NadC)	ST1, ST47 and ST62
993,691	synon	A->G	*LPC_2487 /lpp0870*	N-acetylglucosaminyltransferase, (MurG)	ST1, ST47 and ST62
993,901	synon	T->C	*LPC_2487 /lpp0870*	N-acetylglucosaminyltransferase, (MurG)	ST1, ST47 and ST62
993,949	synon	C->T	*LPC_2487 /lpp0870*	N-acetylglucosaminyltransferase, (MurG)	ST1, ST47 and ST62
1,020,223	nonsynon	T->G	*LPC_2461 /lpp0896*	anthranilate phosphoribosyltransferase, (TrpD)	ST1, ST47 and ST62
1,021,056	synon	G->A	*LPC_2459 /lpp0898*	ABC transporter, ATP binding protein, (LptB)	ST1, ST23 and ST37
1,023,474	synon	G->A	*LPC_2455 /lpp0902*	polysialic acid capsule expression protein, (kdsD)	ST1, ST23 and ST37
1,023,522	synon	A->C	*LPC_2455 /lpp0902*	polysialic acid capsule expression protein, (kdsD)	ST1, ST23 and ST37
1,024,718	synon	A->G	*LPC_2454 /lpp0903*	toluene tolerance ABC transporter, (Ttg2A)	ST1, ST23 and ST37

1,026,117	synon	C->T	*LPC_2453 /lpp0904*	toluene tolerance protein, (Ttg2B)	ST1, ST23 and ST37
1,042,356	nonsynon	G->T	*LPC_2433 /lpp0923*	cytochrome c-type biogenesis protein, (CcmF)	ST1, ST23 and ST37
1,042,572	synon	G->A	*LPC_2432 /lpp0924*	cytochrome C biogenesis protein, (CcmG)	ST1, ST23 and ST37
1,042,596	synon	T->C	*LPC_2432 /lpp0924*	cytochrome C biogenesis protein, (CcmG)	ST1, ST23 and ST37
1,042,692	synon	G->A	*LPC_2432 /lpp0924*	cytochrome C biogenesis protein, (CcmG)	ST1, ST23 and ST37
1,042,749	synon	A->G	*LPC_2432 /lpp0924*	cytochrome C biogenesis protein, (CcmG)	ST1, ST23 and ST37
1,042,767	synon	C->T	*LPC_2432 /lpp0924*	cytochrome C biogenesis protein, (CcmG)	ST1, ST23 and ST37
1,055,741	synon	G->A	*LPC_2418 /lpp0937*	NAD(P) transhydrogenase subunit beta, (PntB)	ST1, ST23 and ST37
1,060,955	nonsynon	C->G	*LPC_2413 /lpp0942*	diguanylate kinase (GGDEF domain)	ST1, ST47 and ST62
1,061,021	nonsynon	C->T	*LPC_2413 /lpp0942*	diguanylate kinase (GGDEF domain)	ST1, ST47 and ST62
1,078,092	synon	G->A	*LPC_2397 /lpp0957*	hypothetical protein, Sel-1 repeat protein	ST1, ST23 and ST37
1,081,123	synon	C->T	*LPC_2394 /lpp0960*	A/G specific adenine glycosylase, (MutY)	ST1, ST37 and ST62
1,081,129	synon	T->C	*LPC_2394 /lpp0960*	A/G specific adenine glycosylase, (MutY)	ST1, ST37 and ST62
1,081,513	synon	A->T	*LPC_2393 /lpp0961*	conserved hypothetical protein, (AsmA)	ST37, ST47 and ST62
1,086,849	intergenic	G->A	intergenic	N/A	ST1, ST37 and ST62
2,717,362	intergenic	A->G	intergenic	N/A	ST1, ST37 and ST62

