6. Application of WGS to nosocomial investigations of Legionnaires' disease

Declaration of work contributions

Julian Parkhill and Timothy Harrison supervised this work. Collaborators at PHE (UK) and the Reference Center for *Legionella* (France) provided samples, typing data and epidemiological data. Baharak Afshar, Massimo Mentasti and Christophe Ginevra performed the culture and DNA extraction of all newly sequenced isolates. Simon Harris, Sophie Jarraud and Victoria Chalker provided valuable advice. I conducted the bioinformatics analyses, interpreted the data and generated the figures.

Publication

The following work has been accepted for publication:

David, S., Afshar, B., Mentasti, M., Ginevra, C., Podglajen, I., Harris, S. R., Chalker, V. J., Jarraud, S., Harrison, T. G. & Parkhill, J. Seeding and establishment of *Legionella pneumophila* in hospitals; implications for genomic investigations of nosocomial Legionnaires' disease. *Clinical Infectious Diseases* [in press]

6.1 Introduction

While the majority of cases are community-acquired, Legionnaires' disease is also recognised as an important cause of hospital-acquired pneumonia (Lin *et al.*, 2011). Nosocomial cases have been reported from many hospitals around the world and occur both sporadically and as part of outbreaks (Cordes *et al.*, 1981; Arnow *et al.*, 1982; Graman *et al.,* 1997; Kool *et al.*, 1998; Palmore *et al.*, 2009). Most nosocomial cases are linked to the inhalation or aspiration of contaminated drinking water (Blatt *et al.*, 1993) although sources such as decorative fountains, humidifiers and cooling towers have also been implicated (Palmore *et al.*, 2009; Bou & Ramos, 2009; Yiallouros *et al.*, 2013; Osawa *et al.*, 2014). Elderly and immunocompromised patients, or those with underlying conditions, are most at-risk of infection and have the highest mortality rate once infected (Guiguet *et al.*, 1987).

The frequent colonisation of hospital water systems with *Legionella* is often attributed to the large and complex pipe networks in which it can be difficult to maintain sufficient water temperatures to successfully control the bacteria (Orsi *et al.*, 2014). The extensive network of pipe surfaces is also prone to the accumulation of biofilms that promote the growth of *Legionella*. It is recognised that, once colonised, it can be extremely difficult to eradicate *Legionella* from a water system (Rangel-Frausto *et al.*, 1999; Borella *et al.*, 2005; Cristino *et al.*, 2012). Thus the strategy for preventing Legionnaires' disease cases in a hospital or elsewhere is focused on controlling the bacteria so that they are present only at very low concentrations. In addition to water temperature regulation, other control strategies have been used with varying success including copper-silver ionisation, water chlorination, point-of-use filtration and UV irradiation (Lin *et al.*, 2011).

As a result of the difficulties in controlling *Legionella*, there have been an increasing number of reports of long-term colonisation of hospital water systems, often with persistence of the same strain (Lepine *et al.*, 1998; Rangel-Frausto *et al.,* 1999; Perola *et al.*, 2005; Pancer *et al.*, 2013). In particular, ST1 has been shown to colonise several hospitals worldwide and has often been implicated as the cause of nosocomial Legionnaires' disease (Reimer *et al.*, 2010; Pancer *et al.*, 2013; Cassier *et al.*, 2015).

However, since ST1 isolates are detected commonly in environmental sources, both within hospitals and elsewhere (Harrison *et al.*, 2009; Kozak-Muiznieks *et al.*, 2014; Cassier *et al.*, 2015), the source of infection in possible nosocomial cases is often unresolved with SBT. Recently, a method of subtyping of ST1 isolates using spoligotyping has been developed that, with a reported index of discrimination of 79.7%, can be a useful complementary genotyping tool for discriminating ST1 isolates (Ginevra *et al.*, 2012; Gomgnimbou *et al.*, 2014). Nevertheless, even with a combinatory approach, some investigations still remain inconclusive.

This thesis chapter uses WGS, which was demonstrated in *Chapter 5* to provide substantially higher resolution than current typing methods, to examine suspected links between multiple hospital water systems and cases of Legionnaires' disease caused by ST1. In particular, a detailed investigation is performed of seven cases associated with an anonymous hospital, Hospital A (Essex, UK), which occurred between 2007 and 2011. Deep environmental sampling of this hospital allowed comparison with another previously studied and deeply sampled hospital, The Wesley Hospital/Hospital B (Queensland, Australia), that was found to be colonised by a single, although surprisingly diverse, population of ST1 using WGS (although the study did not describe the strain as ST1) (Bartley *et al.*, 2016). It also aims to understand the evolutionary context and the similarity of hospital populations within the global phylogeny of ST1, and finally to assess the implications of these results for future WGS-based investigations of nosocomial-associated infections.

6.2 Materials & Methods

6.2.1 Bacterial isolates

WGS data from an internationally sampled collection of 229 ST1 or ST1-derived isolates were used in this study (**Appendix Table 26**). These include 81 used in *Chapters 3 & 4*, 91 that are newly sequenced for this study and 57 that have been published in other studies. ST1-derived isolates refer to isolates of other STs that have been previously

shown to be closely related to, and to be evolved from, ST1 isolates (see *Chapter 3*). The collection includes 99 environmental isolates from the water systems of 17 hospitals spanning five countries (UK, France, Spain, Denmark, Australia). Multiple environmental isolates were obtained from five of these hospitals (Hospital A, *n*=38; The Wesley Hospital/Hospital B, *n*=39; Hospital C, *n*=5; Hospital D, *n*=3; Hospital E, *n*=2), while a single environmental isolate was obtained from the remaining 12 hospitals. Forty-two clinical isolates from patients with confirmed or suspected links to 20 different hospitals, including ten hospitals from which we also obtained one or more environmental isolates, were also included. Of the remaining 88 isolates in the collection, 47 are from or associated with community-acquired sources of Legionnaires' disease (i.e. non-hospital related), three were sampled from a cruise ship, while the sampling context of 38 isolates is unknown. Culture and DNA extraction of all isolates was performed as described in *Chapter 2 (Materials & Methods)*.

6.2.2 Whole genome sequencing

Isolates were sequenced by the core sequencing facilities at PHE using the Illumina HiSeq platform with 100bp paired-end reads or at the WTSI using the Illumina MiSeq platform with 150bp paired-end reads. Library construction was performed as described in *Chapter 2 (Materials & Methods)*. Raw reads for all newly sequenced isolates were deposited in the ENA under the study accession numbers ERP003631 and ERP015468, and individual run accession numbers are provided in **Appendix Table 26**.

