# 5. Evaluation of an optimal WGS-based typing scheme for *L. pneumophila*

## Declaration of work contributions

This project was conceived and supervised by Julian Parkhill and Timothy Harrison. Massimo Mentasti and Baharak Afshar performed culture and DNA extraction of all newly sequenced isolates. Martin Aslett assisted with the installation of the Bacterial Isolate Genome Sequence Database (BIGSdb) software. Rediat Tewolde assisted with the validation stages of the extended MLST methods. Simon Harris, Anthony Underwood and Norman Fry provided valuable advice throughout the project. Except where specified, I conducted the bioinformatics analyses, interpreted the data and generated all figures.

## Publication

The following work has been published:

David, S., Mentasti, M., Tewolde, R., Aslett, M., Harris, S. R., Afshar, B., Underwood, A., Fry, N. K., Parkhill, J. & Harrison, T. G. Evaluation of an optimal epidemiologic typing scheme for *Legionella pneumophila* with whole genome sequence data using validation guidelines. *Journal of Clinical Microbiology* **54**, 2135-2148 (2016).

## 5.1 Introduction

Human infection with *L. pneumophila* usually arises by inhalation of contaminated aerosols from an environmental source (Muder *et al.*, 1986). While the majority of legionellosis infections occur sporadically (Beaute *et al.*, 2013), outbreaks can also occur. Thus, when one of more cases are recognised, it is vital to rapidly establish the source of infection so that corrective measures can be implemented and further cases prevented. Identification of the source requires a combination of epidemiological information (e.g. knowledge of the patient's exposures) and microbiological characterisation of clinical and epidemiologically linked environmental isolates.

As detailed in *Chapter 1*, many microbiological methods have been used over the years for the epidemiological "typing" of *L. pneumophila* including PFGE (Luck *et al.*, 1991; Luck *et al.*, 1994; De Zoysa & Harrison, 1999), AFLP analysis (Valsangiacomo *et al.*, 1995), and mAb subgrouping (Helbig *et al.*, 1997). However, the current gold standard is SBT, a method analogous to MLST, in which isolates are assigned a ST based on the sequence of seven genes (Gaia *et al.*, 2003; Gaia *et al.*, 2005; Ratzow *et al.*, 2007; Mentasti *et al.*, 2014). This was developed by the ESGLI and is now in routine use in *Legionella* reference laboratories worldwide. The major advantage of SBT is the ease with which data can be exchanged between laboratories. This is particularly useful due to the high proportion of travel-associated cases of legionellosis (ECDC, 2015).

However, as detailed in *Chapter 3* and other studies (Borchardt *et al.*, 2008; Harrison *et al.*, 2009; Tijet *et al.*, 2010), a small number of STs are responsible for a large proportion of legionellosis cases. For example, in Europe, over 40% of epidemiologically unrelated isolates reported to the SBT database prior to April 2015 belonged to one of five STs (1, 23, 37, 47 and 62). Thus SBT can lack discriminatory power and outbreak investigations involving commonly reported STs sometimes remain unresolved.

Meanwhile, WGS is playing an increasingly prominent role in surveillance and outbreak investigations of bacterial pathogens due to the very high resolution that can be achieved (Didelot *et al.*, 2012; Kwong *et al.,* 2015). For this reason, its use in molecular typing schemes has also been considered in various recent studies (Leopold *et al.*, 2014;

Kohl *et al.*, 2014; de Been *et al.*, 2015). Importantly, the cost and turn-around time of WGS has fallen dramatically due to the emergence of NGS technologies and in some public health laboratories, including PHE (UK), WGS now costs as little as SBT whilst yielding considerably more information.

The feasibility of using WGS for the investigation of local point-source outbreaks of *L. pneumophila* has been demonstrated in several studies (Reuter *et al.*, 2013; Levesque *et al.*, 2014; Graham *et al.*, 2014; Moran-Gilad *et al.*, 2015), as described in *Chapter 1*. While most of these have used a SNP/mapping-based approach for comparing isolates, one study also described the development and use of an extended MLST scheme which compared isolates using the number of allele differences (Moran-Gilad *et al.*, 2015). Nevertheless, all studies have demonstrated high similarity between outbreak isolates with differences of <15 SNPs described between isolates from one point-source outbreak (Reuter *et al.*, 2013), and larger differences between isolates that are temporally and spatially disconnected from the outbreaks.

However, no studies have yet evaluated the feasibility of using WGS in a standardised and portable typing scheme that could be used by the *Legionella* community in a way that permits easy exchange of data. Thus the aim of this thesis chapter is to compare the performance of different WGS-based methods for the epidemiological typing of *L. pneumophila* and ultimately propose the optimal methodology for future development. The WGS-based methods include: i) SNP/mapping-based; ii) extended MLST using various numbers of genes; iii) gene presence/absence; iv) a kmer-based method. They were evaluated using a set of published criteria (van Belkum *et al.*, 2007), which include typability (*T*), reproducibility (*R*), epidemiological concordance (*E*), discriminatory power (*D*) and stability (*S*).

## 5.2 Materials & Methods

### 5.2.1 Bacterial isolates

The WGS-based methods were primarily tested using a collection of 106 clinical and environmental *L. pneumophila* sg 1 isolates (**Appendix Table 8**). This collection, known as the typing panel, was established by the ESGLI for the purpose of evaluating new typing methods and all isolates have been extensively characterised in previous studies (Fry *et al.*, 1999; Fry *et al.*, 2000; Fry *et al.*, 2002; Gaia *et al.*, 2003). The isolates were recovered from ten European countries and include an epidemiologically "unrelated" panel (79 isolates) and an epidemiologically "related" panel (44 isolates), with 17 isolates in both panels. As one isolate (EUL 112) produced a different ST to the one recorded (using both *in silico* and traditional SBT), it was replaced with another to which it was epidemiologically related (EUL 114) and which yielded the expected ST. Of these 106 isolates, 92 have been previously sequenced and analysed in *Chapters 3 & 4* of this thesis, while 14 are newly sequenced for this study.

A further 229 clinical and environmental isolates were also analysed (**Appendix Table 9**). These comprise six non-sg1 isolates, 28 isolates from well-defined point-source outbreaks in the United Kingdom (BBC, Portland Place (1988), Barrow-in-Furness (2002) and Hereford (2003)), and an additional 195 isolates from major disease-associated STs (ST 1, 37, 42, 47 and 62). The latter comprises isolates studied in *Chapters 3 & 4* of this thesis, although those without good epidemiological information were excluded. These also include both epidemiologically "unrelated" isolates together with additional sets of "related" isolates. Of the 229 additional isolates, 6 have been previously published in other studies, 220 have been sequenced and analysed in *Chapters 3 & 4* of this thesis, and 3 are newly sequenced for this study.

All newly sequenced isolates were subjected to culture and DNA extraction, performed by Massimo Mentasti and Baharak Afshar (PHE), followed by WGS at the WTSI. The methods used are described in *Chapter 2 (Materials & Methods)*.

### 5.2.2 Study design

Each WGS-based method was evaluated according to official typing criteria outlined by the European Society for Clinical Microbiology and Infectious Diseases (ESCMID) Study Group on Epidemiological Markers (ESGEM) (van Belkum *et al.*, 2007), and as in previous typing studies for *L. pneumophila* (Fry *et al.*, 1999; Fry *et al.*, 2000; Fry *et al.*, 2002; Gaia *et al.*, 2003; Gaia *et al.*, 2005). The evaluation criteria comprise typability (*T*), reproducibility (*R*), epidemiological concordance (*E*), discriminatory power (*D*), and stability (*S*). Typability is defined as the proportion of isolates that can be assigned to a type with a given method. In this study, specific criteria were defined for each of the tested methods that isolates must fulfil in order to be deemed typable (see individual sections on the methods). Reproducibility was defined as the proportion of sequencing replicate pairs that were assigned to the same type (or in which no differences were observed) with a given method. Epidemiological concordance was calculated as the proportion of epidemiologically "related" sets of isolates that were assigned to the same type (or in which no differences were observed) using each method. The index of discrimination was calculated for each of the methods using Simpson's index of diversity (Hunter and Gaston, 1988). Finally, the stability of each WGS-based method was assessed using three sets comprising isolates recovered from the same patient. The first set comprises two isolates recovered fifteen days apart. The second comprises three isolates recovered either from a sputum sample *via* direct plating, from a sputum sample *via* amoebal co-culture or from a faeces sample. The third set includes three isolates picked from single colonies on a primary isolation plate.

### 5.2.3 *De novo* assembly

*De novo* assemblies were constructed from the Illumina sequence reads of all isolates used in this study as described in *Chapter 2 (Materials & Methods)*. Quality metrics were generated to assess the quality of the assemblies and are provided in **Appendix Table 10**. The mean number of contigs is 39.9 (range, 12-140), the mean N50 value is 249,103bp (range, 81,272-2,134,649bp) and the mean length is 3,476,414bp (range, 3,229,839-3,710,927bp).

### 5.2.4 Mapping/SNP-based analysis

Due to the high diversity of the *L. pneumophila* species, it is inappropriate to use a single reference genome for mapping all isolates. Therefore, KmerID (available from https://github.com/phe-bioinformatics/kmerid), was first used to select the closest reference genome to each isolate by comparing the raw sequence reads against a collection of pre-defined reference genomes (**Appendix Table 11**). These included published complete genomes of *L. pneumophila*, as well as four genomes (EUL 28, EUL 120, EUL 165, H044120014) sequenced on the PacBio RSII sequencer, as described in *Chapter 4*. If no closely related reference genome was found to a particular isolate (i.e. the percentage kmer similarities to all references were lower than 90%), a *de novo* assembly of that isolate was used instead and added to the reference genome collection. **Appendix Table 12** lists the reference genomes used for all isolates and the depth of coverage achieved.

