1. Introduction

1.1 The history and classification of *L. pneumophila*

1.1.1 The first recognised outbreak caused by *L. pneumophila*

In July 1976, an explosive outbreak of severe pneumonia caused by an unknown agent occurred in the USA (Fraser *et al.*, 1977). Intriguingly, most of the 182 cases had attended an American Legion convention in Philadelphia before returning home and falling sick (Fraser *et al.*, 1977). The mysterious illness became known as Legionnaires' disease, named after its first known victims, 29 of whom died. In the months that followed, the Centers for Disease Control and Prevention (CDC) performed an investigation to determine the etiologic agent of Legionnaires' disease by examining the patients' serum and tissue specimens (McDade *et al.*, 1977). Using a fluorescent-antibody test with survivors' serum, they demonstrated a causative role for a Gram-negative bacterium, now known as *Legionella pneumophila*, which was subsequently isolated from the lung tissues of four fatal cases (McDade *et al.*, 1977).

It is now known that people primarily become infected with *L. pneumophila* by inhaling aerosols produced from contaminated water (Muder *et al.*, 1986). Subsequent investigation found *L. pneumophila* in the cooling towers of the air conditioning systems at the convention hotel in Philadelphia. It is thought that the air-conditioning system circulated *L. pneumophila* throughout the hotel where it infected both hotel guests and even passers-by on the street.

1.1.2 Earlier isolation of *L. pneumophila*

After the formal recognition of *L. pneumophila* in 1976-77, scientists realised that it had not suddenly emerged but had been causing disease for at least several decades. The earliest isolation of the bacterium dates back to 1947 (McDade *et al.*, 1977). An organism had been isolated from a guinea pig that was inoculated with the blood of a patient with a respiratory illness and, at the time, designated a "rickettsia-like agent".

The results of serologic, cultural and DNA hybridisation studies later identified the organism as *L. pneumophila* (McDade *et al.*, 1977) and this particular isolate is now known as OLDA1.

L. pneumophila was also shown to have likely caused a number of unexplained outbreaks of respiratory disease in the years prior to its discovery (McDade *et al.*, 1977; Glick *et al.*, 1978; Osterholm *et al.*, 1983). The earliest recognised of these was an outbreak that occurred in Austin, Minnesota (USA) in 1957 in which 78 people developed pneumonia. A large number of the cases (~60%) worked at a meatpacking plant and investigation of the survivors 22 years later showed that they had significantly higher antibodies to *L. pneumophila* than matched controls (Osterholm *et al.*, 1983).

Shortly after its discovery, *L. pneumophila* was also shown to be responsible for an outbreak of a milder flu-like illness that had occurred in 1968 in Pontiac, Michigan (USA), affecting at least 144 people (Glick *et al.*, 1978). Sera from 32 out of 37 cases later demonstrated seroconversion or diagnostic rises in antibody titres to the bacterium. This milder illness is now known as Pontiac fever (Glick *et al.*, 1978). Together, the two diseases caused by *L. pneumophila*, Pontiac fever and Legionnaires' disease, are known as legionellosis.

1.1.3 *L. pneumophila* classification

Legionella is the sole member of the family Legionellaceae, which belongs to the gamma subgroup of Proteobacteria (**Table 1.1**). *L. pneumophila* is one of 59 species of the genus *Legionella* now described (http://www.bacterio.cict.fr/l/legionella.html), many of which have been associated with disease (Muder & Yu, 2002). Interestingly though, *L. pneumophila* is responsible for the large majority of Legionnaires' disease cases, including 96% of culture-confirmed cases in Europe in 2013 (ECDC, 2015). The second most common cause of Legionnaires' disease, *L. longbeachae*, accounted for just 1.4% of culture-confirmed cases in Europe in the same year, while all other *Legionella* species caused two or fewer cases (ECDC, 2015). However, in Australia, New Zealand and Japan, cases of *L. longbeachae* are just as common as *L. pneumophila* (Whiley & Bentham, 2011).

L. pneumophila has 16 described serogroups (sgs) based on their reactivity with rabbit antisera and sgs 1, 4, 5 and 6 have also been shown to consist of multiple subtypes using monoclonal antibodies (mAbs) (Joly *et al.*, 1986; Helbig *et al.*, 1997). The majority of disease cases are caused by sg 1, although this does not appear to reflect the environmental distribution of sgs. For example, a study of clinical and environmental isolates from England and Wales showed that 97.6% of clinical isolates were sg 1 compared to 55.8% of environmental isolates (Harrison *et al*., 2009). Studies have shown that the *lag-1* gene, which encodes an LPS epitope and is found only in sg 1 isolates, is also overrepresented in sg 1 clinical isolates (Helbig *et al*., 1995; Kozak *et al.*, 2009) and has thus been associated with increased virulence. However, it is not understood why sg 1 and the *lag-1*-positive strains are responsible for a high proportion of cases. It could be that they are more pathogenic to humans, more easily aerosolised or more suited to colonisation of man-made water systems (Mercante & Winchell, 2015). Three subspecies of *L. pneumophila* have also been proposed (subsp. *pneumophila*, subsp. *fraseri*, subsp. *pascullei*) based on DNA homology and multilocus enzyme typing (Brenner *et al.*, 1988).