6.2.3 Mapping of sequence reads and phylogenetic analyses

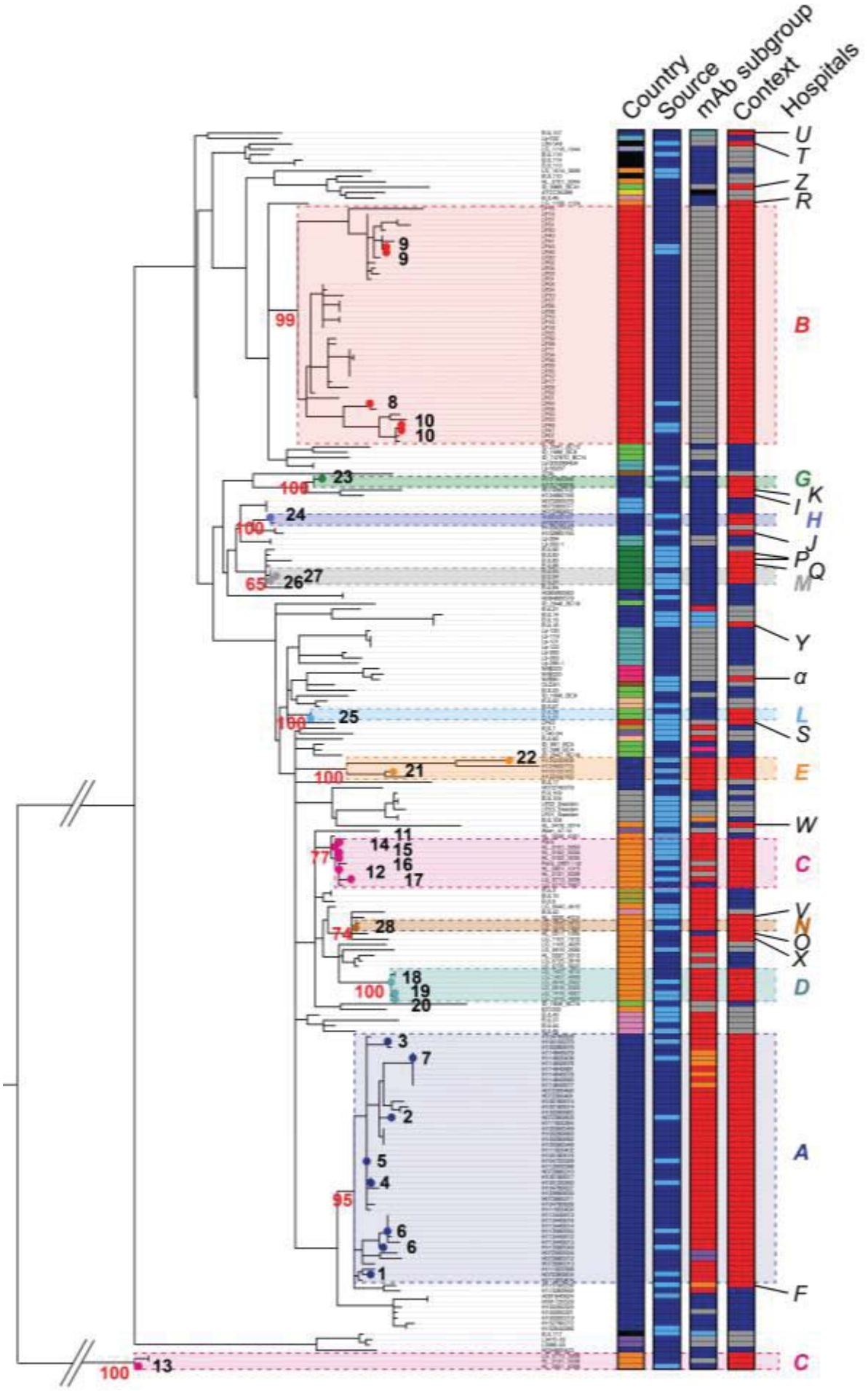
Sequence reads were mapped to the Paris (ST1) reference genome (Cazalet *et al.*, 2004) using SMALT v0.7.4 (available from: http://www.sanger.ac.uk/science/tools/smalt-0) and bases were called as described in *Chapter 2 (Materials & Methods)*. Recombined regions were identified and removed from the alignment using Gubbins (Croucher *et al.*, 2015). A maximum likelihood tree was generated using the variable sites that remained as described in *Chapter 2 (Materials & Methods)*.

6.3 Results

6.3.1 Hospital populations comprise distinct lineages of *L. pneumophila* ST1

The phylogenetic context of 99 environmental isolates sampled from the water systems of 17 hospitals, together with 42 clinical isolates from Legionnaires' disease patients with confirmed or suspected hospital-acquired infections, was investigated within an internationally sampled collection of 229 *L. pneumophila* ST1 or ST1-derived genomes (**Appendix Table 26**). To construct a phylogenetic tree, sequence reads were first mapped to the complete genome of the Paris strain (an ST1) (Cazalet *et al.*, 2004) and a total of 62,395 SNPs were identified amongst all isolates. Since recombination has been previously shown to account for a large proportion of the diversity within single STs, including ST1 (in *Chapter 3*), Gubbins was used to identify and remove regions from the genome alignment that have been affected by recombination. A total of 382 putative recombined regions, containing 97.2% of the total SNPs (but affecting an average (mean) of just 5.1% of each genome (range, 0.85-14.5%)), were identified and removed. The remaining 1,741 SNPs, representing only those that have arisen via *de novo* mutation, were used to construct a phylogenetic tree (**Figure 6.1**). Numbers of SNP differences between isolates that are provided from here on represent only those that have arisen via *de novo* mutation and exclude those in recombined regions, unless stated otherwise.

Using the phylogenetic tree, it was first investigated whether five hospitals (A-E) from which multiple ST1 isolates were obtained have been colonised by distinct or mixed ST1 populations. **Figure 6.1** shows that the 38 environmental isolates sampled from the water system of Hospital A (Essex, UK) between 2007 and 2012 indeed cluster together, demonstrating the existence of a single ST1 population. Re-analysis of the 39 isolates obtained from the water supply of The Wesley Hospital/Hospital B (Queensland, Australia) in 2013 with the wider collection of ST1 isolates supports previous findings that this hospital has also been colonised by a distinct ST1 population (Bartley *et al.*, 2016). Similarly, isolates from the water supply of Hospital D (near Marseille, France) (n=3) and Hospital E (London, UK) (n=2) cluster together, although only small numbers of isolates were obtained.


Country
Source
mAb subgroup
Context
Hospitals
U
T
Z
R
B
G
K
I
H
J
P
Q
M
Y
α
L
S
E
W
C
V
N
O
X
D
A
F
C
9
9
8
10
10
23
24
26
27
25
22
21
11
14
15
16
12
17
28
18
19
20
3
7
2
5
4
6
6
1
13
99
100
100
65
100
77
74
100
95
100