The sequence reads of each isolate were mapped to the chosen reference genome and SNPs were identified as described in *Chapter 2 (Materials & Methods)*. For an isolate to be deemed typable, bases must firstly be called in at least 90% positions with respect to the reference genome. Secondly, for two isolates to be assigned to the same type, bases must be called in at least 90% of all variant positions identified amongst isolates that are mapped to the same reference genome. This excludes variants in MGEs. This is to ensure that large amounts of missing data are not accountable for the apparent high similarity between isolates. Isolates that were not considered as typable were still analysed for the purpose of this study but would unlikely be used in a clinical setting.

Maximum likelihood trees of isolates mapped to the same reference genome were constructed as described in *Chapter 2 (Materials & Methods)*.

### 5.2.5 Extended MLST

The total core gene content of the *L. pneumophila* species was defined using Roary (Page *et al.*, 2015) with genome assemblies belonging to 370 *L. pneumophila* isolates (**Appendix Table 13**). These include a published set of isolates that were selected to

represent the known species diversity at the time (Underwood *et al.*, 2013) as well as isolates used in *Chapters 3 and 4* of this thesis and the current chapter. Genes that are shorter than 120bp and those without a start or stop codon were automatically discarded by Roary. Any genes with multiple copies or that contained regions susceptible to sequence-specific errors (i.e. repeat regions) (Nakamura *et al.*, 2011) were also discarded. A total of 1455 core genes remained after these filtering processes and were defined using the Philadelphia-1 type strain genome (Chien *et al.*, 2004) as a reference. These were used in a cgMLST scheme. Nested subsets of 50, 100 and 500 genes were also randomly extracted from the total 1455 core genes and used to generate smaller cgMLST schemes. **Appendix Table 14** lists the genes used in each of the schemes.

An additional two extended MLST schemes were also tested including a ribosomal MLST (rMLST) scheme (Jolley *et al.*, 2012), which uses 53 ribosomal genes present in all bacteria, and another published 1521-gene cgMLST scheme (Moran-Gilad *et al.*, 2015). The six schemes were set up using BIGSdb software (Jolley & Maiden, 2010), with extensive help from Martin Aslett (WTSI). *De novo* assemblies of all isolates were uploaded to BIGSdb and loci were identified using the integrated Genome Comparator tool. This used a BLASTn search with the default parameters including a 70% identity cut-off, a 50% length cut-off and a word size of 15. Loci were considered untypable if they were either absent or truncated due to a contig break in the assembly.

A further quality control (QC) pipeline was used to validate the loci identified by BIGSdb in each of the isolates. This first identified loci that contained 1 or more "N"s, or that contained less than 20 nucleotides (i.e. contained a large deletion in the middle of the gene), and these were considered as untypable in the affected isolates. Secondly, the raw sequence reads were mapped to the extracted loci, and any loci where there was insufficient mapping coverage to validate the allele, or where a discrepancy existed between the mapping data and the assembly in one or more base positions, were deemed untypable. This analysis was performed by Rediat Tewolde (PHE).

Only isolates that contained 100% loci that passed all the QC filters were considered as fully typable for a particular extended MLST scheme. For the purpose of this analysis, isolates with 95-100% typable loci were still analysed with any untypable loci excluded

but these could not be used to yield a "type" in a clinical setting (although the number of allele differences could still be compared with other isolates). Isolates with <95% typable genes for a particular scheme were not analysed.

Pairwise distance matrices based on allelic differences were constructed and used to generate neighbour-net trees that were inferred and visualised using SplitsTree4 (Huson & Bryant, 2006).

### 5.2.6 Gene presence/absence profiling

The same 370 isolates that were used to define the core gene content were also used to identify "accessory" genes (i.e. genes not present in all isolates) using Roary (Page *et al.*, 2015). 200 genes were identified that are present in 150 to 250 isolates and these were used in a gene presence/absence scheme (**Appendix Table 15**). The reference sequences, defined using a variety of genomes, have been deposited in the ENA under the accession numbers, FJOD01000001-FJOD01000200.

The presence or absence of the 200 accessory genes in the *de novo* assemblies of all isolates was scored using an in-house script at the WTSI. Using SMALT (v0.7.4), this tries to map each of the 200 genes to the assembly and calculates the sequence similarity and percentage coverage (length) of any match. Genes with matches of ≥90% nucleotide similarity and that covered ≥90% length were considered as present, while loci that failed to meet either of these criteria were considered as absent. An exception was loci with matches of ≥90% nucleotide similarity but that had a length of between 20-90% of the gene and that were found at a contig break. Such loci were considered as untypable.

As with the extended MLST schemes, only isolates with 100% typable loci for a particular scheme were considered as fully typable and could be used to yield a "type" in a clinical setting. However, for the purpose of this study, isolates with 95-100% typable loci were still analysed with the untypable loci excluded.

### 5.2.7 Kmer-based analysis

KmerID was used to compare isolates using the kmer content of the *de novo* assemblies. For each pair of isolates, a dissimilarity score was generated which represents the Jaccard distance between the kmer sets (i.e. the number of distinct kmers found in both assemblies over the overall number of distinct kmers found in either of the assemblies). A kmer length of 18 bases was used. Isolates were only deemed typable by this method if the lengths of their *de novo* assemblies were in the normal range (± 3 standard deviations (SD) of the mean length of all assemblies used in this study i.e. between 3,215,920bp and 3,736,908bp) and the number of contigs comprised ≤3 SDs over the mean (93 contigs). As previously, isolates with assemblies that failed to meet these criteria were still analysed although the results would unlikely be used in a clinical setting.

## 5.3 Results

Various WGS-based typing methods were tested for the epidemiological typing of *L. pneumophila* including: i) SNP/mapping-based; ii) extended MLST using various numbers of genes; iii) gene presence/absence; iv) a kmer-based method. Amongst the extended MLST schemes tested were newly designed cgMLST schemes that use 50, 100, 500 or 1455 core genes and previously published schemes using 1521 core genes (Moran-Gilad *et al.*, 2015) and 53 ribosomal genes (rMLST) (Jolley *et al.*, 2012). The typing guidelines produced by the ESGEM (van Belkum *et al.*, 2007) were used to evaluate the different methods and, in particular, five performance criteria were considered: typability (*T*), reproducibility (*R*), epidemiological concordance (*E*), discriminatory power (*D*) and stability (*S*). The methods were tested using a total of 335 isolates, which comprise the standard typing panel (*n*=106) (**Appendix Table 8**), used in all previous typing studies of *L. pneumophila* (Fry *et al.*, 1999; Fry *et al.*, 2000; Fry *et al.*, 2002; Gaia *et al.*, 2003; Gaia *et al.*, 2005), and an additional 229 isolates (**Appendix Table 9**).

### 5.3.1 Typability

The first stage in typing isolates with the SNP/mapping-based method was to determine the closest reference genome to each isolate using KmerID (see *5.2.4*). Twenty-seven reference genomes were used for mapping the total collection of isolates (*n*=335), while 25 were used for just the typing panel (*n*=106) (**Appendix Table 12**), reflecting the high diversity of the *L. pneumophila* species. Isolates that were mapped to different reference genomes could not be compared but were automatically assigned to different types. Meanwhile, those that were mapped to the same reference genome were compared and differentiated into types based on the number of SNP differences. Isolates were only considered typable by this method, firstly, if bases had been called at over 90% positions in the reference genome. Secondly, in order to classify an isolate into the same type as another, bases must be called in at least 90% of total variant positions identified in all isolates mapped to the same reference genome, to the exclusion of those in MGEs. Using these criteria, 100% of typing panel isolates (*T*=1) and 98.3% (225) of the additional 229 isolates (*T*=0.983) were considered typable (**Table 5.1 and Appendix Tables 12 and 16**).

Isolates were initially typed with the six extended MLST schemes using the Genome Comparator tool in BIGSdb, which takes *de novo* assemblies as input. The application of all six schemes to the typing panel revealed that just two loci belonging to the 1521-gene cgMLST scheme were absent or truncated in two isolates, which were thereby considered untypable with this scheme (**Appendix Table 17**). Otherwise, all loci from the six schemes were identified in all typing panel isolates. Furthermore, application of the six schemes to the additional 229 isolates revealed that all loci were identified in 94.8% (1521-gene scheme) to 100% of isolates (50-gene scheme). However, since BIGSdb provides no QC stages and could be prone to mis-classification of alleles due to assembly errors and artefacts, all loci identified by BIGSdb were further subjected to an in-house QC pipeline at PHE. The pipeline was developed and implemented in this study by Rediat Tewolde (PHE). It identifies any alleles containing one of more "N"s or that comprise less than 20 bases (i.e. contain a large deletion in the middle of the gene), two scenarios that are not flagged up by BIGSdb. The pipeline also validates all alleles by mapping the raw sequence reads to the extracted loci and highlights any alleles with poor coverage meaning that one or more bases cannot be called, or discrepant bases.