Table 1.1. Classification of *L. pneumophila.*

Domain	Bacteria
Phylum	Proteobacteria
Class	Gammaproteobacteria
Order	Legionellales
Family	Legionellaceae
Genus	*Legionella*
Species	*Legionella pneumophila*

1.1.4 Microbiological characteristics of *L. pneumophila*

L. pneumophila is a Gram-negative, non-encapsulated coccobacillus. It is typically 0.3-0.9μm wide and 1.3μm long, although much longer, multinucleated filaments of *L. pneumophila* have also been described (Rodgers, 1979). It is aerobic, non-fermentative and requires L-cysteine and iron salts for optimal growth. Colonies of *L. pneumophila* are grey-white with a characteristic "ground-glass" appearance, and green or pink/purple iridescent edges. Free-living *L. pneumophila* is motile by means of a single, polar flagellum (**Figure 1.1**).

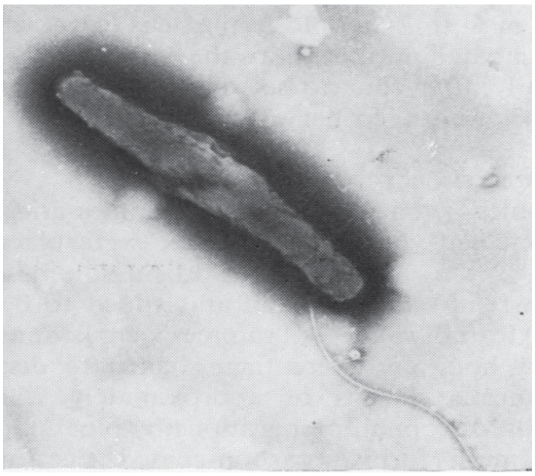

Figure 1.1. Electron microscope image of *L. pneumophila*. The flagellum is negatively stained with 1% phosphotungstic acid. Figure obtained with permission from Rodgers *et al.* (1980).

1.2 The life cycle and pathogenesis of *L. pneumophila*

1.2.1 Protozoa: the natural hosts of *L. pneumophila*

While *L. pneumophila* does occasionally cause human infection, humans are not considered to be the natural host of the bacterium. Rather, *Legionella spp*. including *L. pneumophila* have co-evolved with and parasitize unicellular protozoa which together with *Legionella spp.* are found in natural aquatic and soil environments (Rowbotham, 1980). *L. pneumophila* has a broad host range, having been shown to survive and replicate inside 15 types of protozoa including amoebae of the genera, *Acanthamoebae, Hartmannella* and *Naegleria*, ciliates of the genus *Tetrahymena* and one species of slime mould (Rowbotham, 1980; Fields *et al.*, 1984; Fields, 1996; Fields *et al.*, 2002; Molmeret *et al.*, 2005). Ensminger *et al.* (2012) propose that the broad host range keeps *L. pneumophila* in a state of evolutionary stasis whereby the organism remains a generalist rather than adapting to any specific protozoan species.

1.2.2 The "accidental" infection of humans

The ability of *L. pneumophila* to infect and replicate inside a wide range of protozoa has likely equipped it with the ability to also replicate inside human monocytes and alveolar macrophages which share common features with protozoa (Newton *et al.*, 2010). Humans usually become infected when they breathe in contaminated aerosols from an environmental source (Muder *et al.*, 1986) although one probable case of person-to-person transmission has also recently been reported (Correia *et al.*, 2016). However, the infection of humans by *L. pneumophila* is thought to be "accidental" and usually an evolutionary dead-end for the pathogen. Thus it is the relationship of *L. pneumophila* with protozoa, not humans, that is thought to have shaped the evolution of *L. pneumophila* (Albert-Weissenberger *et al.*, 2007).