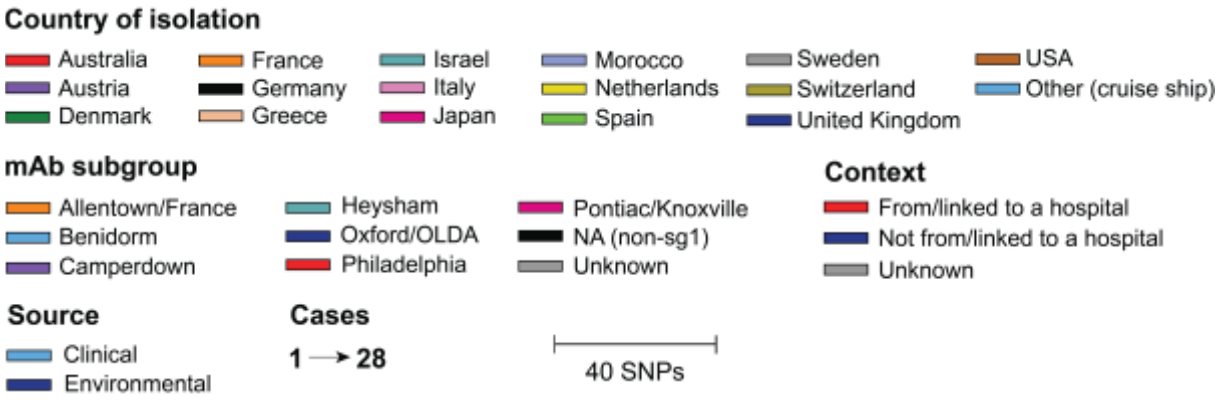


Figure 6.1. Maximum likelihood tree of 229 ST1 and ST1-derived isolates including those from or associated with hospitals (previous and current page). The tree was constructed using 1,741 SNPs identified after the removal of recombined regions. Environmental isolates from and clinical isolates linked to 27 different hospitals are included. Isolates from or potentially linked to the water systems of ten of these hospitals (from which at least one environmental isolate and one clinical isolate was obtained) are coloured within the tree itself. Clinical isolates from 28 suspected cases linked to these ten hospitals are indicated by small circles (coloured according to the hospital) and numbered within the tree. Clinical isolates obtained from the same patient have the same number. Bootstrap values obtained for nodes from which isolates from the ten hospitals are descended are shown in red.

Interestingly though, environmental isolates from Hospital C (Paris, France) (*n*=5) form two clusters, which differ by up to 300 SNPs, although each cluster comprises hospital isolates that are distinct from environmental isolates sampled elsewhere. This discovery of two distinct clusters is concordant with previous typing results obtained by spoligotyping (Gomgnimbou *et al.*, 2014). Both lineages were detected in 2000-2001 and 2007, demonstrating long-term co-existence of two ST1 populations within the hospital water system. Nevertheless, these results suggest that all five hospitals have been colonised by a limited number of distinct ST1 populations rather than a complex mixture. This is an important prerequisite for using WGS to support or refute the hospital acquisition of cases.

6.3.2 WGS can be used to support or refute links between Legionnaires' disease cases and hospital water systems

It was next investigated whether the WGS data supports the confirmed or suspected links between hospital water systems and Legionnaires' disease cases. In particular, a detailed examination was performed of seven Legionnaires' disease cases that occurred between 2007 and 2011 (**Figure 6.2**), all of which are considered to have been acquired from Hospital A. Another six cases with suspected links to the hospital also occurred between 2002 and 2010 but as no clinical isolates were obtained, further genomic investigation could not be performed. The links between the hospital and the seven cases for which clinical isolates were obtained were made on the basis of epidemiological information (**Table 6.1**) and using the molecular typing methods, SBT and mAb subgrouping. All clinical isolates, except one obtained from the most recent case (November, 2011) were typed as ST1, mAb subgroup Philadelphia, which is an uncommon strain in England (Harrison *et al.*, 2009). Isolates obtained from the hospital water supply shortly after each incident were also characterised as ST1, Philadelphia, which supported hospital acquisition. Meanwhile, the clinical isolate from the most recent case was typed as ST1, mAb subgroup Allentown/France, and environmental isolates of the same type were also obtained from the hospital water supply shortly after the incident, again supporting hospital acquisition. Here, the eight clinical isolates from these cases (two of which come from a single patient) were compared with the 38 environmental isolates sampled from the hospital water supply, within the context of the large collection of sequenced ST1/ST1-derived isolates. Importantly, the collection includes contemporary ST1 isolates from or associated with another seven hospitals (E-K) and community-acquired sources in the local area of London/East of England. Phylogenetic analyses show that all eight clinical isolates are nested within and thus derived from the clade of isolates sampled from the water supply of Hospital A (i.e. have evolved from the MRCA of the hospital isolates) (**Figures 6.1 and 6.3**). Assuming that the ST1 population in the hospital water supply has not spread out of the hospital to elsewhere (a scenario that has not been observed with any hospital in this study using phylogenetic evidence), this finding provides strong evidence that the infections were indeed acquired from the hospital (**Table 6.1**). Furthermore, each of the clinical isolates differ by just 0-4 SNPs from the closest hospital isolate, providing further supporting

evidence of hospital acquisition. Crucially, both of these findings were facilitated by the recovery and analysis of a large number of hospital isolates.

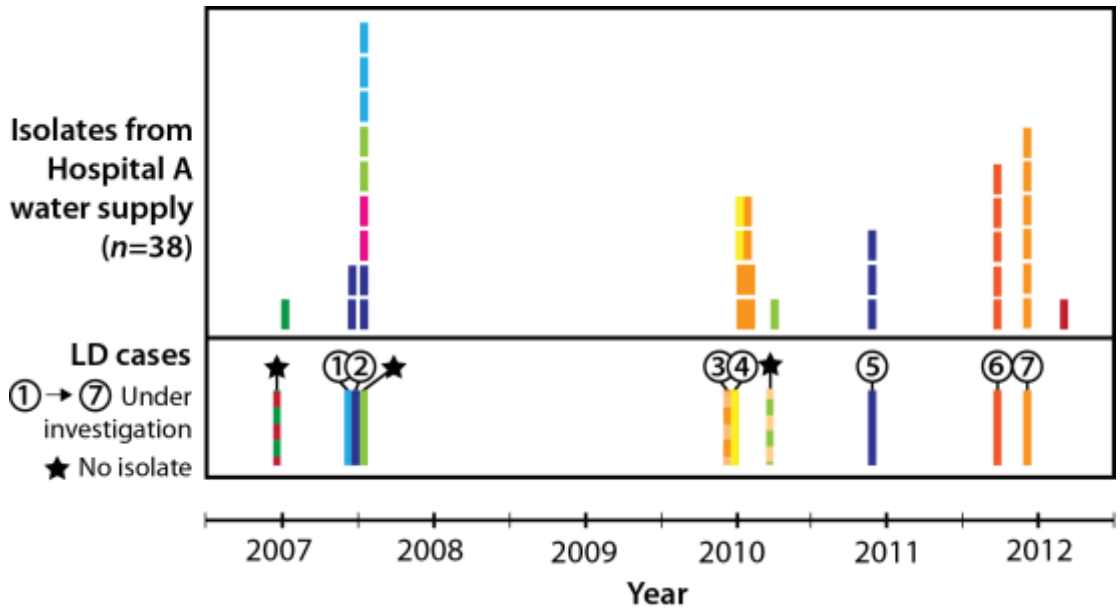


Figure 6.2. Time frame of legionellosis incidents and collection of environmental isolates at Hospital A. The time frame in which ten cases of Legionnaires' disease that were considered to have been acquired from Hospital A between the end of 2006 and 2011 is shown (bottom panel). Clinical isolates were obtained from seven of these cases, as indicated. Environmental isolates were also obtained between 2007 and 2012 from the hospital water supply, usually after each Legionnaires' disease incident (top panel). Isolates are coloured according to the hospital ward(s) in which the patient stayed (clinical isolates) or they were sampled from (environmental isolates).