A final approach used to search for evidence of convergent evolution between the five disease-associated STs was to identify genes with a higher than expected nucleotide similarity between the five STs compared with the rest of the species representatives. This is a potentially more powerful approach that takes into account all evolution that has occurred during the formation of the five STs rather than relying on signals of selection on the individual, sometimes short, branches leading to each of the lineages. First, a total of 1888 genes that are present in all 32 species representatives were identified excluding three isolates belonging to the *L. pneumophila fraseri* subspecies (ST154, ST336 and ST707), which were omitted from this analysis. For each of the 1888 genes, an alignment was created using one representative isolate from each of the five disease-associated STs, and excluding all other species representatives. The nucleotide diversity, a value first described by Nei and Li (1979), was calculated for each of the alignments containing the five isolates. Interestingly, many genes were found to possess very low nucleotide diversity values (meaning they are highly similar) and some genes are indeed identical between the five representative isolates (i.e. the nucleotide diversity is 0) (**Figure 3.7**). Most of these localise to a large region about a quarter of the way along the genome.

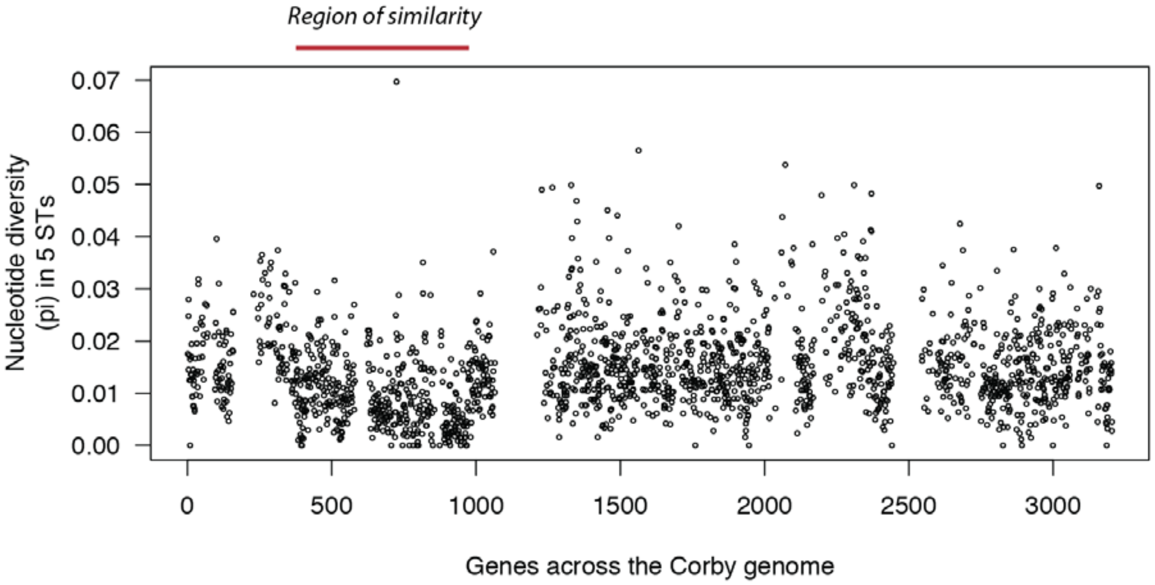


Figure 3.7. Nucleotide diversity of the five STs across the genome. Nucleotide diversity values were calculated for each of the 1888 core genes (i.e. those present in all isolates excluding ST154, ST336 and ST707) using an alignment containing a single representative isolate from each of the five disease-associated STs. Genes that are not present in all species representatives (excluding ST154, ST336 and ST707) were omitted from the analysis and thus values for these genes are not shown.

To test whether each gene possesses a significantly lower nucleotide diversity between the five representative isolates of STs 1, 23, 37, 47 and 62, than would be expected given the overall phylogenetic distance between the five STs and the conservation of each gene across the subspecies, nucleotide diversity values were calculated for all possible combinations of any five STs amongst the set of species representatives (see *Materials & Methods*). All nucleotide diversity values were adjusted using the median value for all

1888 core genes obtained for a particular combination of five STs, thus accounting for the overall phylogenetic distance between any five given STs. Nucleotide diversity values obtained for all possible combinations of five STs were then compared to the value obtained for the five disease-associated STs and p-values were derived. Multiple testing was accounted for using the Benjamini-Hochberg method. Sixty-four genes were found to contain significantly lower nucleotide diversity (higher similarity) in the five disease-associated STs than would be expected given their overall phylogenetic relatedness and gene conservation across the species ($p<0.05$) (**Table 3.4**). All 64 genes are located in a region of 725.1kb (*lpp0536/LPC_2873* to *lpp1176/LPC_0640* (**Figure 3.8**).