1.2.3 The intracellular life cycle of *L. pneumophila*

The intracellular life cycle of *L. pneumophila* has been studied *in vitro* using protozoa such as *Acanthamoeba castellani*, *Hartmannella vermiformis* and *Naegleria spp*., as well as human macrophage and epithelial cells. In all host cells studied, the primary mechanisms of infection and replication appear to be the same although the mechanisms of cell entry and exit can vary (Gao *et al.*, 1997; Vogel & Isberg, 1999). This likely reflects the high conservation between the cellular pathways of protozoa and human phagocytes targeted by *L. pneumophila* (Molofsky & Swanson, 2004).

The life cycle of *L. pneumophila* consists of at least two discrete phases: a replicative phase inside the host cell in which the bacteria are in an unflagellated form, and the post-exponential phase during which the bacteria are in a flagellated, motile form and escape from the host cell (Albert-Weissenberger *et al.*, 2007). While the intracellular life cycle begins by the phagocytosis of *L. pneumophila* by host cells, there is debate as to whether phagocytosis is driven by the host or bacterium (Newton *et al.*, 2010). Once internalised, by immediately altering the phagosomal membrane, *L. pneumophila* forms a safe compartment called the *Legionella*-containing vacuole (LCV) (**Figure 1.2a**), thus evading digestion by the conserved endocytic pathway of eukaryotic cells (**Figure 1.2b**). This pathway would normally result in the fusion of the phagosome with the lysosome,