Table 6.1. Genomic evidence to support 28 suspected links between hospital water systems and Legionnaires' disease cases, from which at least one hospital isolate and one clinical isolate was obtained and analysed using WGS. Different types of genomic evidence were categorised (A-D). A: The clinical isolate(s) is derived from the MRCA of the hospital isolates, and differs by <5 SNPs to the closest hospital water isolate. Strong evidence that the infection was hospital-acquired. B: The clinical isolate(s) is derived from the MRCA of the hospital isolates, but differs by >5 SNPs to the closest hospital water isolate. Good evidence that the infection was hospital-acquired. C: The clinical isolate(s) clusters most closely with hospital isolates, and is <5 SNPs different from the closest hospital isolate. However, the clinical isolate(s) is not derived from the MRCA of the sampled hospital isolates. Acquisition from elsewhere cannot be ruled out on the basis of genomic evidence alone. D: The clinical isolate(s) clusters most closely to and differs by <5 SNPs from the hospital isolate. However, the recovery of only one hospital isolate prevents the determination of whether the clinical isolate is derived from hospital isolates. Acquisition from elsewhere cannot be ruled out on the basis of genomic evidence alone.

Suspected hospital	**Date of incident**	**Known exposures during the incubation period (~18 days prior to onset of symptoms)**	**Clinical isolate(s)**	**Does the clinical isolate cluster most closely with a hospital water isolate? (no. of SNPs)***	**Genomic evidence**
Hospital A, Essex, UK	May 2007	Hospital A (11-18 days), home	H072360604 (case 1)	Yes (4 SNPs)	A
	May 2007	Hospital A (~12 days)	H072360603 (case 2)	Yes (3 SNPs)	A
	December 2009	Hospital A (~4 days), home and local area	H100120270 (case 3)	Yes (1 SNP)	A
	December 2009	Hospital A (~7 days), home and local area	H100120260 (case 4)	Yes (0 SNPs)	A
	November 2010	Hospital A (~7 days), home and local area	H104720329 (case 5)	Yes (0 SNPs)	A
	August 2011	Hospital A (at least 10 days)	H113580549, H113580550 (case 6)	Yes (3 and 0 SNPs, respectively)	A

	November 2011	Hospital A (at least 10 days)	H114820438 (case 7)	Yes (0 SNPs)	A
The Wesley Hospital/ Hospital B, Queensland, Australia	October 2011	Hospital B	LP44 (case 8)	Yes (1 SNP)	A
	May 2013	Hospital B only	LP45, LP46 (case 9)	Yes (both 1 SNP)	A
	June 2013	Hospital B only	LP47 and LP48 (case 10)	Yes (1 and 2 SNPs, respectively)	A
Hospital C, Paris, France	March 2002	Hospital C (13 days) & another hospital near Paris (5 days)	Paris (case 11)	Yes (2 SNPs different to isolate from Hospital C). No isolates obtained from the other hospital.	C (acquisition from other hospital cannot be ruled out)
	December 2000	Hospital C only	HL 0051 1015 (case 12)	Yes (0 SNPs)	C
	December 2000	Hospital C (~17 days)	HL 0051 4008 (case 13)	Yes (4 SNPs)	C
	December 2000	Hospital C (~12 days)	HL 0101 3003 (case 14)	Yes (1 SNP)	C
	December 2000	Hospital C (~4 days), home	HL 0102 3034 (case 15)	Yes (2 SNPs)	C
	December 2000	Hospital C (~4 days), home	HL 0102 3035 (case 16)	Yes (2 SNPs)	C
	March 2007	Hospital C only	LG 0713 5006 (case 17)	Yes (3 SNPs)	A
Hospital D, near Marseille, France	April 2009	Hospital D (~4 days), home	LG 0918 2002 (case 18)	Yes (0 SNPs)	C
	April 2014	Hospital D (~3 days), home (~3	LG 1416 4007 (case	Yes (1 SNP)	A

		days)	19)		
	April 2014	Hospital D (~5 days)	LG 1416 4008 (case 20)	Yes (1 SNP)	A
Hospital E, London, UK	June 2010	Hospital E (at least 10 days)	H103120165 (case 21)	Yes (7 SNPs)	B
	October 2012	Hospital E (less than 10 days)	H124240908 (case 22)	Yes (33 SNPs)	B
Hospital G, Cambridge-shire, UK	April 2010	Hospital G (less than 10 days)	H101460286 (case 23)	Yes (2 SNPs)	D
Hospital H, London, UK	June 2009	Hospital H (at least 10 days)	H092520167 (case 24)	Yes (1 SNP)	D
Hospital L, Cáceres Province, Spain	April 1994	Hospital L	EUL 55 (case 25)	Yes (0 SNPs)	D
Hospital M, Copenhagen, Denmark	October 1992	Hospital M only	EUL 93 (case 26)	Yes (0 SNPs)	D
	December 1992	Hospital M only	EUL 94 (case 27)	Yes (1 SNP)	D
Hospital N, near Marseille, France	April 2010	Hospital N only	LG 1019 1002 (case 28)	Yes (1 SNP)	D