**Table 5.1. Typability of the WGS-based methods.** The typability of isolates using the SNP-based method was calculated assuming that one or more differences between isolates constitute different types (as different thresholds can alter the typability). NA – not applicable

| **Typing method** | **Typability (*T*)** | | **Gene-based schemes (typing panel isolates only)** | | |
|---|---|---|---|---|---|
| | *Typing panel only (n=106)* | *All isolates (n=335)* | *% isolates with >=98% genes typeable* | *% isolates with >=95% genes typable* | *Number (and %) of genes with 100% typability* |
| SNP-based | 1 | 0.988 | NA | NA | NA |
| rMLST (53) | 0.906 | 0.899 | 100 | 100 | 50 (94.3%) |
| cgMLST (50) | 0.991 | 0.988 | 99.1 | 100 | 48 (96.0%) |
| cgMLST (100) | 0.991 | 0.988 | 100 | 100 | 98 (98.0%) |
| cgMLST (500) | 0.972 | 0.973 | 100 | 100 | 495 (99.0%) |
| cgMLST (1455) | 0.868 | 0.916 | 100 | 100 | 1444 (99.2%) |
| cgMLST (1521) | 0.396 | 0.379 | 100 | 100 | 1462 (96.1%) |
| Gene presence/ absence | 0.415 | 0.522 | 98.1 | 100 | 179 (89.5%) |
| Kmer-based | 1 | 0.997 | NA | NA | NA |

The application of these criteria led to the rejection of more isolates as untypable (**Appendix Table 17**) and, consequently, at least one typing panel isolate lacked a full allelic profile with all six extended MLST schemes. Overall, the larger the scheme, the higher the likelihood of a sequencing or assembly artefact occurring in at least one gene, and thus the lower proportion of fully typable isolates. For example, while 99.1% of typing isolates are fully typable using the 50-gene scheme (*T*=0.991), this percentage decreases to 86.8% using the 1455-gene scheme (*T*=0.868) (**Table 5.1**). While the 53-gene rMLST scheme performed poorly for its size with only 90.6% of typing panel isolates fully typable (*T*=0.906), this can be mostly explained by a single gene (*lpg0328*) that could not be validated by the QC stage due to the absence of long enough flanking regions in the *de novo* assemblies. The previously published 1521-gene cgMLST scheme

also performed poorly and allowed the full typing of just 39.6% of isolates (*T*=0.396). All six extended MLST schemes were also tested using the additional 229 isolates, which yielded similar typability scores (**Table 5.1**).

Although a substantial number of isolates were considered untypable by one or more extended MLST schemes, it was found that 94.3-99.2% of loci from the six schemes were typable in all typing panel isolates (**Table 5.1**). Additionally, in every scheme tested, over 96% of loci could be successfully typed in every typing panel isolate. These results indicate that a low number of problematic loci account for the incomplete profiles. Indeed, of the 1865 loci used in all six schemes combined, just 61 could not be successfully typed in one or more isolates (**Appendix Table 18**). The majority of these (59) are used in the 1521-gene cgMLST scheme, explaining the low overall typability of this scheme, although 11 are also used in the newly designed 1455-gene cgMLST scheme. While 25 of the 61 untypable loci were problematic in more than one typing panel isolate and should almost certainly be excluded from any future typing scheme, 36 were unsuccessfully typed in just one isolate. Finally, various quality metrics such as the mean mapping coverage, the number of contigs in the *de novo* assemblies, and the N50 value of the assemblies, were compared between isolates that yielded a complete allelic profile in all six extended MLST schemes and those that did not. Interestingly, no significant differences were found (Student's unpaired t-test, $p>0.05$) (**Appendix Table 19**), indicating that these metrics cannot be used to predict typability.

Isolates were typed using the gene presence/absence method by determining the presence or absence of 200 accessory genes in the *de novo* assemblies. Profiles were constructed using a series of "0"s and "1"s, each unique combination of which produced a different type. Genes that were found to have ≥90% nucleotide similarity to the reference gene, but which were located at the ends of contigs were deemed untypable (see *5.2.6*). This method yielded a low overall typability score with only 41.5% typing panel isolates classed as fully typable together with 57.2% of the additional 229 isolates (**Table 5.1 and Appendix Table 20**). Despite this, each typing panel isolate contained ≥97.5% genes that were successfully typed and, overall, 89.5% of the 200 accessory genes could be typed in every typing panel isolate. Thus, similarly to the extended MLST schemes, a small proportion of genes were responsible for poor overall typability. Of the

21 genes that could not be successfully typed in one or more typing panel isolates, 15 were untypable in two or more (**Appendix Table 21**).

The last method by which isolates were typed was the kmer-based method, which calculates the dissimilarity between all pairs of isolates using the *de novo* assemblies (see *5.2.7*). Isolates with a score below a particular threshold were assigned to the same type. For an isolate to be considered typable, however, the length of the *de novo* assembly must fall within the normal range for *L. pneumophila* (±3 SD of the mean of all assemblies used in the study) and the number of contigs in the assembly must not exceed a threshold (i.e. 3 SDs over the mean). Based on these criteria, all typing panel isolates were considered typable by this method together with all but one of the additional 229 isolates, H063860003 ($T$=0.997) (**Table 5.1 and Appendix Tables 10 and 16**).

### 5.3.2 Reproducibility

To test the reproducibility ($R$) of the WGS-based methods, six typing panel isolates (EUL 27, 33, 69, 75, 92, 111) that belong to a variety of STs were sequenced twice. Along with the remainder of the typing panel, they were sequenced at the WTSI using the same protocols, as described in *Chapter 2 (Materials & Methods)*. No differences were found amongst any sequencing pairs by either the SNP/mapping-based method, any of the six extended MLST schemes and the gene presence/absence method, following implementation of the QC measures described (**Table 5.2 and Figure 5.1**). Furthermore, the dissimilarity scores calculated using the kmer-based method were extremely low (<0.001), demonstrating only very small differences between the assemblies, and were of the same order of magnitude in all six pairs (**Table 5.2 and Figure 5.1**). Overall these results indicate that all the tested WGS-based methods are reproducible and all were assigned $R$ values of 1 (**Table 5.2**).

**Table 5.2. Number of differences identified between sequencing replicates using each of the WGS-based methods.** For each of the extended MLST schemes, both the number of differences identified by BIGSdb software (pre-QC) and the number identified after all alleles are validated by the QC stages are given, the latter in brackets. For the gene/presence absence method, the numbers of differences identified before and after the exclusion of partially present genes on contig boundaries are given, the latter in brackets. The difference between replicates as calculated by the kmer-based method is expressed using the Jaccard dissimilarity score.

| **EUL number** | **Number of differences between replicates** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SNP-based** | **rMLST (53)** | **cgMLST (50)** | **cgMLST (100)** | **cgMLST (500)** | **cgMLST (1455)** | **cgMLST (1521)** | **Gene presence /absence** | **Kmer-based** |
| | *SNPs* | *Alleles* | *Alleles* | *Alleles* | *Alleles* | *Alleles* | *Alleles* | *Genes* | *Jaccard distance* |
| 27 | 0 | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 1 (0) | 0 (0) | 0.00029 |
| 33 | 0 | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0.00050 |
| 69 | 0 | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 1 (0) | 0.00051 |
| 75 | 0 | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 1 (0) | 0 (0) | 0.00028 |
| 92 | 0 | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 1 (0) | 0.00052 |
| 111 | 0 | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 4 (0) | 0 (0) | 0.00064 |
| | | | | | | | | | |
| **Reprod. (*R*)** | **1** | **1** | **1** | **1** | **1** | **1** | **0.5 (1)** | **0.66 (1)** | **1** |