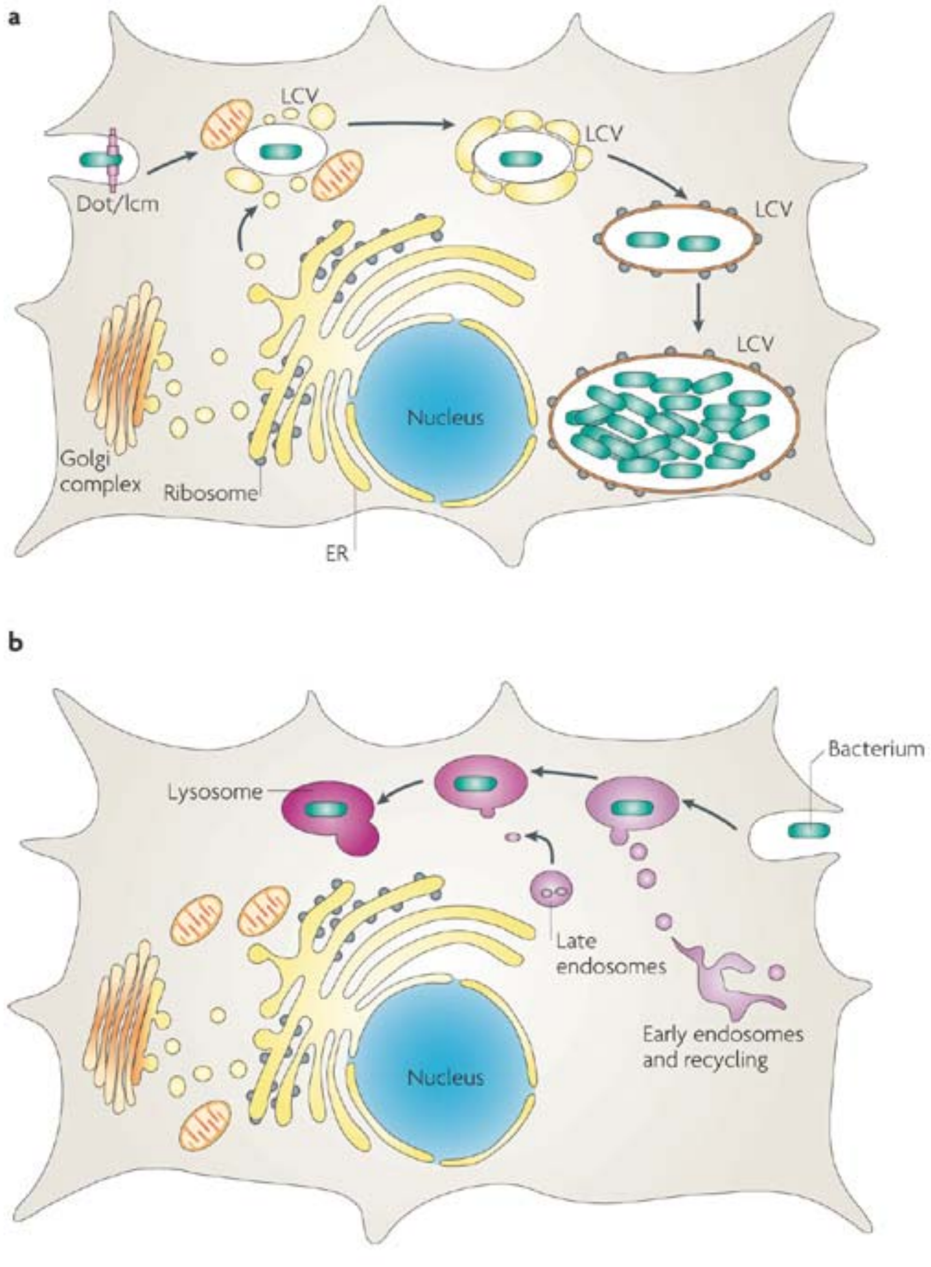

Figure 1.2. The *Legionella*-containing vacuole. a) *L. pneumophila* enters the host cell by phagocytosis, evades delivery to the lysosomal network and immediately establishes a compartment known as the LCV. Endoplasmic reticulum (ER)-derived vesicles and subsequently mitochondria surround the LCV. The vesicles form layers of membrane around the compartment and become studded with ribosomes, giving the LCV an appearance similar to rough ER. *L. pneumophila* replicates inside the LCV before lysing the cell. b) The default trafficking pathway of a non-pathogenic bacterium. The bacterial phagosome fuses with early and late endosomes and finally lysosomes where the bacteria are degraded. Figure reproduced with permission from Isberg *et al.* (2009).

acidification of the vacuole and degradation of the microbe (Isberg *et al.*, 2009). Instead, proteins characteristic of late endosomes and lysosomes are not present on the LCV (Horwitz & Maxfield, 1984) and the luminal pH remains neutral (Horwitz & Maxfield, 1984; Sturgill-Koszycki & Swanson, 2000). Within minutes of uptake, the vacuole becomes surrounded by endoplasmic reticulum (ER)-associated proteins, ER-derived vesicles and, later, mitochondria. The vesicles form a layer of membranes surrounding the vacuole (Isberg *et al.*, 2009) that subsequently becomes studded with ribosomes, giving the vacuole the appearance of rough endoplasmic reticulum (RER) (Tilney *et al.*, 2001). Within this disguised compartment, *L. pneumophila* replicates to high numbers (**Figure 1.3**) (Horwitz, 1983). Crucially, it is protected from the cellular immune system and provided with energy and nutrients for replication (Xu & Luo, 2013). Once nutrient concentrations and host cell viability declines, *L. pneumophila* undergoes a switch from its replicative form to the flagellated, transmissive form. The bacteria rupture the LCV and exit the host cell using pore-forming toxins (Molmeret & Kwaik, 2002) and can then be internalised by neighbouring cells for further rounds of infection.

1.2.4 The Dot/Icm secretion system

The mechanisms through which *L. pneumophila* subverts host cell processes to establish infection and replication have been studied intensely. This work has uncovered a remarkable array of virulence factors, the most notable of which is the Dot/Icm (defect in organelle trafficking/intracellular multiplication) type IVB secretion system. This apparatus has been found to be conserved across all *Legionella* species studied (Feldman *et al*., 2005). It consists of 27 proteins that span the bacterial and phagosomal membranes (**Figure 1.4**) (Christie *et al.*, 2005), and almost all of these have been shown to be essential for successful establishment of the LCV and intracellular replication (Isberg *et al.*, 2009). In *L. pneumophila*, the system secretes over 300 effector proteins into the host cell (Harding *et al.*, 2013), which make up approximately 10% of the protein-coding capacity (Cazalet *et al.*, 2004). Remarkably, a recent study of 38 different *Legionella* species detected a total of 5885 putative Dot/Icm effectors belonging to over 600 orthologous gene families (Burnstein *et al.*, 2016). Most gene families were found in fewer than ten species while only seven were found to be shared by all species.