Table 3.4. Highly similar genes in the five STs. Genes that have a significantly lower nucleotide diversity (higher similarity) in the five major disease-associated STs than expected, given the overall phylogenetic distance between the five STs and the conservation of each gene across the *L. pneumophila pneumophila* subspecies.

Gene	**Alternative name**	**Product/function**
lpp0536	*poxF*	phenol hydroxylase
lpp0542	*rpoN*	RNA polymerase signma-54 factor RpoN
lpp0548	*hflK*	protease subunit HflK specific for phage lambda cII repressor
lpp0550	*purA*	adenylosuccinate synthetase (IMP-aspartate ligase) (AdSS) (AMPSase)
lpp0561	*ctpA*	carboxy-terminal protease
lpp0615		hypothetical protein
lpp0618		stearoyl-CoA-9-desaturase
lpp0619		hypothetical protein
lpp0626	*spmB*	spore maturation protein B
lpp0627		peptidase, M23/M37 family
lpp0643	*fthC*	5-formyltetrahydrofolate cyclo-ligase
lpp0653	*sufC*	ATP transporter, ABC binding component, ATP-binding protein
lpp0655	*sufS/csdB*	selenocysteine lyase
lpp0658	*lysS*	lysyl tRNA synthetase
lpp0661	*phtB*	major facilitator family transporter

lpp0665		hypothetical protein
lpp0676		transmembrane protein
lpp0677		conserved hypothetical protein
lpp0679		hypothetical protein conserved within *Legionellae*
lpp0680	*comA*	DNA uptake/competence protein ComA
lpp0707	*phtF*	major facilitator transporter PhtF
lpp0757	*tdh*	threonine(-3-)dehydrogenase
lpp0758		ABC transporter ATP-binding protein
lpp0759	*enhA*	enhanced entry protein EnhA
lpp0760		predicted transporter component (contains sulphur transport domain)
lpp0761		predicted transporter component
lpp0801		DNA helicase, SNF2/RAD54 family domain protein
pp0810	*lipA*	lipoic acid synthetase
lpp0865		acyl CoA dehydrogenase, short chain specific
lpp0866		choloylglycine hydrolase/Peptidase C59 family
lpp0867	*ppsA*	phosphoenolpyruvate synthase
lpp0874	*mreC*	rod shape determining protein MreC
lpp0877		hypothetical protein conserved within Legionellae
lpp0878	*icd*	isocitrate dehydrogenase, NADP-dependent
lpp0880	*clpA*	ATP binding protease component ClpA
lpp0883		lipopolysaccharide biosynthesis glycosyltransferase
lpp0887		peptidase, M23/M37 family
lpp0888	*xseA*	exonuclease VII, large subunit
lpp0890		periplasmic protein
lpp0891		diguanylate cyclase/phosphodiesterase, GGDEF and EAL domain
lpp0892		conserved hypothetical protein
lpp0893		flavin containing monooxygenase
lpp0907	*rsbV*	conserved hypothetical protein
lpp0911	*lolD*	ABC transporter, ATP binding protein
lpp0890		periplasmic protein
lpp0913		membrane fusion protein
lpp0914		hypothetical protein conserved within Legionellae
lpp0918	*ccmA*	heme exporter protein CcmA
lpp0920	*ccmC*	heme exporter protein CcmC

lpp0922	*ccmE*	cytochrome c-type biogenesis protein CcmE
lpp0931	*acdA*	acyl CoA dehydrogenase, short chain specific
lpp0932		3-hydroxyisobutyryl Coenzyme A hydrolase
lpp0933		enoyl-CoA hydratase/carnithine racemase
lpp0934		hypothetical protein

