# 7. Conclusions and future directions

## 7.1 A restatement of the research questions and aims

*L. pneumophila* is an environmental bacterium and is thought to "accidentally" infect humans when the opportunity arises. Human infection usually occurs by inhalation of contaminated aerosols produced by man-made water systems (Muder *et al.*, 1986). Analysis of SBT data revealed that >40% of Legionnaires' disease cases in Europe are caused by just five STs, although >2000 STs have now been reported to the SBT database. Intriguingly, four of these five STs are only rarely found in commonly expected sources of *L. pneumophila* (Harrison *et al.*, 2009). The geographical distribution of these STs ranges from being very restricted (ST47 in North West Europe) to global (ST1). Prior to this study, it was not understood when these STs emerged, nor how rapidly they have spread across countries and continents. Thus the first major aim of this thesis was to use the high resolution of WGS to understand their evolution, emergence and spread. Signs of convergent evolution were also searched for that could explain their predominance in human disease. Since recombination was found to account for almost all the diversity observed within some lineages, the second results chapter aimed to characterise the details of this process (in particular, of homologous recombination) and further understand its biological impact.

Due to the high prevalence of some STs in clinical infections, some outbreak investigations can go unresolved. In addition to its power in evolutionary studies, WGS also offers a highly promising typing tool due to its extremely high resolution. Furthermore, recent decreases in its cost and turnaround time now make it the typing tool of choice in some laboratories. While several studies have demonstrated the feasibility of using WGS for investigating community outbreaks of Legionnaires' disease, there is currently no WGS-based typing scheme described that would allow comparison of results from different laboratories. This is critical given the high proportion of travel-associated cases (ECDC, 2013). Thus, the aim of the third chapter was to compare different WGS-based typing methodologies and propose the optimal method for future

development and implementation. Finally, the last chapter explored whether WGS could be successfully used in nosocomial-based investigations of Legionnaires' disease, which had been explored in few studies prior to this thesis (Levesque *et al.*, 2014; Bartley *et al.*, 2016).

## 7.2 Key findings and future directions

### 7.2.1 Five major disease-associated STs have emerged recently and spread rapidly

By analysing multiple isolates belonging to each of five major disease-associated STs (1, 23, 37, 47 and 62), it was found that the five STs have emerged both recently and independently within the context of the *L. pneumophila* species. In each of the STs, isolates from different countries (and in the case of ST1, different continents) were found to often possess very few SNP differences, in contrast to the high number of SNPs found within the *L. pneumophila* species. This suggests that they have spread recently and relatively rapidly in the context of *L. pneumophila* evolution. The finding that ST47 isolates, which account for ~25% of Legionnaires' disease cases in North West Europe (Harrison *et al.*, 2009; Vekens *et al.*, 2012; Euser *et al.*, 2013), differ by a maximum of 19 SNPs, was particularly remarkable. The findings strongly challenge the idea that humans are "accidentally" infected by any strain that happens to be present in an environment. Instead, they suggest that disease cases predominantly arise by infection with specific clones that are more efficient at human infection. The mechanism by which these *L. pneumophila* clones are spreading is unknown, but the possibility of transmission *via* humans was raised. A region comprising genes that are highly similar in the five STs was also identified, which could be contributing to their increased disease propensity.

Given the importance of these five STs in human disease, future studies are required to identify their environmental niche in order to minimise human exposure (particularly that of STs 23, 37, 47 and 62, which are rarely found in the environment). Elucidating the mechanisms by which these clones are spreading should also be a priority. Finally,

further genomics studies comparing larger collections of clinically important strains (such as these five STs, and others from different parts of the world) with environmental strains that never or rarely cause disease will also be crucial to further understand the genomic basis for increased disease propensity. These studies should explore diversity in both the core and accessory genomes, the latter of which has been little studied in this thesis.

### 7.2.2 Homologous recombination is a major driver of *L. pneumophila* evolution

The analysis of multiple major disease-associated STs revealed that >96% of SNPs had arisen from recombination events in some lineages. By disentangling homologous and non-homologous recombination (i.e. MGEs), it was subsequently found that the former accounts for 33-80% of SNPs in the affected lineages. Remarkably, while homologous recombination events have occurred far less frequently, they have brought in up to 94x as many SNPs as *de novo* mutations. These results have confirmed previous findings that homologous recombination plays a very important role in the evolution of *L. pneumophila* (Sanchez-Buso *et al.*, 2014). Numerous hotspots of homologous recombination were also identified which included outer membrane proteins, the LPS locus and Dot/Icm effectors, and these provide interesting clues to the selection pressures faced by *L. pneumophila*. Inference of the origin of the recombined regions showed that isolates have most frequently imported DNA from isolates belonging to their own clade, but also occasionally from other major clades of their subspecies (*L. pneumophila pneumophila*). Indeed, it was shown that the horizontal exchange of genes between the five disease-associated STs described in the first results chapter, which belong to different major clades of the subspecies, was likely a critical factor in their emergence. However, acquisition of recombined regions from another subspecies, *L. pneumophila fraseri*, was rarely observed, suggesting the existence of a recombination barrier and/or the possibility of ongoing speciation between the two subspecies.