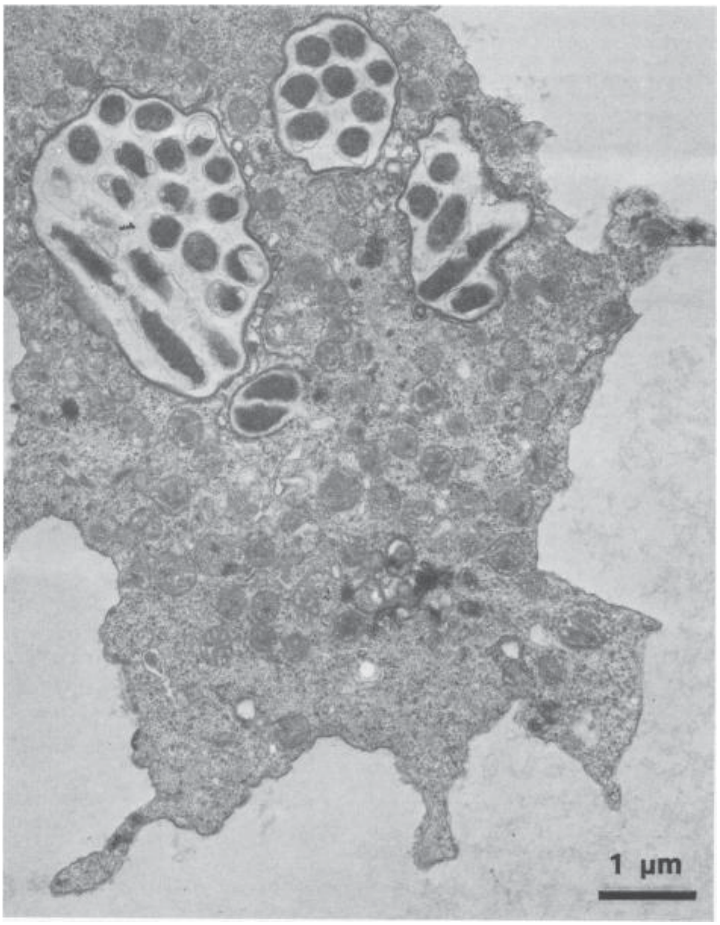

Figure 1.3. Infection of *Acanthamoeba castellani* with *L. pneumophila*. Transmission electron micrograph of *L. pneumophila* contained within LCVs inside *Acanthamoeba castellani* at 48h post-infection. Figure reproduced with permission from Holden *et al.* (1984).

The *L. pneumophila* effectors have been shown to manipulate a wide range of host cell processes such as membrane trafficking, apoptosis, ubiquitination, and innate immune signalling to achieve survival and replication. Interestingly, many effectors share sequence similarity with eukaryotic proteins or possess typical eukaryotic domains such as ankyrin repeats, and this characteristic feature has been an important means of effector identification (Cazalet *et al.*, 2004; Chen *et al.*, 2004; Chien *et al.*, 2004; de Felipe *et al.*, 2005; Habyarimana *et al.*, 2008; Kubori *et al.*, 2008; Pan *et al.*, 2008). These effectors, often termed eukaryotic-like proteins, may have arisen through horizontal gene transfer and/or convergent evolution (Gomez-Valero *et al.*, 2011) and likely manipulate host cell processes through molecular mimicry.