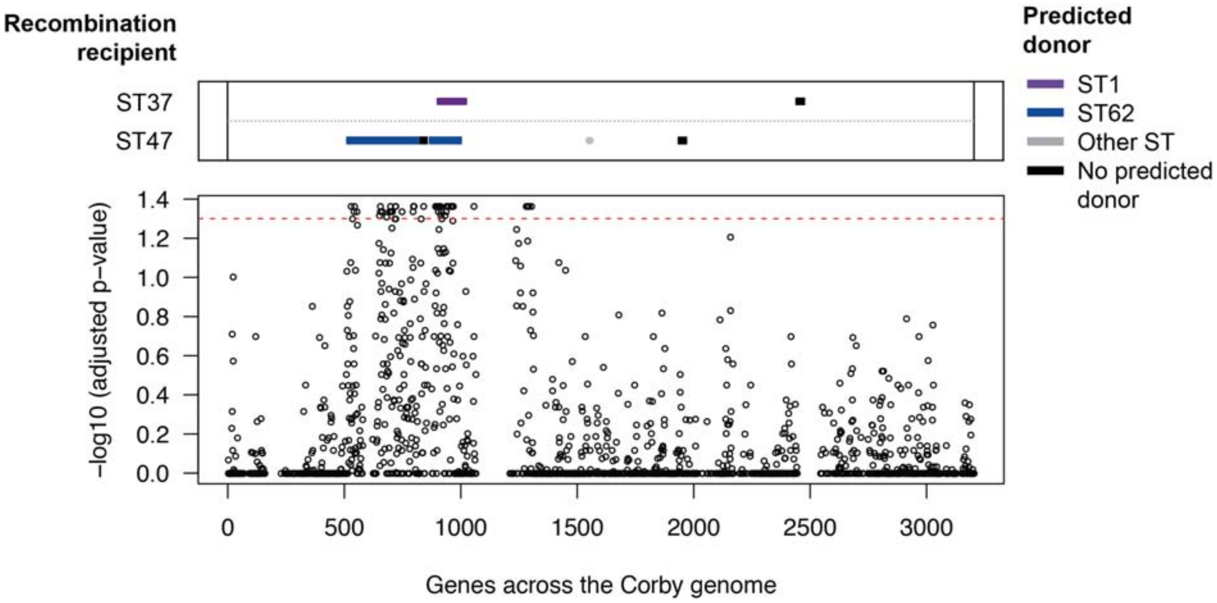


Figure 3.8. Similarity of genes across the five STs and recombination events that have occurred on the branches leading to STs 37 and 47. The bottom plot shows log-transformed p-values derived from testing whether representative isolates from STs 1, 23, 37, 47 and 62 have lower than expected nucleotide diversity values (i.e. higher similarity) in individual core genes, taking into account their nucleotide diversity across all 1888 core genes, and the overall conservation of the gene across the species representatives. Isolates from the *L. pneumophila fraseri* subspecies (ST154, ST336 and ST707) were excluded from the analysis together with ST5 and ST152, which are nested within ST1, and Philadelphia/ST36, Alcoy/ST578 and ST42, which belong to strains that are regularly associated with disease. The core genes are ordered as in the Corby genome. Genes that are not present in all species representatives (excluding ST154, ST336 and ST707) were omitted from the analysis and thus values for these genes are not shown. The red dotted line indicates the significance threshold when the Benjamini-Hochberg method is used to account for multiple testing. The top plot shows the location (with respect to the Corby genome) and predicted donor lineages of recombined regions detected on the branches leading to STs 37 and 47. Recombined regions that were detected in accessory regions

of the ST37 and ST47 genomes and which have no counterpart in the Corby genome are not shown.

Individual gene alignments of some of these 64 loci were used to construct maximum likelihood trees comprising the 32 species representatives, and these confirmed that the five disease-associated STs do indeed cluster together (**Figure 3.9**). This is in contrast to their position in a tree constructed using the whole genome and thus indicates that these genes with high similarity have been independently acquired through convergent evolution.