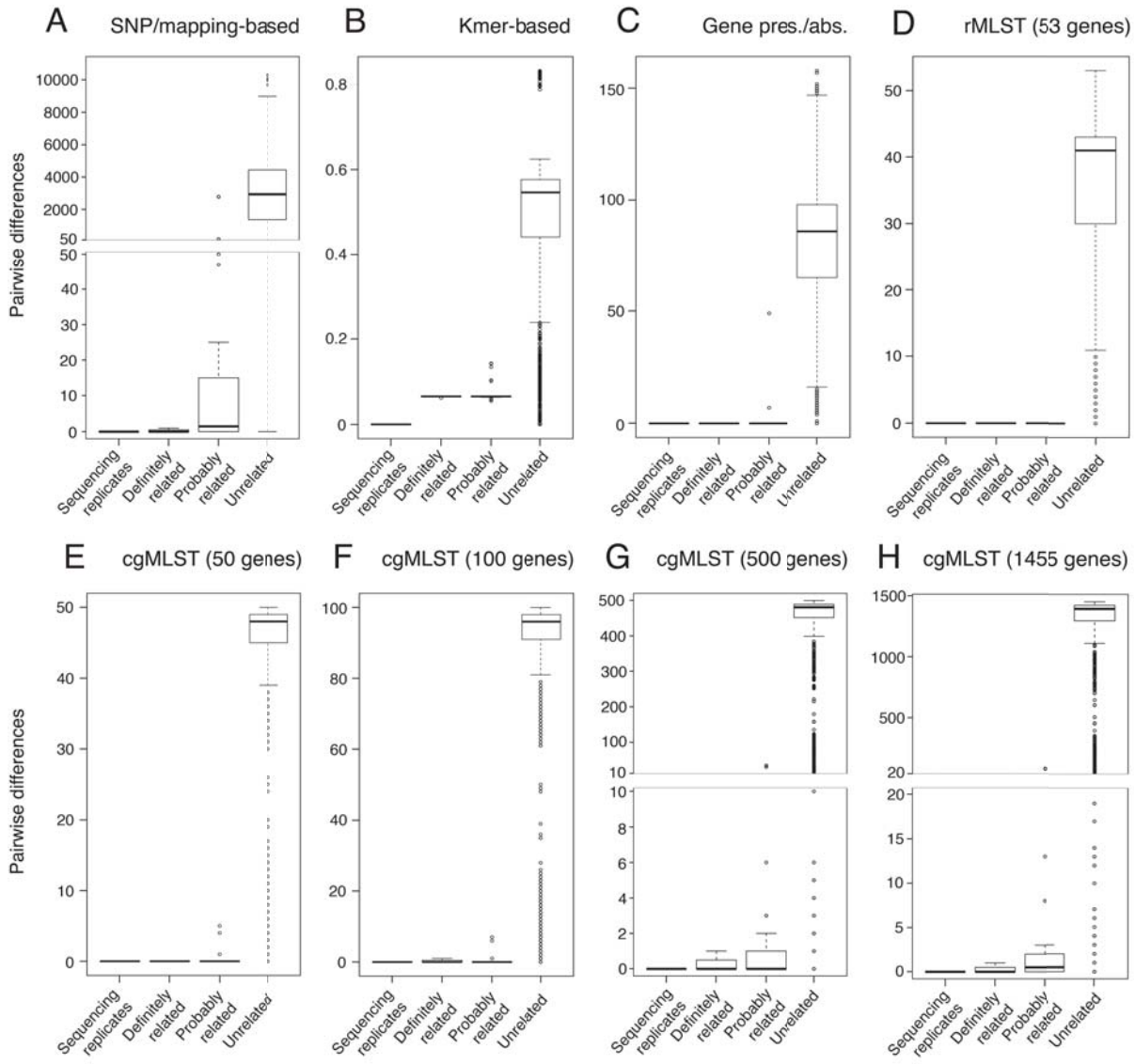


**Figure 5.1. Pairwise differences between typing panel isolates using different WGS-based methods (A-H).** Included are sequencing replicates (6 pairs), "definitely related" isolates (10 isolates comprising 4 sets), "probably related" isolates (34 isolates comprising 13 sets) and 79 epidemiologically "unrelated" isolates.

### 5.3.3 Epidemiological concordance

Seventeen epidemiologically "related" sets comprising 44 isolates from the typing panel were first used to assess the epidemiological concordance of the WGS-based methods. These isolates include four "definitely related" sets comprising a total of ten isolates, and 13 "probably related" sets comprising a total of 34 isolates (**Appendix Table 8**). Those

considered “definitely related” are either replicates or were recovered from the same patient, while those considered “probably related” were either associated with a point source outbreak or were isolated in a similar time and geographical location. Thus, the latter may not necessarily be genotypically related. All 17 sets are concordant (i.e. no differences found between isolates from the same set) using mAb subgrouping, RFLP analysis and AFLP analysis (Fry *et al.*, 1999; Fry *et al.*, 2000), and both 3- and 6-allele SBT, the latter of which were used before the introduction of the current gold standard 7-allele SBT. However, later testing of the sets by 7-allele SBT revealed that one set (EUL 37, 44 and 45) is discordant by this method (EUL 37 and 44 are ST1 while EUL 45 is ST72), suggesting that these isolates could have been falsely linked. In this analysis, all isolates that were not deemed typable using any of the WGS-based methods were still included except those with <95% successfully typed loci in the extended MLST or gene presence/absence schemes. All isolates from the typing panel were therefore included, although it is important to note that those yielding small amounts of missing data may lead to a slight over-estimation of the epidemiological concordance values.

Using each of the WGS-based methods, isolates that are identical were first assigned to the same type while those containing one of more differences were assigned to different types. This is the simplest means of classification that is currently used by the *L. pneumophila* SBT scheme and other bacterial MLST schemes, and also permits the highest level of discrimination to be attained for a particular method. However, since not even sequencing replicates were found to be identical using the kmer-based method, isolates were assigned to types with this method by first defining a threshold equivalent to the largest difference observed between sequencing replicates (0.00064). Single linkage clustering was then used to classify isolates, and isolates with a dissimilarity score equal to or less than 0.00064 were clustered into the same type, together with isolates linked to the cluster by at least one isolate. **Table 5.3** and **Figure 5.2A** show the epidemiological concordance (*E*) values achieved by the different WGS-based methods when these criteria are applied.

**Table 5.3. Index of discrimination (*D*) and epidemiological concordance (*E*) of the current and tested WGS-based typing methods.** The number of types and *D* values were calculated using 79 epidemiologically "unrelated" isolates (panel 1) from the typing panel. The *E* values were calculated using a total of 44 epidemiologically "related" isolates (panel 2) from the typing panel that include both "definitely related" (subdivision I) and "probably related" (subdivision II) isolates.

| **Typing method** | **Thres-hold** | **No. of types** | **Index of discrimination (*D*)** | **Epidemiological concordance (*E*)** | | |
|---|---|---|---|---|---|---|
| | | | | **Subdivision I ("definitely related")** | **Subdivision I & II ("definitely related" and "probably related")** | **Subdivision I & II (excluding EUL37/44/45) *** |
| SBT | 0 | 40 | 0.940 | 1 (4/4) | 0.941 (16/17) | 1 (16/16) |
| SBT + mAb subgrouping | 0 | 43 | 0.968 | 1 (4/4) | 0.941 (16/17) | 1 (16/16) |
| SNP/ mapping-based | 0 | 78 | 0.999 | 0.750 (3/4) | 0.353 (6/17) | 0.375 (6/16) |
| | 1 | 77 | 0.999 | 1 (4/4) | 0.471 (8/17) | 0.500 (8/16) |
| rMLST (53) | 0 | 44 | 0.972 | 1 (4/4) | 1 (17/17) | 1 (16/16) |
| cgMLST (50) | 0 | 57 | 0.990 | 1 (4/4) | 0.941 (16/17) | 1 (16/16) |
| cgMLST (100) | 0 | 59 | 0.991 | 0.750 (3/4) | 0.824 (14/17) | 0.875 (14/16) |
| | 1 | 53 | 0.983 | 1 (4/4) | 0.941 (16/17) | 1 (16/16) |
| cgMLST (500) | 0 | 71 | 0.997 | 0.750 (3/4) | 0.529 (9/17) | 0.563 (9/16) |
| | 1 | 67 | 0.990 | 1 (4/4) | 0.824 (14/17) | 0.875 (14/16) |
| cgMLST (1455) | 0 | 75 | 0.998 | 0.750 (3/4) | 0.471 (8/17) | 0.500 (8/16) |
| | 1 | 72 | 0.996 | 1 (4/4) | 0.647 (11/17) | 0.688 (11/16) |
| cgMLST (1521) | 0 | 76 | 0.999 | 0.750 (3/4) | 0.412 (7/17) | 0.438 (7/16) |
| | 1 | 72 | 0.996 | 1 (4/4) | 0.529 (9/17) | 0.563 (9/16) |
| Gene presence/ absence | 0 | 53 | 0.976 | 1 (4/4) | 0.882 (15/17) | 0.938 (15/16) |
| Kmer-based | 0.00064 | 71 | 0.996 | 0 (0/4) | 0 (0/17) | 0 (0/16) |
| | 0.065 | 41 | 0.945 | 1 (4/4) | 0.824 (14/17) | 0.875 (14/16) |

*The set of "probably related" isolates comprising EUL 37, 44 and 45 is not epidemiologically concordant via 7-allele SBT, suggesting these may be falsely linked isolates, and thus *E* values were also calculated excluding this set.

Only three methods achieved full epidemiological concordance (*E*=1) for the four "definitely related" sets, which were rMLST, 50-gene cgMLST and the gene presence/absence method. The same three methods also performed well using the 13 "probably related" sets, with rMLST achieving full concordance (*E*=1), 50-gene cgMLST achieving concordance for 12 of the 13 sets (the exception being EUL 37, 44 and 45, which is discordant by SBT) (*E*=0.923), and the gene presence/absence method achieving concordance for 11 of the 13 sets (the two exceptions being EUL 37, 44 and 45, and EUL 19, 22, 23 and 24) (*E*=0.846) (**Table 5.3**). However, the larger the number of genes used in the extended MLST schemes, the lower the epidemiological concordance. For example, only 4 of the 13 "probably related" sets were concordant with the largest scheme (1521-gene cgMLST) (*E*=0.308) (**Table 5.3**). A neighbour-net tree inferred from the pairwise allelic differences between typing panel isolates shows the epidemiological concordance obtained using the 100-gene scheme (**Figure 5.3**). Finally, only 5 of the 13 "probably related" sets were concordant using the SNP/mapping-based method, whilst not a single set was concordant using the kmer-based method with a threshold set at 0.00064 (**Table 5.3**).

Since at least "definitely related" isolates should be classified into the same type by a given typing scheme, a new approach was tested for each of the WGS-based methods that failed to produce full epidemiological concordance for the four "definitely related" sets. As used with the kmer-based method previously, single linkage clustering was applied to classify isolates using the smallest threshold possible that would provide epidemiological concordance for at least the "definitely related" isolates. The resulting thresholds were one allele difference in the extended MLST schemes with 100, 500, 1455 or 1521 loci, one SNP difference using the SNP/mapping-based method, and a dissimilarity score of 0.065 using the kmer-based method. Applying the methods with these new thresholds increased the epidemiological concordance of the "probably related" sets and, notably, the number of "probably related" sets that were concordant using the kmer-based method increased from 0 to 11 (**Figure 5.2B**). However, many

sets were still discordant using the extended MLST schemes (particularly those with larger numbers of genes) and the SNP-based scheme.