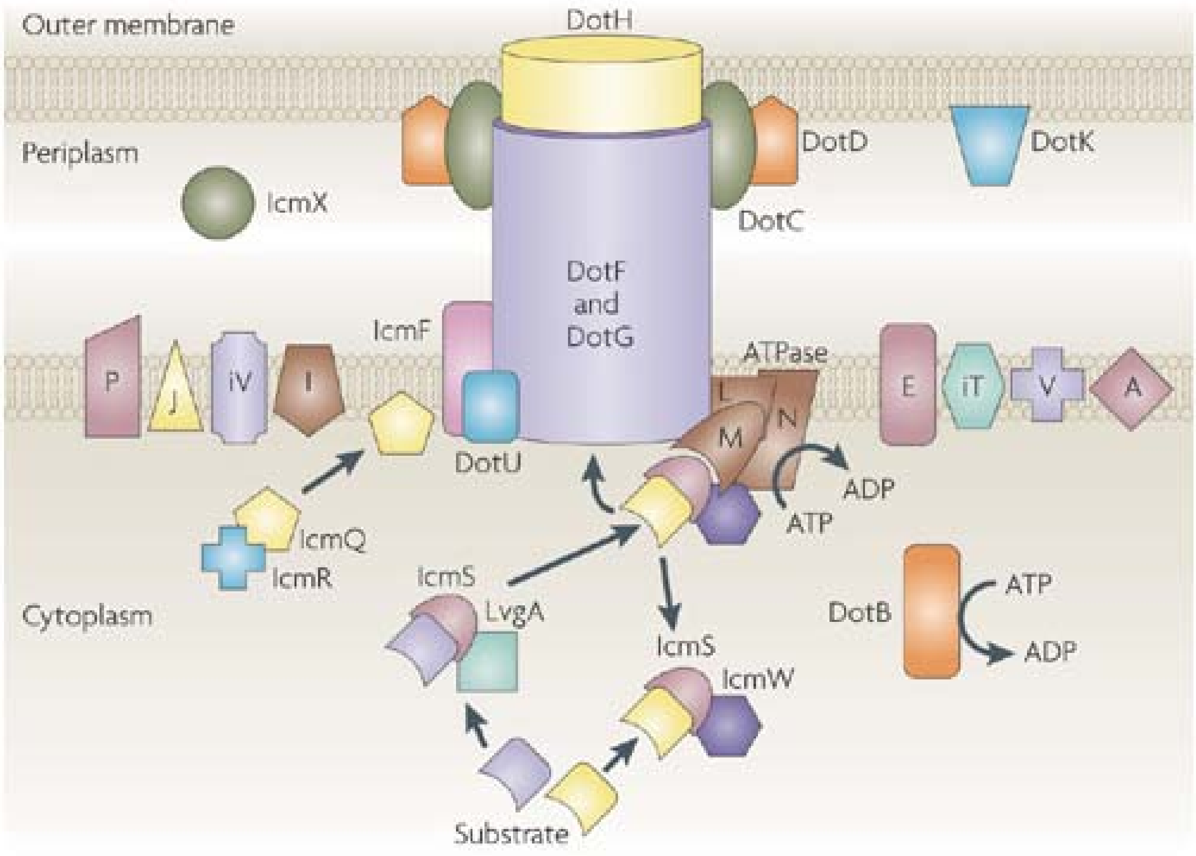

Figure 1.4. Dot/Icm machinery. The components of the Dot/Icm (defect in organelle trafficking/intracellular multiplication) machinery in the bacterial cell envelope. Figure reproduced with permission from Isberg *et al.* (2009).

1.3 Disease caused by *L. pneumophila*

1.3.1 Legionnaires' disease

Legionnaires' disease is an acute and sometimes severe pneumonia caused by species of the genus, *Legionella*. It accounts for 2-5% of community-acquired pneumonia (Lim *et al.*, 2001; von Baum *et al.*, 2008) and is also recognised as an increasingly important cause of hospital-acquired pneumonia (Lin *et al.*, 2011).

The incubation period of Legionnaires' disease is typically between 2 and 10 days (Diederen, 2008) but may be much longer (Lettinga *et al.*, 2002). The symptoms are non-specific and include fever, non-productive cough, headache, myalgia, diarrhoea and delirium (Tsai *et al.*, 1979). Therefore, it is not possible to clinically distinguish Legionnaires' disease from other types of pneumonia such as that caused by pneumococcal bacteria (Edelstein, 1993). This highlights an important role for prompt microbiological testing in suspected cases as prompt administration of effective antibiotics are crucial for successful treatment of Legionnaires' disease (Phin *et al.*, 2014). Since *L. pneumophila* is an intracellular pathogen, antibiotics that can penetrate cells are required such as macrolides, fluoroquinolones or those of the cyclin families (Mykietiuk *et al.*, 2005; Blazquez Garrido *et al.*, 2005; Mandell *et al.*, 2007; Haranaga *et al.*, 2007; Griffin *et al.*, 2010; Garau *et al.*, 2010). Specifically, azithromycin and levofloxacin are recommended, both in healthy and immunocompromised individuals (Phin *et al.*, 2014). However, the dose and route of administration (oral or intravenous) is determined by disease severity, patient consciousness, and any underlying risk factors or further complications (Phin *et al.*, 2014). Crucially, beta-lactam antibiotics have poor intracellular penetration and are not effective at treating infection by *Legionella*.

The mortality rate of Legionnaires' disease is typically 8-12%, and thus within a similar range as for other bacterial pneumonias. However, it can depend on a range of factors including promptness of specific antibiotic treatment, the patient's underlying health, whether the patient is a smoker and whether cases are sporadic, nosocomial or part of a large outbreak (Dominguez *et al.*, 2009). While many people are exposed to *Legionella spp.*, only a very small proportion develops Legionnaires' disease (Keller *et al.*, 1996; Den Boer *et al.*, 2002; Sabria *et al.*, 2006; Beyrer *et al.*, 2007) demonstrating the low efficiency of infection (Isberg *et al.*, 2009). For example, at a flower show in the Netherlands in 1999, of 77,061 visitors that attended, 188 became ill giving an attack rate of 0.24% (Den Boer *et al.*, 2002).

1.3.2 Pontiac fever

Pontiac fever is a less reported, less serious form of legionellosis and often identified only when cases occur as part of an outbreak or cluster (Glick *et al.*, 1978; Kaufmann *et al.*, 1981; Tossa *et al.*, 2006). It is generally characterised by fever, chills, myalgia and headache (Kaufmann *et al.*, 1981). It has a shorter incubation period than Legionnaires' disease (usually 6-8h), a high attack rate of up to 95% (Glick *et al.*, 1978) and is more common in younger people (Phin *et al.*, 2014). No deaths or long-term complications have been attributed to Pontiac fever (Fields *et al.*, 2001).