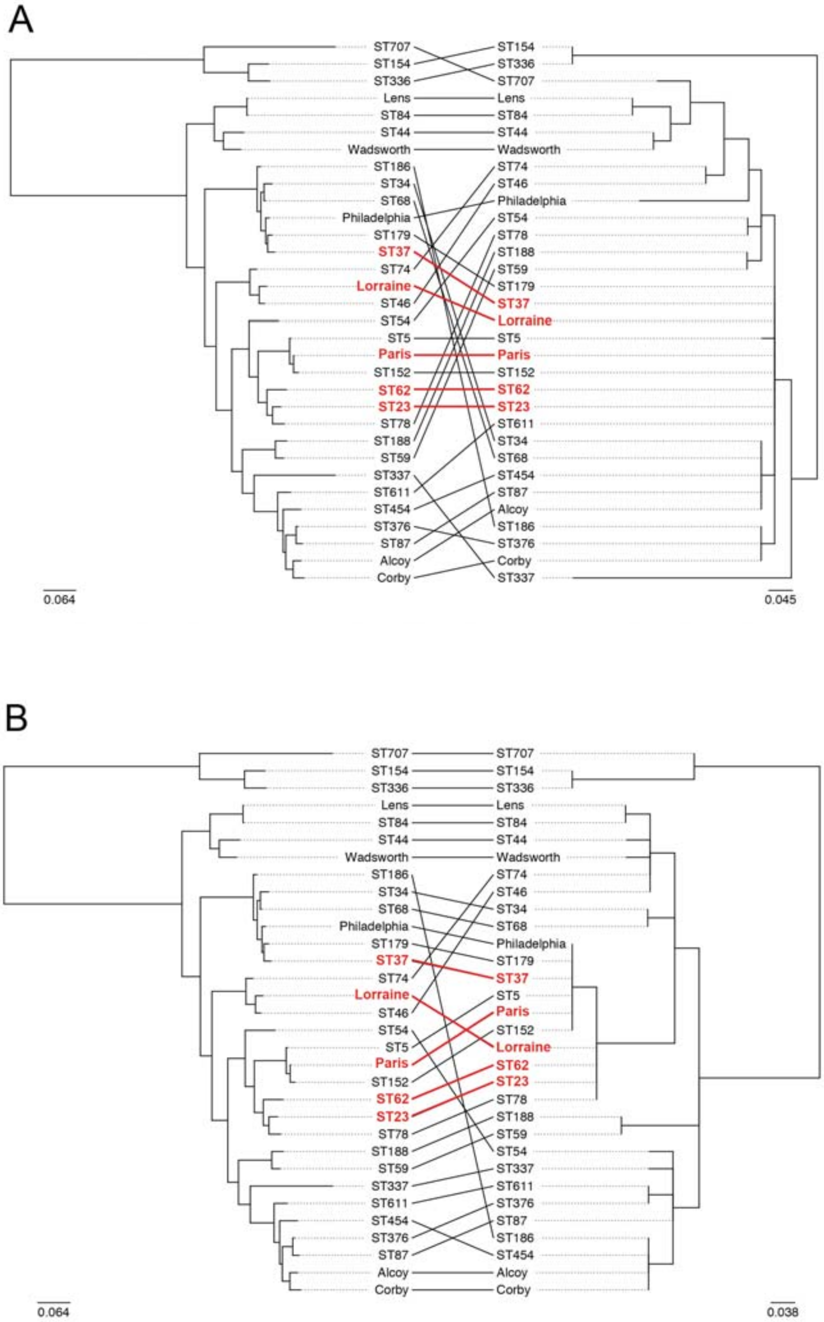


Figure 3.9. Tanglegrams comprising maximum likelihood trees of 32 STs of *L. pneumophila* that are representative of the known species diversity (previous page). The tree on the left hand side of each tanglegram was constructed using the whole genome alignment, generated by mapping sequence reads to the Corby reference genome. The trees on the right hand side were generated using individual gene alignments of either *LPC_2588* (A) or *LPC_2671* (B). Each scale bar represents the number of SNPs per variable site. The STs referred to by strain name belong to the following STs: Lens – ST15; Wadsworth – ST42; Philadelphia – ST36; Lorraine – ST47; Paris – ST1; Alcoy – ST578, Corby – ST51.

Many of the 64 genes identified in this analysis are involved in intracellular infection including those from the cytochrome c maturation (*ccm*) locus (Naylor & Cianciotto, 2004; Viswanathan *et al.*, 2002), PilR which regulates pilin and flagella synthesis, those belonging to the Pht phagosomal transporter family (Sauer *et al.*, 2005) and the enhanced entry protein, EnhA. The latter has been demonstrated to play a role in phagocytic cell entry (Cirillo *et al.*, 2000) and aquatic survival (Li *et al.*, 2015).