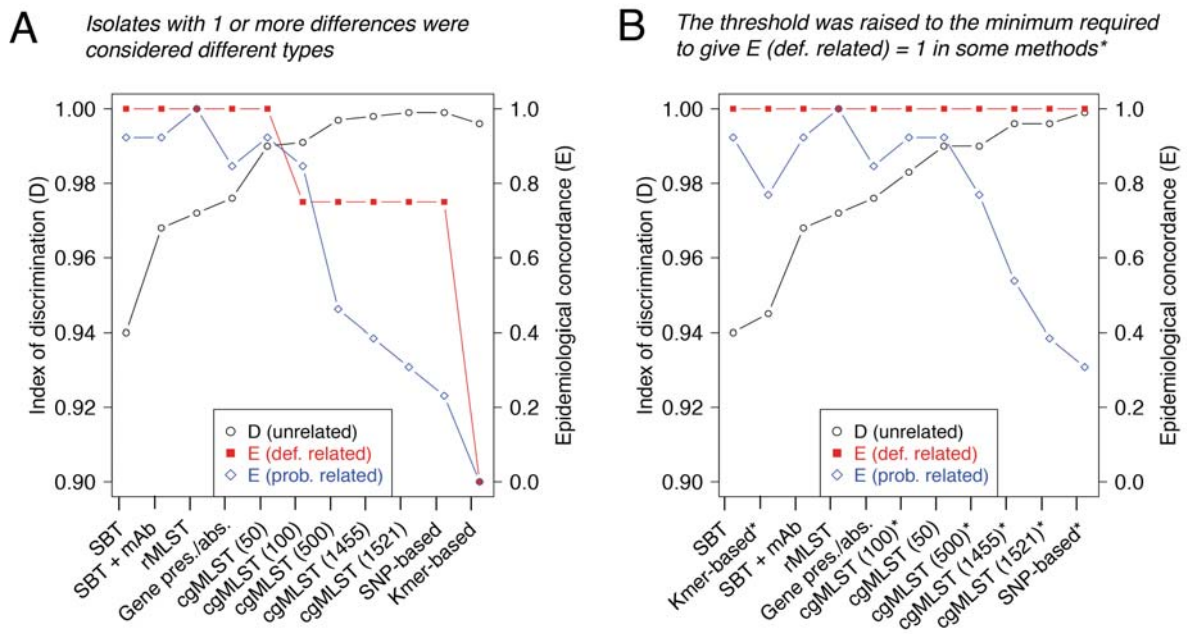


**Figure 5.2. Index of discrimination (*D*) and epidemiological concordance (*E*) of the current and WGS-based methods.** A) Isolates from the typing panel (*n*=106) were classified as the same type if they shared no differences and a different type if they shared 1 or more differences, except using the kmer-based method where isolates were categorised into types using single-linkage clustering with a threshold equal to the maximum difference detected between sequencing replicates. B) The *D* and *E* values of each of the current and WGS-based methods when single linkage clustering was used for some methods with a threshold that maintains *E* of at least "definitely related" isolates at 1. The threshold is one allele difference using the cgMLST schemes with 100 or more genes, one SNP using the SNP-based method, and 0.065 using the kmer-based method. Using the rMLST scheme, the 50-gene cgMLST scheme and the gene presence/absence scheme, isolates were classified as different types if they shared 1 or more differences (as in A).

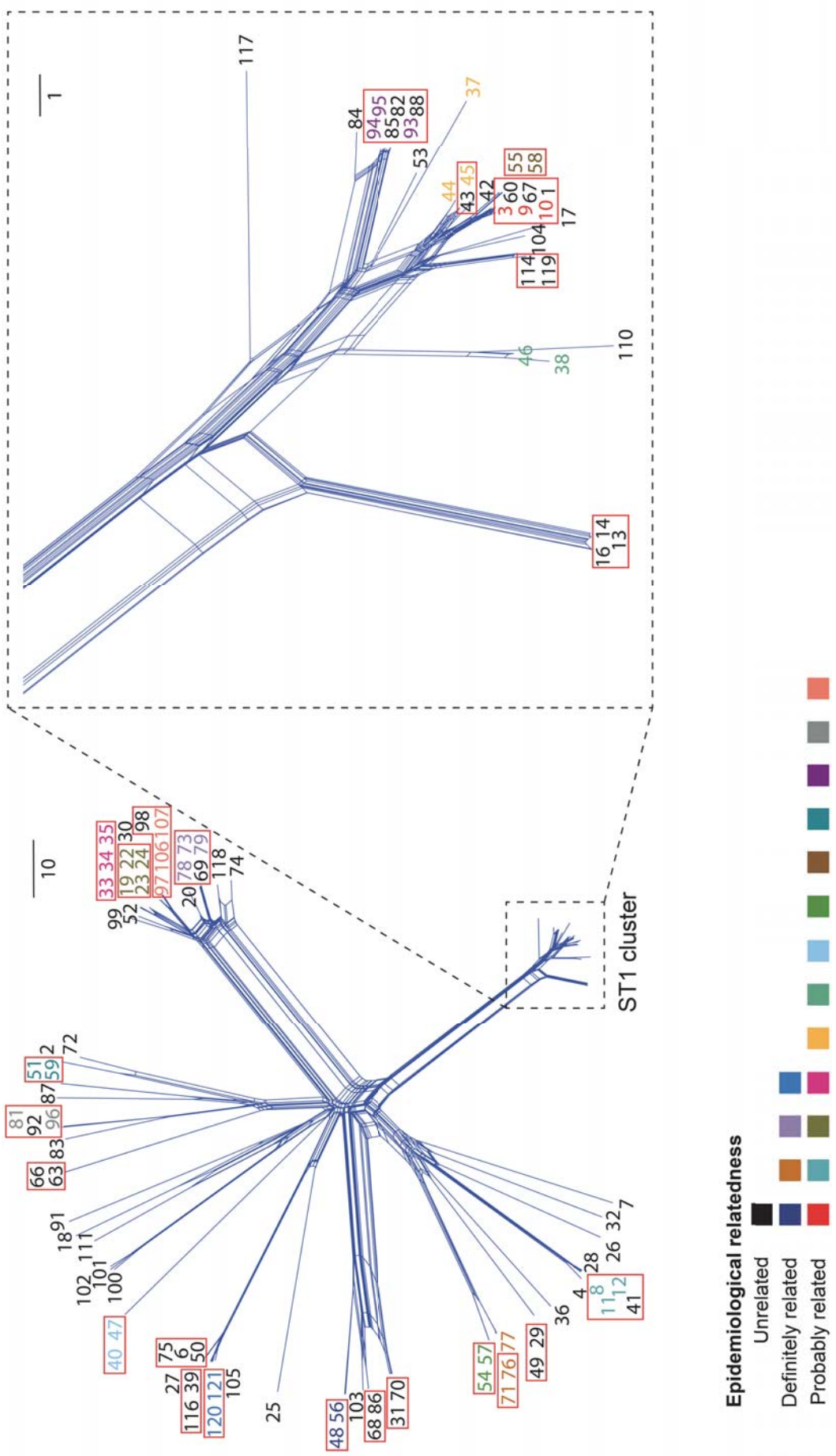
1
10
ST1 cluster
Epidemiological relatedness
Unrelated
Definitely related
Probably related

**Figure 5.3 Neighbour-net tree of the typing panel isolates constructed using the 100-gene cgMLST scheme (previous page).** Isolates (*n*=106) are labelled according to their "EUL" number and coloured by their epidemiological relatedness as indicated in the key. Isolates belonging to the same type (i.e. with no allele differences) are enclosed in a red box. The ST1 cluster, comprising both ST1 isolates and isolates derived from ST1, is shown at a higher resolution on the right. The scale bars indicate the number of allelic differences.

To establish the extent to which the clustering thresholds would need to be further raised to maintain complete epidemiological concordance, the numbers of differences between isolates belonging to "probably related" sets were determined (**Table 5.4**). This also highlighted sets with differences far greater than the majority, which may comprise isolates that were falsely linked. Using the SNP/mapping-based method, the differences identified between "probably related" isolates were very wide ranging, from 0 to 2786. The set with the highest number of SNP differences was, unsurprisingly, that comprising EUL 37, 44 and 45 (range, 179-2786), which was found to be discordant using even 7-allele SBT. However, another set (EUL 19, 22, 23 and 24) included a clinical isolate (EUL 19) that differs by 25-50 SNPs to the remaining three isolates in the set. Meanwhile, differences between the three remaining isolates are only 0-1 SNPs, suggesting the fourth isolate may have been incorrectly linked to the cluster. Isolates belonging to the remaining 11 "probably related" sets share a similar number of SNP differences, ranging from 0 to 16 (**Table 5.4**). As expected, the number of allelic differences between EUL 37, 44 and 45 found using the extended MLST schemes with either 100, 500, 1455 and 1521 genes were also substantially larger than those identified in most sets (**Table 5.4**). Interestingly though, the maximum differences found between isolates belonging to EUL 19, 22, 23 and 24 were just 3 and 8 using the 1455-gene and 1521-gene cgMLST schemes, respectively, which is not out of the range observed in other "probably related" sets. This suggests that the majority of the SNPs identified between these isolates using the SNP/mapping-based method are in *L. pneumophila* "accessory" regions present in the reference genome, but which have been excluded from even the largest of the extended MLST schemes. Thus, excluding only EUL 37, 44 and 45, the number of allelic differences observed between isolates in the remaining "probably related" sets were 0-1, 0-3, 0-8 and 0-13 using the extended MLST schemes with 100, 500, 1455 and 1521 genes, respectively. Using the gene

presence/absence scheme, the only two sets to be discordant are also EUL 37, 44 and 45, and EUL 19, 22, 23 and 24. These contain up to 7 and 49 differences, respectively, indicating differences in gene content that are particularly prominent in the latter set. Finally, the same two sets are also discordant using the kmer-based method by which they were assigned dissimilarity scores of 0.10-0.13 (EUL 37, 44 and 45) and 0.057-0.14 (EUL 19, 22, 23 and 24). Both sets include scores that are substantially larger than those observed between other "probably related" isolates. One further "probably related" pair (EUL 51 and 59) was also discordant by the kmer-based method, but was assigned a dissimilarity score only slightly higher than the threshold of 0.065.