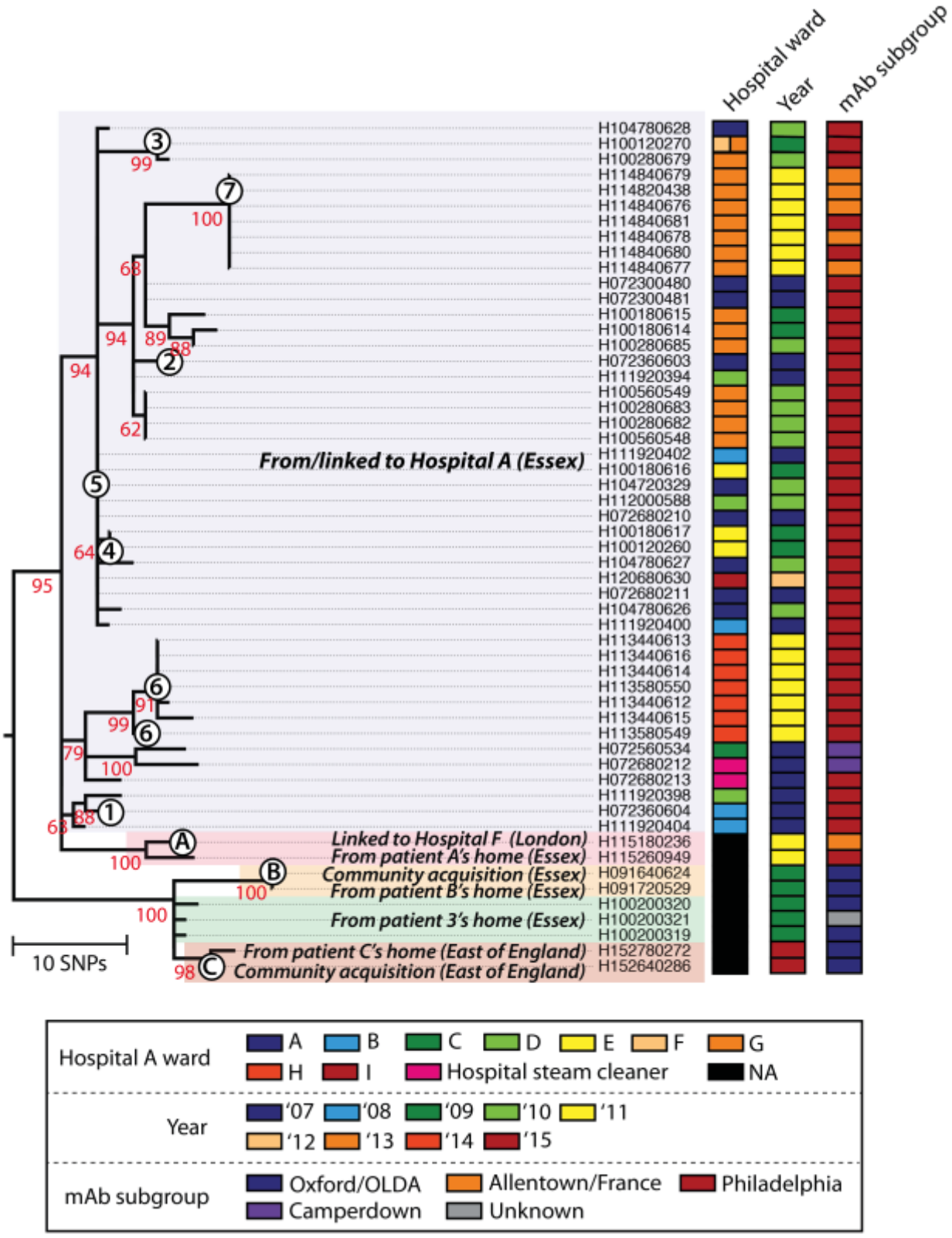


Figure 6.3. Phylogeny of isolates from Hospital A and the surrounding area. A zoomed-in section of the maximum likelihood tree presented in Figure 6.1 is shown, comprising environmental isolates from and clinical isolates linked to Hospital A. Clinical isolates from

seven cases linked to Hospital A are indicated by small circles and numbered 1-7 (two isolates were obtained from case 6). Closely related isolates sampled from nearby homes are also shown, including the home of a patient (case 3) who spent part of their incubation period in Hospital A as well as three homes of patients who had no epidemiological link to Hospital A. Clinical isolates from these latter three patients are indicated by small circles and labelled A-C. Clinical isolate A was obtained from a patient whose incubation period was spent both at home and in the Hospital F, while isolates B and C are from patients with no known epidemiological links to hospitals.

Interestingly, some isolates from or associated with community sources of *L. pneumophila* in the local area of London/East of England, as well as a clinical isolate from a patient who spent part of their incubation period in Hospital F (London, UK), also cluster closely with isolates from or associated with Hospital A (**Figure 6.3**). For example, just 13 SNPs were found between an isolate sampled from the water supply of Hospital A in 2007 (H111920404) and an isolate sampled in 2011 from the nearby home of a patient with no known epidemiological link to Hospital A (H115260949). Also closely related to the Hospital A isolates are three isolates (H100200319, H100200320, H100200321) obtained from the home of a patient (case 3) who spent their incubation period both at home and in Hospital A. The investigation at the time ruled out the home as a potential source since the mAb subgroup of two of the three home isolates was Oxford/OLDA rather than Philadelphia (unusually, the third home isolate did not react with any antibodies from the typing panel). WGS also supports this conclusion since the clinical isolate (H100120270) obtained from the patient is nested within the clade of hospital isolates and has just one SNP difference with the closest hospital isolate (H100280679), while it is 26 SNPs different from the closest home isolate. However, it is an important observation that the isolates from or associated with the hospital are so closely related to epidemiologically unrelated isolates from the local area.

Examination of other suspected links between cases and hospitals further demonstrated how the interpretation and strength of evidence obtained is highly dependent on both sampling and contextual information (**Table 6.1 and Figure 6.4**). For example, our phylogenetic analyses confirmed previous findings (Bartley *et al.*, 2016) that the three Legionnaires' disease cases associated with The Wesley Hospital/Hospital B (from

which five clinical isolates were obtained) were most likely acquired within the hospital since the clinical isolates are nested within, and thus derived from, the clade of hospital isolates and differ by just 1-2 SNPs from the closest hospital isolate (**Figure 6.1 and Table 6.1**). Similarly to the investigation of cases associated with Hospital A, the large number of hospital isolates obtained and analysed facilitated these findings. Furthermore, investigation of two cases associated with Hospital E (one in 2010, one in 2012) revealed that while the two clinical isolates each cluster most closely with a single environmental isolate obtained from the hospital water supply shortly after each incident, they differ by 7 and 33 SNPs, respectively, to these hospital isolates. If each pair (comprising one clinical and one contemporary environmental isolate) were analysed alone, an investigation might refute a link between the second case and the hospital due to the large number of SNP differences. However, phylogenetic analysis of both pairs, together with the large collection of ST1 isolates, shows that the four isolates cluster together and that both clinical isolates are derived from the MRCA of the two hospital isolates (which presumably was a hospital isolate itself unless the hospital has been seeded multiple times) (**Figure 6.4**). This provides good evidence to support the hospital acquisition of both infections. On the other hand, several links were investigated between cases where only one environmental isolate from the suspected hospital has been obtained (e.g. Hospital G [Cambridgeshire, UK], Hospital H [London, UK], Hospital L [Cáceres Province, Spain], Hospital M [Copenhagen, Denmark], Hospital N [near Marseille, France]) (**Table 6.1 and Figure 6.4**). In all such cases, the clinical isolates associated with a hospital are more closely related to the environmental isolate from the suspected hospital than from anywhere else, differing by just 0-2 SNPs. However, when only one environmental isolate is obtained, it is impossible to determine whether the clinical isolate is derived from hospital isolates, even if the isolates are very similar or even identical. The genomic basis to support each link is therefore based only upon genomic similarity, which is a weaker form of evidence, since epidemiologically unrelated isolates can also be very similar (particularly those from the same geographical region), as described in *Chapter 5*. This means that acquisition from elsewhere cannot be ruled out, except in the cases where the patient spent their entire incubation period in the hospital.