Genes that are highly similar in the five disease-associated STs are hypothesised to have arisen *via* recombination events before their emergence. While the branches of the species tree leading to STs 1, 23 and 62 were too long to allow for recombination detection, a number of events were identified on the branches leading to ST37 and ST47 using Gubbins. In an attempt to identify the donor lineage of each of these recombined regions, a BLASTn search was performed using each of the recombined sequences as a query and the *de novo* assemblies belonging to all 364 isolates used in this study as a BLAST database. Remarkably, several regions that were imported into the ancestor of ST47 share 100% nucleotide identity with ST62 isolates (**Table 3.5**). Many of these detected recombined regions are situated very closely to each other and may have been imported together. These regions likely imported from an ST62 isolate constitute 11.4% (396,135bp) of the ST47 chromosome. Furthermore, a large recombined region of 90,578bp was predicted to have been imported along the branch leading to the ST37 lineage and this shares 100% nucleotide identity with ST1 isolates (**Table 3.5**). This region, together with the regions imported into the ST47 ancestor from ST62, are all situated within the 725.1kb region found to contain large numbers of genes that are more similar than expected in the five disease-associated STs (**Figure 3.8**). Overall these results demonstrate that particular genomic regions have recently been exchanged between the major disease-associated STs, resulting in a common pool of allelic variants that may be affecting their propensity to cause human disease.

Table 3.5. Recombination events that occurred on the branches leading to STs 47 and 37 and their predicted origin. The start and end of the regions are with respect to the ST47 and ST37 mapping references (Lorraine, EUL 132).

Region start	**Region end**	**Length (bp)**	**Top hit**	**% identity of best hit**
Recombined regions detected on the branch leading to ST47				
530,564	629,905	99,341	ST62	100
636,480	711,192	74,712	ST62	100
719,451	765,675	46,224	ST62	100
772,825	820,139	47,314	ST62	100
848,400	850,454	2,054	ST62	100
888,240	985,915	97,675	ST62	100
990,561	1,006,506	15,945	ST62	100
1,517,688	1,521,080	3,392	ST84	99.59
1,917,592	1,946,487	28,895	No donor found	NA
1,990,956	2,002,446	11,490	ST78/ST62	98.88
2,141,795	2,143,409	1,614	ST44	99.38
2,623,816	2,625,196	1,380	ST62	100
Recombined regions detected on the branch leading to ST37				
2,063,553	2,074,733	11,180	ST74	98.09
2,183,734	2,274,312	90,578	Paris (ST1) & ST5	100

3.4 Discussion

While more than 2000 STs of *L. pneumophila* have now been reported, analysis of the SBT database demonstrated that just five STs (1, 23, 37, 47 and 62) accounted for over 40% of European isolates submitted prior to April 2015. Four of these STs (23, 37, 47 and 62) are also found very rarely in environmental sources, suggesting that their predominance in human infections may be even more pronounced. This thesis chapter aimed to understand the emergence and diversity of these five STs within the context of the species and explore whether they possess common genomic features that could be related to their increased propensity to cause disease. A total of 364 *L. pneumophila* genomes were studied here, including 329 that were newly sequenced, which constituted the largest genome collection of this species sequenced and analysed to date.

Phylogenetic analysis of the five disease-associated STs, together with isolates representative of the species diversity, showed that the five lineages have emerged independently from within a diverse species. Each of the five lineages also possesses very little diversity in the form of vertically inherited (*de novo*) mutations, suggesting that all have emerged recently. Indeed, time-dependent phylogenetic analysis of the ST37 lineage, the only ST to show some temporal signal in SNP accumulation, predicts that the MRCA existed between 1968 and 1985. Applying the substitution rate estimated for the ST37 lineage and that of previously published ST578 lineage (Sanchez-Buso *et al.*, 2014) to the ST1, ST23, ST47 and ST62 lineages, also predicts recent emergence dates for all.