The pathogenesis of Pontiac fever is not well understood, nor is why Pontiac fever and Legionnaires' disease result in clinically and epidemiologically distinct illnesses (Fields *et al.*, 2001). One theory is that Pontiac fever results from exposure to dead *Legionella* (Eickhoff, 1979). However, live legionellae have been recovered from environmental sites associated with point-source outbreaks (Fraser *et al.*, 1979; Girod *et al.*, 1982; Friedman *et al.*, 1987). An alternative hypothesis is that Pontiac fever is caused by hypersensitivity to cellular components of either *Legionella* or the associated amoebae (Rowbotham, 1980; Rowbotham, 1986).

1.3.3 Extra-pulmonary disease

Extrapulmonary infection with *Legionella spp.* is extremely rare and has been associated with surgical patients (Lowry & Tompkins, 1993). It can occur in the presence or absence of Legionnaires' disease. Various clinical manifestations have been reported including sinusitis, cellulitis, pancreatitis, peritonitis and pyelonephritis and brain abscesses (Eitrem *et al.*, 1987; Lowry & Tompkins, 1993; Stout & Yu, 1997). The most common extrapulmonary infections, however, are those of the heart and include myocarditis, pericarditis and prosthetic-valve endocarditis (Nelson *et al.*, 1985; Tompkins *et al.*, 1988). In these cases, there has usually been no accompanying pneumonia and it is thought that contaminated water has been introduced into a postoperative sternal wound or the site of a suture of a drainage tube (Lowry *et al.*, 1991). There have also been rare reports of neural infections associated with encephalomyelitis, cerebellum involvement and peripheral neuropathy (Johnson *et al.*, 1984; Shelburne *et al.*, 2004).

1.4 Microbiological identification and detection

Since the diagnosis of Legionnaires' disease cannot be made on clinical or radiological grounds alone, microbiological testing is required in order that appropriate antibiotic therapy is administered. A number of methods have been developed and used although most are biased towards the detection of *L. pneumophila* sg 1. This is likely a major contributing factor to the under-diagnosis of *Legionella* infections.

1.4.1 Culture methods

Culture is the "gold standard" method for the diagnosis of *Legionella* infections and has the highest specificity of any method. The standard medium is buffered charcoal yeast extract (BCYE) agar supplemented with alpha-ketoglutarate. This provides both L-cysteine and iron, which are required by *L. pneumophila*. Methods can also be used to reduce contaminating flora such as the addition of antibiotics (Wadowsky & Yee, 1981; Edelstein, 1982) and heat and acid treatments (Edelstein *et al.*, 1982; Dennis, 1988). However, since heat and acid treatments can also inhibit the growth of *Legionella spp*., they should be used in combination with untreated samples (Munro *et al.*, 1994). Additionally, BCYE medium lacking cysteine is often used in conjunction with traditional BCYE agar. Colonies that grow on traditional BCYE, but not BCYE without cysteine, are indicative of *Legionella spp.*

L. pneumophila is a slow-growing organism and it usually takes 3-5 days to detect colonies (Murdoch, 2003). Therefore, the culture method can fail to give a timely diagnosis (Reischl *et al.*, 2002). Another problem is that the obtainment of respiratory samples for culture can be difficult due to the characteristic dry cough of Legionnaires' disease (Phin *et al.*, 2014). The sensitivity is also low (approximately 60%) although highly dependent on the type of clinical sample used (Edelstein, 1993; Ramirez & Summersgill, 1994). While sputum samples are the most common clinical specimens obtained, they yield a lower sensitivity than bronchoalveolar lavage (BAL) fluid, bronchial aspirates, lung biopsy and post-mortem tissue samples (Maiwald *et al.*, 1998). The sensitivity of culture can also be low due to the "viable but not culturable" (VBNC) phase of *Legionella* (Hussong *et al.*, 1987). A study showed that sensitivity was increased to 80% when samples were taken within two days of patient admission to hospital (Mentasti *et al.*, 2012). There is no evidence, however, that culturability of different *L. pneumophila* strains is variable since the same study showed that strains detected by culture (~65% of Legionnaires' disease cases) or PCR (~20% cases) show a similar distribution of sequence types (see *1.5.4*). However, the overall low sensitivity, particular for detecting non-pneumophila *Legionella spp*., calls for improved culture methods.

1.4.2 Serologic diagnosis

L. pneumophila was first identified as the etiological agent of Legionnaires' disease in the 1976 Philadelphia outbreak by serology. Since then, various serological methods have been used for the diagnosis of *Legionella* infections. The most widespread are the indirect fluorescent antibody (IFA) test and the enzyme-linked immunosorbent assay (ELISA or EIA) (Wilkinson *et al.*, 1979; Stanek *et al.*, 1983). Tests using paired sera (acute and convalescent) are generally more reliable than those using a single convalescent specimen and require a fourfold antibody titre rise to confirm *Legionella* infection.