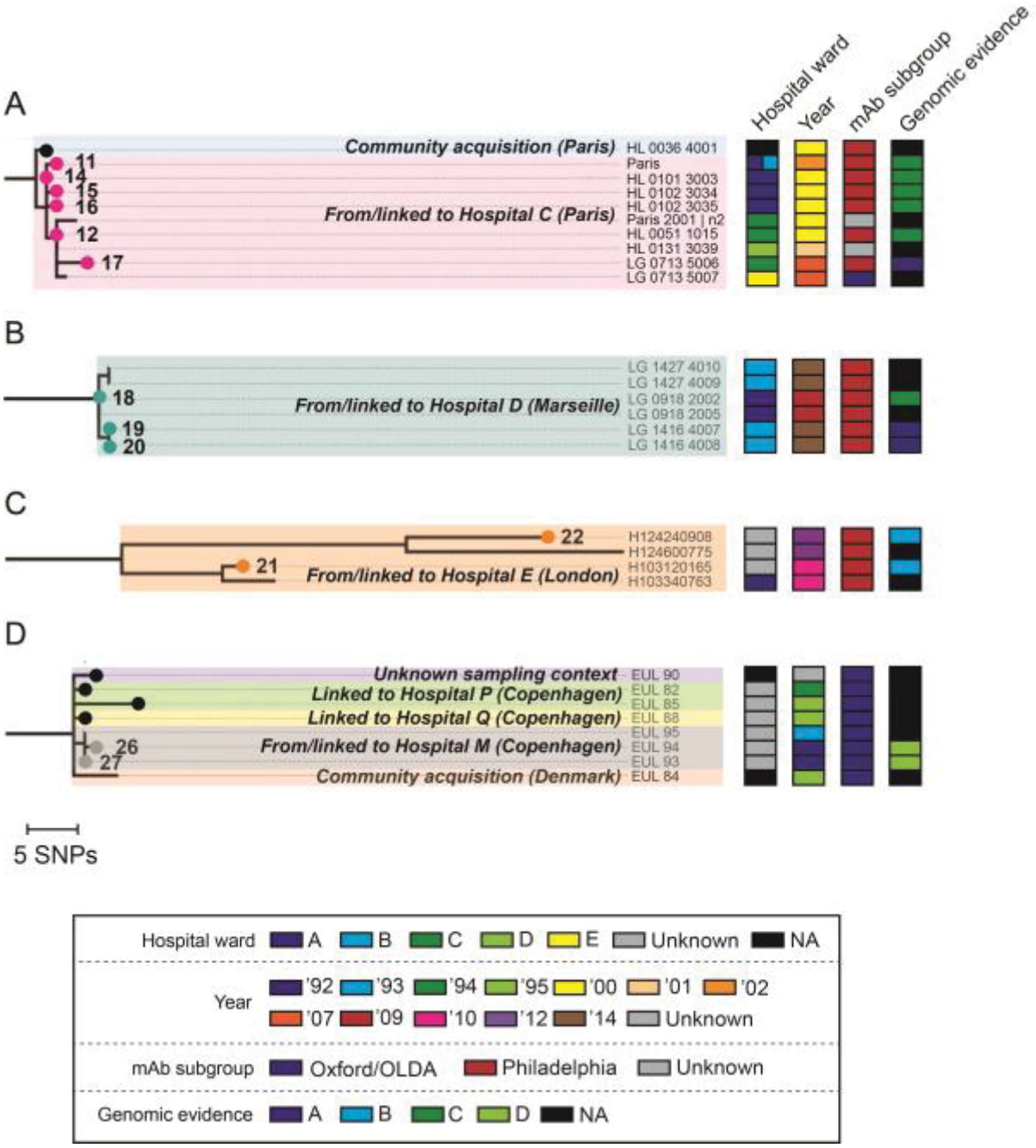


Figure 6.4. A-D) Zoomed-in sections of the maximum likelihood tree presented in Figure 1. All clinical isolates are indicated by small circles, with those from the 28 cases under investigation coloured and numbered as in Figure 1. Where applicable, isolates are additionally coloured in the right hand panel according to the hospital ward(s) in which the patient stayed (clinical isolates) or they were sampled (environmental isolates). Clinical isolates from the 28 cases under investigation are also coloured in the right hand panel by the strength of genomic evidence for hospital acquisition (see Table 1). NA – not applicable.

6.3.3 Substantial diversity within single hospital populations

Despite the colonisation of several hospitals with distinct ST1 populations, it is clear from both the previous study by Bartley *et al.* (2016) and the genomic analyses described here that considerable diversity exists within at least some of these lineages. For example, initial analysis of the ST1 diversity in the Hospital A water supply revealed a total of 1682 SNPs amongst 38 isolates. Gubbins detected the occurrence of seven putative recombination events within the hospital lineage (of which two are just 6bp and 41bp and likely the result of sequencing or mapping artefacts), which, once removed, leaves a total of 72 SNPs between the 38 isolates and a maximum difference of 25 SNPs between any pair. Interestingly, the five larger recombination regions (ranging in size from 1,442bp to 38,021bp) all occurred on the same branch of the phylogenetic tree, affecting the isolates, H072560534 and H072680212, and thus may have been acquired on the same occasion. In comparison, using the same methods, a total of 891 SNPs were identified amongst the 39 environmental isolates sampled from The Wesley Hospital/Hospital B, of which 746 were derived from two recombination events, leaving 145 SNPs generated by *de novo* mutation and a maximum difference of 44 SNPs between any pair. By comparison, between 6 and 339 SNPs were identified between environmental isolates sampled from different hospitals (N and O, and C and E, respectively). The detection of recombination events within the ST1 populations of both hospitals indicates the existence of other (probable non-ST1) *L. pneumophila* strains within each hospital water supply, assuming that the hospital populations have been restricted to the hospital water system and that the hospitals have not been re-seeded with newly recombined strains. Furthermore, a total of 60 SNPs generated by *de novo* mutation were detected between the two isolates sampled from Hospital E in 2010 and 2012, a higher number than that observed between any pair of isolates from either Hospital A or The Wesley Hospital/Hospital B. By contrast, very few pairwise differences (0-3 SNPs) were detected between isolates from the two lineages in Hospital C and one lineage in Hospital D, although only small numbers of environmental isolates were obtained.

As discussed previously, variation with respect to mAb subtypes was also detected within the population of Hospital A. Overall, 32 of 38 environmental isolates from

Hospital A and seven of the eight associated clinical isolates belong to the mAb subtype, Philadelphia (**Figure 6.3**). However, two closely related environmental isolates sampled from the hospital water supply in 2007, which are the same two isolates affected by the five recombination events, were typed as Camperdown. The genetic determinants of the mAb subtypes are not well understood but are presumably located within the LPS locus. Thus, we predict that one of the recombination events that spans the LPS locus, ranging from 923,274bp (*lpp0825*) to 931,183bp (*lpp0831*) with respect to the Paris reference genome, and which introduces a total of 107 SNPs, is the cause of the mAb switch. Intriguingly though, the one clinical isolate and four environmental isolates sampled in 2011 and characterised as mAb subtype, Allentown/France, cluster together in the phylogenetic tree along with two isolates typed as Philadelphia (**Figure 6.3**). No SNPs were identified between all seven isolates, both before and after the removal of recombined regions. Other differences that could explain the differing mAb subtypes were searched for including insertions, deletions and differences in gene content. The only observed difference affecting the LPS locus was a single insertion of a thymine base at 935,649 (which cases a frameshift about 80% through *lpp0835*) in the five Allentown/France isolates, but not the two Philadelphia isolates, and is thus the likely cause of the mAb switch.