L. pneumophila is a naturally competent bacterium (Stone & Kwaik, 1999) and early genomic studies using a small number of isolates predicted that recombination makes an important contribution to its evolution (Gomez-Valero *et al.*, 2011; Coscolla *et al.*, 2011). This prediction was later confirmed by genomic analysis of 45 ST578 isolates in which recombination events were found to account for almost 98% of the SNPs detected in the lineage (Sanchez-Buso *et al.*, 2014). In this study, similar results were observed in the analysis of STs 1, 23, 37 and 62 whereby 96.3%-99.0% SNPs detected in each lineage were found to be imported by recombination events. Interestingly, no recombination regions were detected within the ST47 lineage although events were detected on the internal branch of the species tree leading to the MRCA of ST47. Since a streptomycin-resistant ST47 isolate has been constructed (by collaborators at the Institut Pasteur), the possibility of the ST47 lineage having lost natural competence can be ruled out. Instead, since ST47 is predicted to have emerged only very recently, the absence of recombination may simply reflect the lack of time available for the occurrence of recombination. A second possibility is that ST47 isolates survive in a particular niche in the absence of other *L. pneumophila* lineages, and thus lack the opportunity to recombine with lineages other than their own.

Time-dependent phylogenetic analysis of the ST37 lineage predicted the substitution rate to be 2.07x10^{-7} SNPs/site/year (0.71 SNPs/genome/year), which is slightly higher than the previously estimated rate for the ST578 lineage of 1.39x10^{-7} SNPs/site/year (0.49 SNPs/genome/year) (Sanchez-Buso *et al.*, 2014). Both of these estimates are relatively low in comparison with many bacteria but similar to that of *Mycobacterium tuberculosis*, a notoriously slow-evolving pathogen (Ford *et al.*, 2013). Further support for a low substitution rate in *L. pneumophila* comes from the fact that only 20 vertically inherited SNPs were detected between the OLDA1 isolate, recovered in 1947, and another ST1 isolate, recovered in 1995. Similarly, no SNPs were identified amongst 21 ST47 isolates recovered between 2003 and 2012. Furthermore, linear regression analysis of the root-to-tip distances against sampling time in each of the five disease-associated lineages showed an extremely poor correlation in all lineages except ST37. This observation, together with the low estimates for the evolutionary rate, suggests that *L. pneumophila* may undergo periods of dormancy in which no replication occurs.

Analysis of the phylogeographic structure of the five disease-associated STs showed that isolates from different countries and even continents often cluster together and differ by just a few vertically inherited SNPs (**Figure 3.6**). Meanwhile, isolates from a similar geographical area are often more different. These observations suggest that these disease-associated STs have been involved in multiple long-distance spreading events. One possible spreading mechanism is *via* wind currents, which have previously been reported to disperse the bacteria many kilometres during outbreaks (Addiss *et al.*, 1989; Nygard *et al.*, 2008; Blatny *et al.*, 2011). Transmission *via* ocean currents is also possible and indeed *L. pneumophila* has been detected in seawater by PCR (Palmer *et al.*, 1993). It could also be that human-related activities are responsible for the spread of *L. pneumophila*. The bacteria have been shown to colonise human transport such as cruise ships (Jernigan *et al.*, 1996; Pastoris *et al.*, 1999) and trains (Quaranta *et al.*, 2012), and could also be unwittingly transported with any other man-made objects harbouring water. Compost has also been shown to contain *L. pneumophila*, the transport of which could spread the bacteria (Currie *et al.*, 2014). Of these possibilities, the spread of *L. pneumophila via* man-made environments such as modern transport would also explain the recent emergence of these STs, since they may have adapted to these new niches. However, further work is needed to elucidate the spread of *L. pneumophila* strains including those that cause a large proportion of human disease as well as those that are rarely implicated in disease. It could be extremely interesting if major disease-associated STs were transmitted more frequently across long distances (e.g. across countries and continents) than STs that are rarely implicated in disease. While this could suggest that disease-associated STs have adapted to a new environmental niche that facilitates long-distance spread, another possibility also exists.