Future work could use larger genomic data sets to further explore the recombination hotspots identified here. It was sometimes unclear which genes were driving the hotspots, particularly in lineages where only a small number of recombination events were detected. While there appeared to be some differences in hotspot regions between

lineages, further exploration of these could shed light on differences in infection strategies, host cells or environmental niches.

### 7.2.3 A 50-gene cgMLST scheme is suggested as the optimal WGS-based method for *L. pneumophila* typing

In order to determine the optimal WGS-based approach for *L. pneumophila* typing, various methods were tested using published criteria, which included typability, reproducibility, epidemiological concordance, discriminatory power and stability (van Belkum *et al.*, 2007). Overall, it was suggested that a 50-gene cgMLST scheme would be the most suitable method for future development since it substantially improves upon discrimination achieved by current methods whilst maintaining good epidemiological concordance. However, in order to not lose the large amount of information provided by WGS, the 50-gene scheme could also form part of a larger, hierarchical scheme comprising 50, 100, 500 and ~1500 genes. An ESGLI working group has now been set up to develop and implement this suggested scheme.

A number of challenges lie ahead in the development of this scheme. The first is that the new typing scheme should maintain backwards compatibility with SBT. However, one of the SBT genes, *mompS*, is present in multiple copies and it is currently not always possible to determine the correct *mompS* allele using short-read data (Moran-Gilad *et al.*, 2015). Another major challenge will be ensuring that all alleles are called correctly from the currently imperfect assemblies produced from short-read Illumina data. In this thesis, it was found that alleles are occasionally incorrectly called when only the *de novo* assemblies are used, even when the sequence data is deemed to be of high quality. Yet the files that contain the raw sequence reads (that can be used successfully to confirm or refute the alleles) are large and it is difficult to incorporate these into a web-based pipeline. More generally, data from different sequencing centres is currently of highly variable quality, and robust QC measures must be put in place to ensure high accuracy and reproducibility.

### 7.2.4 WGS can be used to successfully confirm or refute links between Legionnaires' disease cases and hospitals

The last chapter showed that WGS could be used successfully to confirm or refute suspected links between Legionnaires' disease cases (caused by ST1) and hospitals, as was demonstrated in a previous, albeit smaller, study (Bartley *et al.*, 2016). This was facilitated by the presence of distinct populations of *L. pneumophila* in several hospitals, rather than the existence of a complex mixture. However, it was revealed that, in order to confirm or refute a suspected link in future WGS-based investigations, deep environmental sampling would be required. This is firstly because the strains found within hospital water systems were often found to be highly similar to epidemiologically unrelated isolates sampled from the local area around the hospital (e.g. the patients' homes). Secondly, despite the presence of distinct hospital populations, substantial diversity was found within some of these populations. The combination of these two factors means that isolates from the same hospital water supply often have more SNP differences than epidemiologically unrelated isolates separated by geographical location. Thus, without deep sampling and an understanding of the hospital diversity within the context of the local diversity, spurious links could be made on the basis of SNP differences alone. Much stronger evidence of a link comes from the discovery that a clinical isolate is nested within, and thus derived from, a clade of hospital water isolates, in addition to the detection of a low number of SNP differences.

Analysis of a large number of ST1 isolates in the final results chapter also confirmed the previous findings from this thesis that the lineage possesses limited diversity in terms of *de novo* mutations, and that its international spread has occurred over a relatively short time frame within the context of *L. pneumophila* evolution. The finding that ST1 isolates from multiple hospital water systems are very closely related or even identical, despite being sampled several years apart, also reinforced earlier findings that *L. pneumophila* has a very slow mutation rate. It could also be suggestive of a dormancy phase. Overall, the interpretation of WGS data in future investigations would benefit from a deeper understanding of the speed and mechanism by which *L. pneumophila* spreads both locally and globally, and a greater insight into the evolutionary rate and potential dormancy of this bacterium.

## 7.3 Closing remarks

This thesis has demonstrated how WGS can be used to understand the evolution and spread of important bacterial pathogens such as *L. pneumophila*. It has also explored how WGS can be used in a clinical setting for the detection and resolution of outbreaks, and revealed some of the challenges faced in the interpretation of WGS data from a slow-evolving bacterium such as *L. pneumophila*.