6.3.4 Evidence for local microevolution within hospital populations

Given the substantial level of diversity observed amongst isolates sampled from Hospital A, it was explored whether isolates clustered by ward or location in the hospital (**Figure 6.5**), as was shown previously to be the case in The Wesley Hospital/Hospital B (Bartley *et al.*, 2016). **Figure 6.3** shows that there is some clustering by ward and that seven of the eight clinical isolates are most similar to one or more contemporary environmental isolates sampled from the same ward in which the patient was a resident. For example, all five environmental isolates sampled from various outlets in ward H in 2011 cluster together, differing by 0-4 SNPs, and also cluster with two clinical isolates (H113580549, H113580550) obtained from the post-mortem lung tissue of a patient (case 6) who stayed in the same ward. Another example is the clinical isolate, H100120260, obtained from a patient (case 4) who stayed in ward E, which has no SNP differences with an environmental isolate, H100180617, sampled from a shower in the same ward. The one

clinical isolate (H072360603) that is not most similar to an environmental isolate from the same ward in which the patient (case 2) stayed (ward A) nevertheless differs by just 4 SNPs from contemporary isolates from the same ward (H072300480 and H072300481).

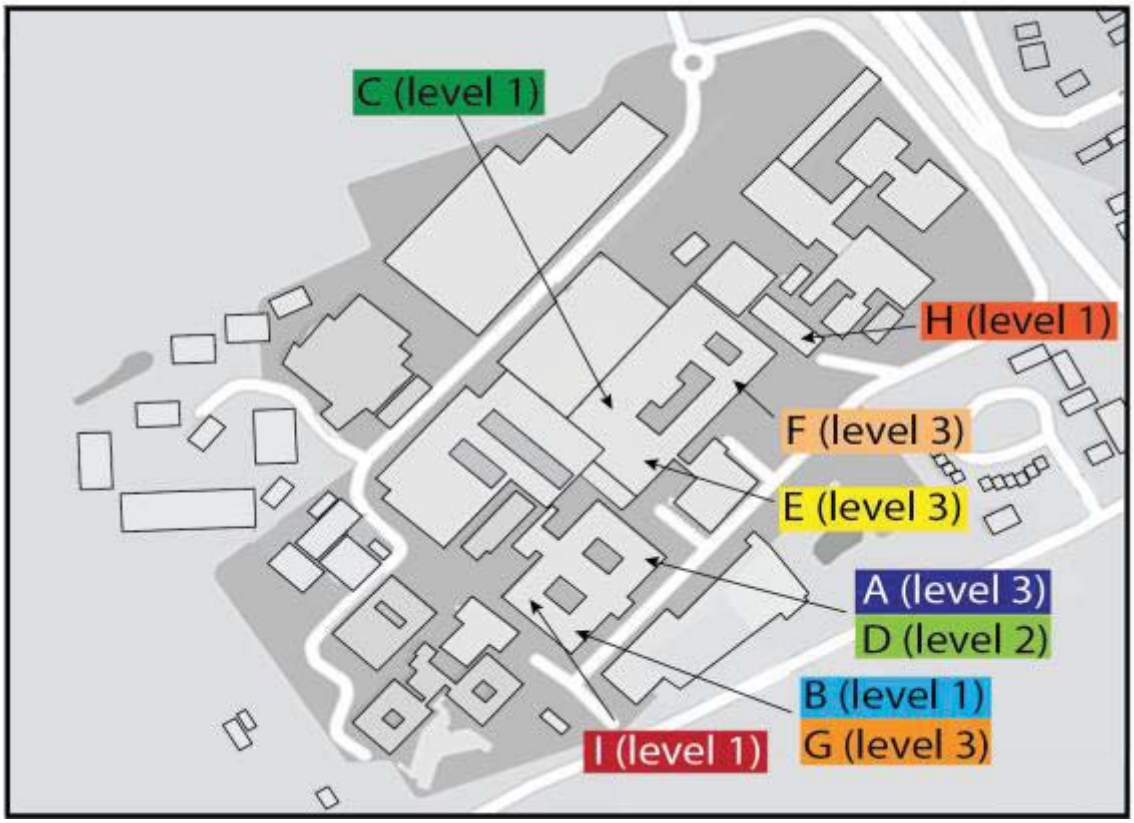


Figure 6.5. A plan of Hospital A. The wards in which the patients stayed are shown, as well as those in which the environmental isolates were obtained.

Putative evidence of ward-specific evolution was also found in Hospital C. For example, four clinical isolates (Paris, HL 0101 3003, HL 0102 3034 and HL 0102 3035) obtained from patients who were treated in the intensive care unit (cardiac surgery) cluster together while one environmental isolate (Paris 2001 I n2) obtained from the nephrology ward also clusters closely with two clinical isolates (HL 0051 1015 and LG 0713 5006) from patients who were treated in this ward (**Figure 6.4**). Furthermore, the

phylogenetic analyses show that both ST1 populations detected within this hospital have co-existed within the same wards.

Evidence of shared adaptation to hospital settings was also investigated by searching for homoplasic SNPs in the lineages of Hospital A and The Wesley Hospital/Hospital B. However, none were found, including in recombined regions, suggesting that any specific adaptations may have been acquired earlier in the evolution of the ST1 lineage.

6.3.5 Long-term stability of hospital strains

Despite the discovery of substantial diversity within single ST1 hospital populations, long-term persistence of some highly similar and even identical strains was also observed. For example, isolates with no SNPs were sampled from the water supply of Hospital A over a period of five years (sampled in 2007, 2010, 2011 and 2012). Long-term persistence was also evident in Hospital C where, for example, two environmental isolates (HL 0131 3038 and LG 0713 5008) with no SNPs were sampled more than five years apart, and in Hospital D where environmental isolates sampled in 2009 and 2014 differ by just 1 SNP.

6.3.6 Evidence for hospital seeding *via* local and international spread of ST1

Phylogeographic analysis of the 229 ST1/ST1-derived isolates demonstrates that there are many examples whereby isolates cluster with epidemiologically unrelated isolates from the same region and/or country (**Figure 6.1**). In addition to the isolates from Hospital A and the surrounding area, another notable example is the six isolates sampled from or associated with three different hospitals in the Greater Copenhagen area (M, P and Q), which are no more than 10km from each other, that differ by 2-8 SNPs (not including pairwise differences between isolates from the same hospital) (**Figure 6.4**). Furthermore, an environmental isolate from Hospital C (Paris, France) is just 3 SNPs different to a clinical isolate (HL 0036 4001) from a patient who lived in Paris but who has no known epidemiological link to the hospital and is assumed to have acquired the infection from a community source (**Figure 6.4**). Another example is an environmental isolate (H092620872) sampled in 2009 from Hospital H that differs by

12 SNPs from a clinical isolate (H102860194) obtained in 2010 from a patient associated with Hospital J (~20km from Hospital H), but with no known epidemiological link to Hospital H. These findings suggest that hospitals have been seeded via the local spread of ST1.