An alternative hypothesis is that humans could contribute to the transmission of some *L. pneumophila* strains. Humans could become infected with *L. pneumophila* from man-made environments that are prone to colonisation and which they come into frequent contact with, such as domestic water systems and spa pools, and may later shed the bacteria back into other similar environments. Thus, since the emergence of these new environmental niches, a subset of strains that colonise these systems may have adapted (or were pre-adapted) to infecting humans by acquiring mutations or genes that facilitate more efficient replication in human cells. Strains that are the most efficient at infecting humans would be more frequently transmitted to other man-made water systems, allowing expansions of the strains. The fact that human vectors would likely spread *L. pneumophila* between similar environmental sites would also enhance the ability of strains to adapt to this particular niche. This scenario would explain the recent emergence of these STs, since it relies on *L. pneumophila* coming into relatively frequent contact with humans in the required infectious dose, which is more likely via modern, man-made water systems than natural sources. It also explains the wide and rapid distribution of strains since infected humans may travel long-distances, for example by air travel, before transmitting to new environments. Finally, the scenario would explain the strong association of the STs in this study with human disease.

Transmission of *L. pneumophila* from humans back into the environment is possible since *L. pneumophila* is regularly isolated from sputum samples of legionellosis patients, and has also been isolated from human feces (Rowbotham, 1998). Thus, contamination of man-made water systems via human respiratory or faecal secretions, or a combination of both, could be a possible mechanism of transmission back to the environment. Interestingly, the first probable case of human-to-human transmission has also recently been reported (Correia *et al.*, 2016). However, this reported case occurred in particularly unusual circumstances whereby the first patient, who was severely ill, was nursed by his mother for several hours in a small and non-ventilated room. Thus, due to the unusual nature of this case, and the fact that person-to-person has never previously been reported, it can be assumed that direct transmission between humans is an extremely rare event and is unlikely to play a major role in the spread of *L. pneumophila*. Instead, it is more likely that transmission between humans would occur indirectly (via an intermediate environmental source).

It could be argued that the frequency of human infection, as measured by the prevalence of Legionnaires' disease, is not high enough for these particular strains to be maintained entirely via replication in human cells. However, it could be that transmission also occurs via humans with Pontiac fever (the prevalence of which is unknown) or with asymptomatic infection. Little is known about the prevalence of asymptomatic infection although one study showed that many people who seroconverted to *Legionella* after attending the scene of a large outbreak had no symptoms (Boshuizen *et al.*, 2001). It could also be that while particular strains have acquired mutations allowing efficient replication in human cells, they also maintain the ability to replicate in other protozoan hosts. Indeed, it has previously been suggested that *L. pneumophila* maintains the ability to replicate in a wide range of host cells, rather than ever adapting to one particular host (Ensminger *et al.*, 2012).

Finally, a number of methods were used in this chapter to explore whether the five disease-associated STs share genomic features that could explain their increased propensity to cause human infection. Sixty-four genes contained within a large region of ~700kb were identified that contain higher than expected nucleotide similarity between representative isolates from STs 1, 23, 37, 47 and 62. Some of these genes have been previously reported to be involved in intracellular infection and virulence. By searching for recombination events on the branches of the species tree leading to each of the five disease-associated STs, it was shown that genes within this region have been horizontally exchanged between these STs prior to their emergence. It is hypothesised that this shared pool of allelic variants that has arisen *via* recombination may be related to the increased disease propensity of these five STs. Future confirmation could come from genomic analyses of other major disease-associated strains together with a comparative collection of strains that are never or that are very rarely implicated in human disease.

In conclusion, this chapter has provided insight into the emergence of multiple, independently evolved, major disease-associated STs of *L. pneumophila*. Remarkably, each of these STs has spread widely and rapidly since their recent emergence. The findings support the idea that humans are not "accidentally' infected by any *L. pneumophila* strain that happens to be present in an environmental source, but rather are infected by specific clones that are more efficient at human infection. Future studies are required to investigate the possible transmission of *L. pneumophila* by humans, as well as other transmission routes, in order to reduce disease burden. Since some of these clones (STs 23, 37, 47 and 62) are found rarely in commonly suspected sources, future studies should also focus on identifying their environmental niche, allowing human exposure to these bacteria to be minimised.