**Table 5.4. Number of differences between isolates from epidemiologically "related" sets using each of the WGS-based methods.** All "related" sets (*n*=17) from the typing panel are included as well as three epidemiologically "related" pairs of non-sg1 isolates, and isolates from a further three point-source outbreaks. Sets in which one or more isolates could not be fully typed by a particular scheme are marked with an asterisk.

| **EUL number/ outbreak** | **Mean (and range) of differences** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SNP-based** | **rMLST (53)** | **cgMLST (50)** | **cgMLST (100)** | **cgMLST (500)** | **cgMLST (1455)** | **cgMLST (1521)** | **Gene pres./ abs.** | **Kmer-based** |
| | *SNPs* | *Alleles* | *Alleles* | *Alleles* | *Alleles* | *Alleles* | *Alleles* | *Genes* | *Jaccard distance* |
| *Typing panel subdivision I sets ("definitely related")* | | | | | | | | | |
| 48, 56 | 0 | 0* | 0 | 0 | 0 | 0 | 0* | 0 | 0.065 |
| 71, 76, 77 | 0.67 (0-1) | 0 | 0 | 0.67 (0-1) | 0.67 (0-1) | 0.67 (0-1) | 0.67 (0-1)* | 0 (0-0) | 0.065 (0.065-0.065) |
| 73, 78, 79 | 0 | 0 | 0 | 0 | 0 | 0 | 0* | 0 (0-0)* | 0.064 (0.061-0.065) |
| 120, 121 | 0 | 0* | 0 | 0 | 0 | 0 | 0 | 0* | 0.0646 |
| *Typing panel subdivision II sets ("probably related")* | | | | | | | | | |
| 3, 9, 10 | 2.67 (0-4) | 0 | 0 | 0 | 0.67 (0-1) | 1.33 (0-2) | 2.67 (1-4)* | 0 (0-0) | 0.065 (0.064-0.065) |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 8, 11, 12 | 10.33 (0-16) | 0 | 0 | 0 | 2 (0-3) | 5.33 (0-8) | 8.67 (1-13)* | 0 (0-0)* | 0.065 (0.065-0.065) |
| 19, 22, 23, 24 | 20.67 (0-50) | 0 | 0 | 0 | 0.5 (0-1) | 1.5 (0-3) | 4 (0-8)* | 24.5 (0-49)* | 0.10 (0.057-0.14) |
| 33, 34, 35 | 1 (0-2) | 0 | 0 | 0 | 0* | 0* | 3.67 (3-5)* | 0 (0-0)* | 0.064 (0.063-0.065) |
| 37, 44, 45 | 1915 (179-2786) | 0 | 3.33 (1-5) | 4.67 (1-7) | 24 (6-35) | 58 (13-82) | 60 (9-87)* | 4.67 (0-7) | 0.11 (0.10-0.13) |
| 38, 46 | 5 | 0 | 0 | 1 | 2 | 3 | 5* | 0 | 0.064 |
| 40, 47 | 0 | 0 | 0 | 0 | 0 | 0 | 0* | 0* | 0.054 |
| 51, 59 | 15 | 0 | 0 | 0 | 1 | 8* | 8* | 0* | 0.066 |
| 54, 57 | 2 | 0 | 0 | 0 | 0* | 0* | 1* | 0* | 0.063 |
| 55, 58 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.063 |
| 81, 96 | 6 | 0 | 0 | 0 | 0 | 1 | 2* | 0* | 0.059 |
| 93, 94, 95 | 0.67 (0-1) | 0 | 0 | 0 | 0 | 0* | 0* | 0 (0-0) | 0.065 (0.065-0.065) |
| 97, 106, 107 | 0 | 0* | 0 | 0 | 0 | 0 | 0* | 0 (0-0)* | 0.065 (0.064-0.065) |
| *3 pairs of epidemiologically "related" non-sg1 isolates* | | | | | | | | | |
| 153, 158 | 0 | 0 | 0 | 0 | 0 | 0 | 0* | 0* | 0.00017 |
| 154, 155 | 3 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0.00020 |
| 156, 159 | 2 | 0 | 0 | 0 | 0 | 1 | 1* | 0* | 0.00065 |
| *Additional point-source outbreaks* | | | | | | | | | |
| Barrow outbreak (*n*=18) | 0.48 (0-2) | 0 | 0 | 0 | 0 | 0 | 0* | 0 (0-0)* | 0.00052 (0.00030-0.00075) |
| BBC outbreak (*n*=5) | 1 (0-2) | 0 | 0 | 0 | 0 | 0-4 (0-1)* | 0.4 (0-1)* | 0* | 0.00052 (0.00039-0.00071) |
| Hereford (*n*=5) | 4 (0-9) | 0* | 0* | 0* | 0.4 (0-1)* | 0.8 (0-2)* | 1.6 (0-4)* | 0 (0-0)* | 0.0061 (0.00044-0.013) |

In addition to the 17 typing panel sets, three sets of isolates from UK outbreaks with well-defined point sources (BBC, Portland Place, 1988; Barrow-in-Furness, 2002; Hereford, 2003) were also used to test the epidemiological concordance of the WGS-based methods (**Table 5.4**). Using the SNP-based method, up to 2 SNPs were found between the 5 isolates linked to the BBC outbreak. No differences were identified using the rMLST scheme, the cgMLST schemes using either 50, 100, or 500 genes, or the gene presence/absence scheme and just one difference was identified using the cgMLST schemes with either 1455 or 1521 genes. Interestingly, unlike any sets in the typing panel, the kmer dissimilarity scores were in a similar range to those of sequencing replicates. Nineteen isolates linked to the Barrow outbreak share up to 2 SNP differences although no differences were found using the rMLST scheme, any of the cgMLST schemes or the gene presence/absence method. The kmer dissimilarity scores were also in a similar range to those of sequencing replicates. Finally, the 5 Hereford outbreak-linked isolates share up to 9 SNPs, but no differences were observed using the rMLST scheme, the cgMLST schemes with 50 and 100 genes, or the gene presence/absence scheme. Up to 1, 2 and 4 differences were found using the cgMLST schemes with 500, 1455 and 1521 genes, respectively. Some, but not all, pairwise kmer dissimilarity scores were in a similar range to the sequencing replicates although all are below the threshold of 0.065. Thus, if isolates were assigned to the same type only if they were identical (or using a threshold of 0.00064 with the kmer-based method), epidemiological concordance for the three sets would be achieved only using rMLST, 50-gene or 100-gene cgMLST, and the gene presence/absence scheme.

The number of differences between isolates belonging to another 17 epidemiologically "related" sets were also analysed, including three non-sg1 sets and 14 sets comprising isolates belonging to some of the major disease-associated STs. Pairwise differences between isolates from these sets, as analysed by all the WGS-based methods, are provided in **Table 5.4** and **Appendix Table 22**, and SNP differences between isolates from sets belonging to major disease-associated STs are also shown in **Figure 5.4**. A SNP-based phylogenetic tree of 74 ST37 isolates also shows the number of SNP differences found between epidemiologically "related" isolates (**Figure 5.5**). Overall, the majority of these additional epidemiologically "related" sets are concordant using rMLST, the cgMLST schemes with either 50 or 100 genes, and the gene

presence/absence scheme, but not with the more discriminatory methods (allowing for no differences between isolates of the same type, or using a threshold of 0.00064 with the kmer-based method).

### 5.3.4 Discriminatory power

The discriminatory power of each of the WGS-based methods was first tested using 79 epidemiologically "unrelated" isolates from the typing panel (**Appendix Table 8**). As in the epidemiological concordance analysis, isolates previously designated as untypable were still used, the result of which could result in the slight under-estimation of discriminatory power. As previously, isolates were firstly assigned to the same type if they are identical, or different types if they contain one or more differences, with the exception of the kmer-based method whereby single-linkage clustering with a threshold of 0.00064 (the maximum difference between sequencing replicates) was used. Secondly, in order to achieve complete epidemiological concordance of at least the "definitely related" isolates, the previously defined thresholds were also used. The indices of discrimination (*D*), calculated using Simpson's index of diversity (Hunter & Gaston, 1988), for all tested methods are shown in **Table 5.3** and **Figure 5.2**. All WGS-based methods had greater discriminatory power than the current gold standard, SBT, as well as the commonly used combination of SBT and mAb subgrouping. The *D* values of all individual loci used in the extended MLST schemes and the gene presence/absence scheme were also calculated and are provided in **Appendix Tables 23, 24 and 25.**

Finally, a disadvantage of the current SBT scheme is that a large proportion of clinical isolates are classified into just a small number of STs such as those described in *Chapters 3 and 4*. Thus, the ability of each of the WGS-based methods to differentiate between isolates belonging to some of the major disease-associated STs (1, 37, 42, 47 and 62) was tested. In this analysis, isolates were classified into the same type if they were identical, and different types if they contained one or more differences. One ST1 isolate, H034800423, was not included in any of the analyses since up to 30% of the loci could not be successfully typed with the extended MLST schemes.