Intriguingly, there are also isolates from distant countries, including those from or associated with hospitals, which differ by a small number of SNPs. For example, just 14 SNPs were identified between an environmental isolate (LG 1139 1124) sampled from Hospital R (France) in 2011 and an environmental isolate (LP25) sampled from The Wesley Hospital/Hospital B (Australia). Just 17 SNPs were identified between a clinical isolate (LP23) associated with Bundaberg Hospital/Hospital S (Australia) in 2011 and an environmental isolate (EUL 58) sampled from Hospital L (Spain) in 1994, and 16 SNPs between a clinical isolate (L00-549) associated with Hospital T (Germany) in 2000 and an environmental isolate (LG 1118 1044) sampled from Morocco in 2009. These findings demonstrate that ST1 strains have spread internationally, as reported in *Chapter 3*, but also that these long-distance spreading events have resulted in the seeding of hospital water systems.

6.4 Discussion

While the possibility of using WGS in investigations of community-acquired Legionnaires' disease has been well explored (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), its potential role in resolving nosocomial-associated investigations has been addressed in only a few studies (Levesque *et al.*, 2014; Bartley *et al.*, 2016). In this thesis chapter, WGS data from 229 *L. pneumophila* isolates belonging (or closely-related) to a major nosocomial-associated strain, ST1, was used to develop a greater understanding of the genomic diversity within hospital populations and how this relates to diversity elsewhere. The overall aim was to determine the feasibility of WGS-based investigations. On the one hand, the findings have revealed the enormous capability of WGS to resolve investigations due to its unparalleled resolution that, for example, can trace source

acquisition to the level of a single hospital ward. On the other hand, this study has also highlighted a number of limitations faced in WGS-based investigations of *L. pneumophila*, attributable to the unusual biology and evolution of this bacterium, which should be considered in the future interpretations of genomic data.

The first caveat is related to the finding, both from this thesis and another study (Sanchez-Buso *et al.*, 2014), that due to the low evolutionary rate of *L. pneumophila*, epidemiologically unrelated isolates exist that are highly similar or even identical at the SNP level. The implication of this, both for community- and hospital-associated investigations, is that while the existence of a low number of SNPs between isolates supports a link, it does not provide absolute evidence of one. Therefore, in the several suspected nosocomial cases that were investigated in this study from which only one clinical isolate was obtained and compared with just one environmental isolate from the hospital, it was impossible to rule out acquisition from elsewhere on the basis of the genomic data alone. However, stronger genomic evidence of a link between a case and a hospital can come from the observation that a clinical isolate is nested within and thus derived from a clade of hospital isolates. Such evidence can be achieved only by obtaining multiple isolates from the hospital and, for example, was successfully used to link seven suspected cases to Hospital A and, previously, three suspected cases to The Wesley Hospital/Hospital B (Bartley *et al.*, 2016). However, even recovery of multiple isolates (especially in low numbers) does not guarantee obtaining this key piece of supporting evidence, as was the case with six cases linked to Hospital C and one case linked to Hospital D. While the clinical isolates clustered closely with hospital isolates and with other clinical isolates associated with the same hospital, the fact that the lineage from which they are derived diverged earlier than the MRCA of the sampled hospital isolates means that acquisition from elsewhere cannot be completely ruled out on the basis of genomic data alone. To improve the chances of observing a clinical isolate nested within a clade of hospital isolates, analysis of 5-10 isolates (and preferably more) from the hospital water system would be recommended. Further work is required to understand the level of *L. pneumophila* diversity within a patient and whether analysing multiple colony picks from a clinical sample could also be useful. While the limited data available from this study (two isolates from one patient), the previous study of The Wesley Hospital/Hospital B (two isolates from one patient)

(Bartley *et al.*, 2016) and *Chapter 5* of this thesis (two or three isolates from three different patients, albeit associated with community acquisition) suggest that only very limited diversity exists between isolates obtained from the same patient (0-3 SNPs), others have shown that patients can be co-infected with multiple *L. pneumophila* variants (Coscolla *et al.*, 2014).

The requirement for deep environmental sampling is also reinforced by the discovery of two highly distinct populations of ST1 within Hospital C (that co-existed even within the same wards), as well as the substantial diversity within individual hospital populations. The combination of the high diversity within hospital populations and the relatively high similarity of hospital populations to isolates from elsewhere means that the number of pairwise SNP differences between isolates from the same hospital water system frequently outnumbers those found between hospital isolates and epidemiologically unrelated isolates from sources elsewhere, particularly within the local area (e.g. nearby homes). The implication of this is that, without deep sampling and a good understanding of the hospital diversity in relation to the local diversity, spurious links could be made on the basis of SNP differences alone. However, the finding that isolates do partially cluster by their ward of isolation suggests that, as expected, the chance of sampling an environmental isolate from the hospital that is very closely related or identical to a potentially linked clinical isolate increases if sampling is performed within the same ward as which the patient stayed.

Finally, this study reinforces the previous finding from *Chapter 3* that the ST1 lineage has surprisingly limited diversity in terms of *de novo* mutations. It has also shown that clinical isolates are interspersed amongst environmental isolates across the ST1 phylogeny, suggesting that ST1 clinical isolates are not pathogenic subtypes of the ST1 lineage, but rather that the entire ST1 lineage is adapted to, or more likely to cause, human infection (assuming that our sampling is representative). The discovery of highly similar ST1 isolates within nearby hospitals (and other community sources) suggests that hospitals may be seeded by the local "endemic" strain of ST1, possibly *via* the public water supply, from which hospital water supplies are generally derived (PHE, 2016). Some hospitals also supplement their water supply with alternative sources such as bore wells or water tankers, which could also introduce *L. pneumophila* into the hospital

water supply. Another possible method of local spread could be *via* contaminated water pipes or other plumbing devices. Nearby hospitals are more likely to use the same manufacturers and thus potentially be contaminated with similar strains. However, it is also quite remarkable that ST1 isolates from Australia and across Europe differ by just a small handful of SNPs. This finding demonstrates that ST1 has spread over long distances, as reported in *Chapter 3*, and subsequently seeded environmental sources including hospital water systems. Possible mechanisms of global spread have already been discussed in *Chapter 3*. The number of SNPs between isolates from distant countries is sometimes similar to or even lower than those between isolates from the same hospital (e.g. Hospital A), which could suggest that these long-distance spreading events have occurred within a similar time frame to that in which the hospital populations have diversified within the hospital water supply. This timeframe could span years to decades considering, for example, that Hospital A was opened in the 1970s, and thus cannot have been colonised for more than ~40 years since the last environmental isolate was obtained in 2012. However, this hypothesis firstly assumes that each hospital has been seeded once, or a limited number of times, and therefore that the observed diversity within hospital populations has been generated completely, or mostly, within the hospital itself since the initial colonisation event(s). Since isolates at least partially cluster by ward in both Hospital A and The Wesley Hospital/Hospital B, this seems a safe assumption for these hospitals. Secondly, the hypothesis also assumes that the evolutionary rate of ST1 remains relatively constant, which may not be the case. It could be that the evolutionary rate is higher in hospital water systems than other environments due to favourable replication conditions, meaning that international dispersal need not be explained by such rapid spread. As suggested in *Chapter 3*, *L. pneumophila* could also undergo periods of dormancy, which would explain our observations of identical or highly similar isolates sampled many years apart. Deepening our understanding of the speed and mechanisms by which *L. pneumophila* has spread locally and globally, and gaining further insights into the evolutionary rate and potential dormancy of this bacterium, will be important for informing future WGS-based investigations.