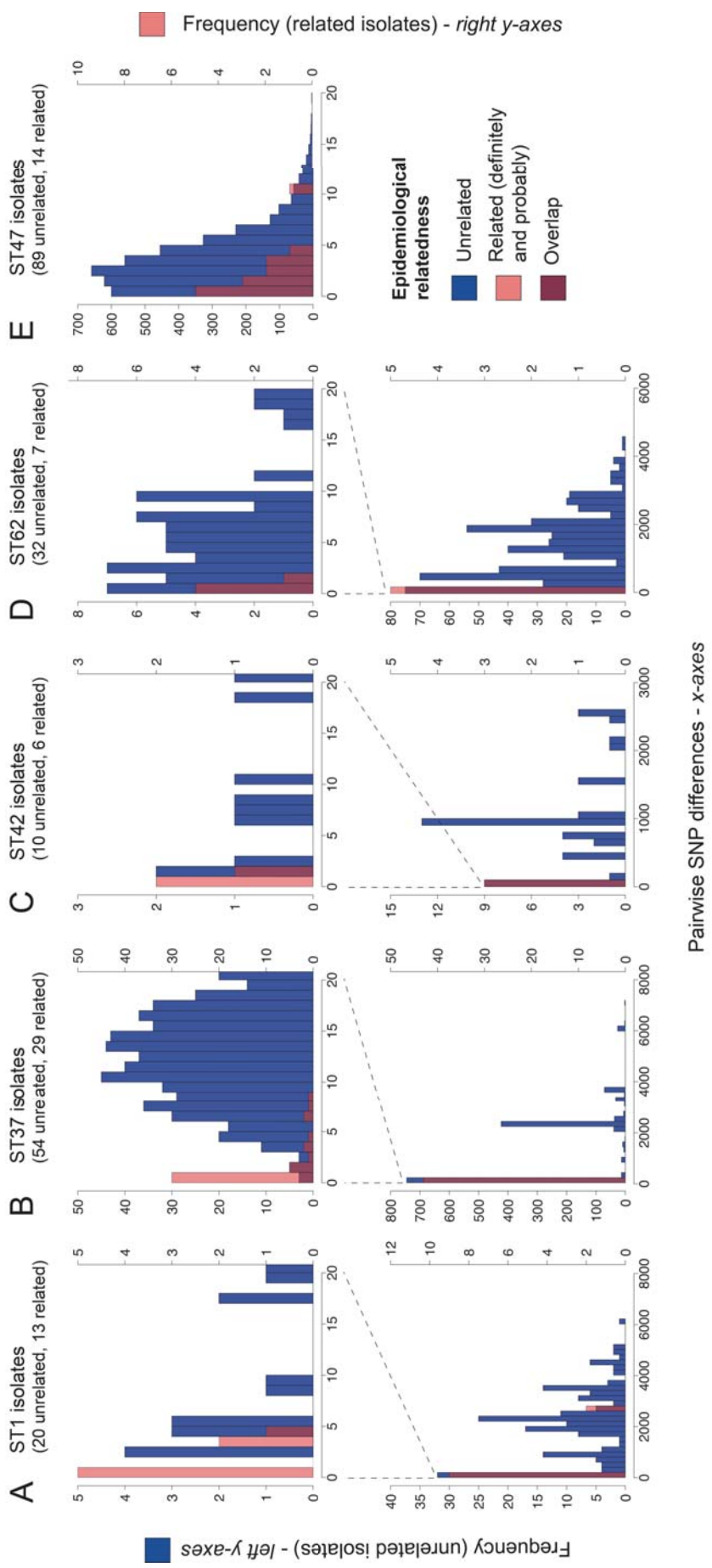
Frequency (related isolates) - right y-axes
A
ST1 isolates
(20 unrelated, 13 related)
B
ST37 isolates
(54 unrelated, 29 related)
C
ST42 isolates
(10 unrelated, 6 related)
D
ST62 isolates
(32 unrelated, 7 related)
E
ST47 isolates
(89 unrelated, 14 related)
Epidemiological relatedness
Unrelated
Related (definitely and probably)
Overlap
Pairwise SNP differences - x-axes
Frequency (unrelated isolates) - left y-axes

**Figure 5.4. Pairwise SNP differences between epidemiologically "unrelated" and "related" isolates belonging to some of the major disease-associated STs (A-E) (previous page).** In A-D, the top histogram shows pairwise SNP differences up to 20 only while the bottom figure presents the full range. The maximum pairwise SNP difference within the ST47 isolates is <20 SNPs and thus only one figure is shown (E). Left and right y-axes represent the frequency of epidemiologically "unrelated" and "related" isolates, respectively. The epidemiologically "unrelated" and "related" isolates are coloured as indicated in the key at the bottom right.

The results of this analysis are provided in **Table 5.5** and demonstrate that all WGS-based methods can differentiate further between "unrelated" isolates that belong to the same ST, as defined by SBT. However, in concordance with the findings of *Chapter 3*, even using the most discriminatory methods (e.g. the SNP/mapping-based method), some epidemiologically "unrelated" isolates were found to be very similar (e.g. <20 SNPs) and some even identical, as shown in **Figure 5.4**. This phenomenon is most notable in the ST47 lineage, which contains isolates recovered up to 20 years apart and from distant regions of the UK and France, and in which all isolates are less than 20 SNPs apart (**Figure 5.4**). The SNP-based phylogenetic tree of 74 ST37 isolates also shows that isolates are separated into three highly clonal groups, with epidemiologically "unrelated" isolates sometimes interspersed with "related" isolates (**Figure 5.5**).


Epidemiological relatedness
Epidemiologically unrelated isolates
Isolates from the same patient
Related clinical and environmental isolates
Related clinical isolates
Clinical isolates from a point-source outbreak (BBC, UK, 1988)
Clinical and environmental isolates from a point-source outbreak (Hereford, UK, 2003)
H113660550
H103260667
H113280076
H080820009
H101820190
LC6058
EUL00069
H101760092
LC6163
H081180019
H062660463
H120240685
H071880001
LC464
H101980130
EUL00073
EUL00079
EUL00078
0 SNPs
H114740454
H115040456
LC5755
H103820081
H092760433
H073900557
H113780240
H073060003
LC0512
H113340664
H113180118
1 SNP
LC6267
LC6268
LC5694
LC5738
LC0798
LC6293
LC5722
LC0583
LC0565
H061140013
EUL00134
LC0794
H104320293
EUL00132
EUL00133
EUL00131
1 SNP (related pair only)
2 SNPs (related set only)
H041380048
H041640791
4 SNPs
H102020414
H100940111
H042960010
H090320386
LC6788
H044260061
H085020185
LC6228
H093160422
H093140322
4 SNPs
H064240448
H083920177
H084140691
5 SNPs
H111580389
H084760449
H034980467
H034800427
EUL00166
EUL00165
H034680033
10 SNPs
LC1127
LC0795
LC0731
LC0782
LC0801
LC0763
LC0732
1 SNP
0.2

**Figure 5.5. Maximum likelihood tree of 74 ST37 isolates with isolates coloured by their epidemiological relatedness (previous page).** The total number of SNPs identified between isolates of each epidemiologically "related" set is indicated. The scale shows the number of SNPs per variable site.

**Table 5.5. Differentiation between isolates from major disease-associated STs.** The number of types that epidemiologically "unrelated" isolates from five STs (1, 23, 37, 42 and 62) are divided into are given, as well as the indices of discrimination achieved by each of the WGS-based methods. Isolates were classified as the same type if they shared no differences and a different type if they shared 1 or more differences, except using the kmer-based method where isolates were categorised into types using single-linkage clustering with a threshold equal to the maximum difference detected between sequencing replicates.

| **ST (no. of unrelated isolates)** | **Number of types/Index of discrimination (*D*)** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SNP-based** | **rMLST (53)** | **cgMLST (50)** | **cgMLST (100)** | **cgMLST (500)** | **cgMLST (1455)** | **cgMLST (1521)** | **Gene pres./ abs.** | **Kmer-based** |
| 1 (20) | 20/1 | 4/0.721 | 10/ 0.879 | 12/ 0.911 | 17/ 0.979 | 20/1.00 | 19/ 0.995 | 5/0.668 | 20/1 |
| 37 (54) | 53/ 0.999 | 9/0.473 | 10/ 0.368 | 12/ 0.426 | 36/ 0.909 | 50/ 0.997 | 49/ 0.999 | 13/ 0.592 | 54/1 |
| 42 (10) | 10/1 | 4/0.778 | 4/0.733 | 5/0.800 | 10/1.00 | 10/1.00 | 10/1.00 | 5/0.667 | 10/1 |
| 47 (89) | 66/ 0.958 | 1/0 | 2/0.022 | 2/0.022 | 18/ 0.365 | 41/ 0.857 | 40/ 0.848 | 5/0.229 | 89/1 |
| 62 (32) | 30/ 0.994 | 4/0.333 | 8/0.790 | 12/ 0.849 | 21/ 0.942 | 28/ 0.980 | 31/ 0.998 | 15/ 0.915 | 27/ 0.982 |

### 5.3.5 Stability

Finally, the stability of the WGS-based methods was tested using three sets of "definitely related" isolates that were recovered from the same patient. The first includes EUL 48

and EUL 56, which were recovered fifteen days apart from a legionellosis patient. The second set contains three isolates, one of which was recovered by direct plating from a sputum sample (EUL 71), the second of which was recovered using amoebal co-culture from a sputum sample (EUL 76), and the third of which was isolated from a faeces sample of the same patient (EUL 77). The remaining set contains three isolates (EUL 73, 78 and 79) picked from single colonies on a primary isolation plate. No differences were found between isolates belonging to two of these sets (EUL 48 and 56; EUL 73, 78 and 79) using all methods except the kmer-based method, which yielded dissimilarity scores that were very small, but greater than the scores observed between sequencing replicate pairs (**Table 5.4**). However, isolates from the remaining set (EUL 71, 76 and 77) were identical (thereby stable) using only the less discriminatory methods (rMLST, 50-gene cgMLST and the gene presence/absence method) (**Table 5.4**).

## 5.4 Discussion

Since a large proportion of legionellosis cases are caused by a small number of common STs, as defined by the current gold standard typing method (SBT), some outbreak investigations remain unresolved. An increasing number of public health laboratories are therefore turning to WGS, the cost of which has decreased substantially in recent years. Several studies have now shown the feasibility and added value of using WGS for the investigation of local legionellosis outbreaks (Reuter *et al.*, 2013; Levesque *et al.*, 2014; Graham *et al.*, 2014; McAdam *et al.*, 2014; Moran-Gilad *et al.*, 2015; Sanchez-Buso *et al.*, 2016). However, there is currently no standardised method that allows results from different laboratories to be compared. Thus, the aim of this thesis chapter was to evaluate and compare several WGS-based methods for the typing of *L. pneumophila* and to determine the most suitable approach for future development. The evaluation criteria used were those defined by the ESGEM (van Belkum *et al.*, 2007) and include typability, reproducibility, epidemiological concordance, discriminatory power and stability.

For each of the four WGS-based methods tested (SNP/mapping-based, extended MLST, gene presence/absence and kmer-based), specific criteria were defined that must be met for isolates to be deemed typable. These were primarily intended to reject isolates

with low quality sequence data, and thus could change on re-sequencing, although in some cases they could also be linked to the intrinsic properties of an isolate. Using the SNP-based and kmer-based methods, 98.8% and 99.7% isolates were considered typable. With the newly designed cgMLST schemes with 50 to 1455 core genes, it was found that the more genes included in the scheme, the fewer the isolates that yielded complete profiles. For example, 99.1% isolates were fully typable using the 50-gene cgMLST scheme, compared with 86.8% using the 1455-gene scheme. Furthermore, just 39.6% isolates were fully typable by the published 1521-gene cgMLST scheme, as well as only 41.5% with the newly designed gene presence/absence scheme. However, further analysis showed that, despite the overall low typability scores of some of these schemes, the vast majority of genes were typable in every isolate tested. A small minority that were untypable were often so in multiple isolates, suggesting a problem with the choice of gene rather than with a single isolate. Therefore, these results suggest that the typability of gene-based schemes could be further improved upon with the elimination of problematic loci that may be difficult to sequence or assemble. The results also demonstrated that sole dependence on the BIGSdb software to extract alleles from *de novo* assemblies and determine a type can commonly result in mis-classification. BIGSdb fails to recognise alleles containing "N"s or deletions in the middle of the gene, and unwittingly assigns an allele number to such cases. Validation of the alleles using mapping data, another step that is not currently part of standard practice or permissible using BIGSdb (due to the large size of raw sequence files), also highlighted loci with discrepancies that should not be used for defining a type.

All WGS-based methods tested were found to be highly reproducible based on the re-sequencing of six isolates from the same DNA and using the same sequencing and library preparation protocols at the same centre. Importantly though, reproducibility was highly dependent on the implementation of the robust QC filters used for each method, such as the detection of MLST alleles containing "N"s or deletions. Indeed, given high quality data and robust QC filters, high reproducibility is a major advantage of sequencing-based methods and another study reported average differences of ≤0.39 (SNPs and indels) between sequencing replicates (Salipante *et al.*, 2015). However, further studies are needed to test the reproducibility of the WGS-based methods when isolates are sequenced at different centres, using different technologies (e.g. with

sequence data from the PacBio RSII and the MinION), different library preparation methods, and at different times (e.g. after prolonged storage or passaging).

Using the 79 epidemiologically "unrelated" isolates from the typing panel, the WGS-based methods all demonstrated greater discriminatory power than the current gold standard typing method, SBT, as well as the combination of SBT and mAb subgrouping, which is also frequently used. However, the indices of discrimination achieved by rMLST ($D$=0.972) and the gene presence/absence scheme ($D$=0.976) were only slightly higher than that achieved by the combination of SBT and mAb subgrouping ($D$=0.968), and these methods therefore provided minimal gains. The discriminatory power of the 50-gene cgMLST scheme was substantially higher ($D$=0.990) and the 100-gene scheme only improved upon this slightly ($D$=0.991). Meanwhile, almost total differentiation was achieved using the cgMLST scheme with 500 or more genes, and the SNP-based and kmer-based methods.

As expected, a trade-off between discriminatory power and epidemiological concordance was observed. The rMLST and gene presence/absence schemes, which achieved the lowest discrimination of the WGS-based methods, both demonstrated high epidemiological concordance. When isolates were classified into different types if they possessed one or more differences, all "definitely related" and "probably related" isolates from the typing panel and isolates from well-defined point source outbreaks (e.g. the BBC, Barrow and Hereford outbreaks) were classified into the same type. However, in addition to their greater discriminatory power, the 50-gene and 100-gene cgMLST schemes also achieved acceptable levels of epidemiological concordance, although the 100-gene scheme differentiated between isolates in the "definitely related" set of EUL 71, 76 and 77 due to the presence of a single SNP. Meanwhile, when one or more differences between isolates yielded different types, the more discriminatory WGS-based methods such as the 500-gene, 1455-gene and 1521-gene cgMLST schemes, and the SNP-based and kmer-based methods, showed poor epidemiological concordance as many "definitely" and "probably related" isolates could be distinguished between. These methods must therefore be used with a threshold that specifies the number of differences allowed between isolates of the same type. Thus, thresholds that maintained the epidemiological concordance of at least the "definitely related" sets were tested and

allowed for one SNP difference using the SNP/mapping-based method, one allele difference using the cgMLST schemes with 100 or more genes, and a Jaccard distance of 0.065 using the kmer-based method. Since many "probably related" isolates could still be differentiated between, these cut-offs would likely need increasing further although further work would be required to determine a suitable level. Up to 15 SNPs were described between clinical and environmental isolates from a point-source outbreak (Reuter *et al.*, 2013) while the authors of the study in which the 1521-gene cgMLST scheme was designed and tested suggested a cut-off of four allele differences (Moran-Gilad *et al.*, 2015). However, thresholds must also be used in conjunction with a clustering algorithm such as the single linkage clustering method implemented in this chapter. Whilst this is quite feasible with a pre-defined number of isolates, clustering becomes problematic when isolates are continuously being added to a collection. Each time an isolate is added, the clustering would likely need to be re-run and could change the groupings. Clustering could also mis-represent relationships by drawing arbitrary boundaries between groups of isolates and, for example, isolates on the edge of a group could possess fewer differences to those in another group than to those in their own group. Thus, it is more appropriate to assign types using a less discriminatory method that does not require the use of thresholds and clustering to maintain high epidemiological concordance.

Finally, it is important to ensure that a new WGS-based typing scheme for *L. pneumophila* could maintain backwards compatibility with the current gold standard method, SBT. This is firstly because many laboratories will likely lack the capacity to perform WGS on any or all of their isolates for several years. Ideally though, it should be possible to compare results from the laboratories that continue to use SBT with others that replace the use of SBT with WGS. Secondly, it is not always possible to culture *L. pneumophila* and therefore perform WGS, most usually due to contamination of the sample with background microbiota. There is a nested-PCR-based protocol, however, that allows SBT to be performed directly from clinical samples (http://bioinformatics.phe.org.uk/legionella/legionella_sbt/php/protocols/ESGLI%20NESTED%20SBT%20GUIDELINE%20v2.0.pdf), which could be used when WGS is not possible. Thus, in order to maintain backwards compatibility, the seven SBT alleles should be determined from the WGS data, regardless of the primary WGS-based typing

procedure. Whilst this is possible for six of the seven SBT genes, the presence of multiple copies of the *mompS* gene that are occasionally different means that it is not always possible to correctly determine the allele number using short-read data. Thus, until this problem is resolved with long-read sequencing technology, it may be necessary to perform PCR and Sanger sequencing of the *mompS* gene.

Overall, the analyses presented in this chapter suggest that the most appropriate typing scheme for *L. pneumophila* is a 50-gene cgMLST scheme since it substantially improves upon the discrimination offered by current methods whilst maintaining high epidemiological concordance without the requirement for thresholds and clustering methods. The relatively low number of genes also decreases the possibility that an isolate will contain an untypable gene and thereby lose the ability to be assigned a type. However, in order not to lose the large amount of information provided by WGS, the 50-gene scheme could also be used as part of a larger, hierarchical scheme comprised of the 7 SBT genes, and increasing numbers of core genes (e.g. 50, 100, 500 and ~1500). Whilst some differences between "related" isolates would be expected when defining types using the larger numbers of genes, this approach would allow the extremely high discriminatory power of such schemes to be exploited when needed. They could also be very useful for differentiating between isolates belonging to highly clonal STs such as ST47.

An ESGLI working group comprising representatives from more than ten national reference laboratories for *L. pneumophila* has been established with the aim of designing and implementing a cgMLST-based scheme. Novel gene sets are being selected based on criteria calculated in this study such as discriminatory power and typability as well as other factors such as gene size and the genomic position. The development of a working scheme will also require the establishment of a central database, similar to the current SBT database, which assigns allele numbers and types to sequence data, and which can be searched by members of the research and public health community. The final result should be a standardised and portable scheme that can resolve a higher proportion of legionellosis outbreaks than SBT, and which has the potential to become the new gold standard typing method.