The major disadvantage of using serology is that seroconversion to *Legionella spp*. is highly variable between infected patients. For example, while approximately 25-40% of patients seroconvert within a week of developing symptoms, about 10% do not seroconvert until up to 9 weeks post-disease onset and as many as 20-30% of patients do not seroconvert at all (Harrison & Taylor, 1988; Edelstein, 1993; Maiwald *et al.*, 1998). The specificity of serological methods to detect *L. pneumophila* may also be reduced by cross-reactions with other species including *Pseudomonas aeruginosa*, *Campylobacter spp*., *Rickettsia spp* and *Coxiella burnetti*, among others (Harrison & Taylor, 1988; Edelstein, 1993; Musso & Raoult, 1997), and results must be interpreted with some caution. Thus, while serological methods are useful tools in epidemiological studies of *L. pneumophila*, they are now rarely used for clinical diagnosis and decision-making (Murdoch, 2003).

1.4.3 Direct fluorescent antibody testing

Direct fluorescent antibody (DFA) testing using fluorochrome-conjugated antibody to stain clinical specimens has been used as a rapid method of *L. pneumophila* diagnosis. Crucially, *L. pneumophila* can be detected in respiratory secretions by DFA even after several days of antibiotic therapy (Fields *et al.*, 2002). Another advantage of DFA over culture techniques is that it can detect VBNC *L. pneumophila* (Bangsborg *et al.*, 1990). The sensitivity of DFA testing depends on the type of specimen used (typically sputum, BAL or lung biopsy tissue), the disease severity and the experience of the staff (Edelstein, 1993). Thus estimates are highly variable and have ranged from 27% to 70% (Edelstein, 1987; Edelstein, 1993; Ramirez & Summersgill, 1994). The specificity is very high (>95%) although false positive results can occur if clinical samples are mixed with contaminated reagents during the testing procedure (Haldane *et al.*, 1993). Cross-reactions with other bacteria have also been reported, occasionally leading to false positive results (Cherry *et al.*, 1978; Flournoy *et al.*, 1988; Roy *et al.*, 1989).

1.4.4 Urine antigen detection

Urine antigen testing is an established tool that is used in the majority of laboratories for the diagnosis of *L. pneumophila* in conjunction with culture methods. Several commercial kits are available (Dominguez *et al.*, 1998; Harrison & Doshi, 2001) and the vast majority of cases (including 70-80% in Europe) are now diagnosed with this method (ECDC, 2015). The advantages of this method are that it is quick, urine samples are easy to obtain, *L. pneumophila* antigens are detectable early during the course of infection (Kohler *et al.*, 1984) and it has high specificity of up to 100% (Aguero-Rosenfeld & Edelstein, 1988; Birtles *et al.*, 1990). The main disadvantage is that the urine antigen test detects sg 1 only. Therefore, a negative urinary antigen result cannot exclude infection by *L. pneumophila* non-sg 1 isolates or other *Legionella spp*. While *L. pneumophila* sg 1 is predicted to cause approximately 85% of all Legionnaires' disease cases (Beaute *et al.*, 2013), our estimates may be biased due to the heavy reliance on this test. Additionally, the sensitivity of the urine antigen test for patients even with *L. pneumophila* sg 1 may not always be high, with estimates varying between 60 and 100% (Dominguez *et al.*, 1998; Dominguez *et al.*, 1999; Yzerman *et al.*, 2002).

1.4.5 PCR-based detection

Real-time polymerase chain reaction (PCR) is now the molecular method of choice due to its high specificity, sensitivity and rapidity (Phin *et al.*, 2014). One *L. pneumophila*-specific PCR targeting the *mip* gene demonstrated 100% specificity and 30% greater sensitivity than culture (Mentasti *et al.*, 2012).

1.5 Typing methods & outbreak investigations

Since the occurrence of a Legionnaires' disease case implies a source of contaminated water that could infect more people, rapid establishment and control of the source is of high priority. Source identification can be difficult in sporadic (single) cases especially since the incubation period of Legionnaires' disease can be long and variable. It becomes easier though when two or more cases of Legionnaires' disease occur in a similar place or time. Epidemiological information is crucial, and by tracing the recent movements of the patients, putative sources can be identified. This information is used in conjunction with molecular typing methods that aim to determine whether patients are infected with the same strain and whether the clinical isolates match environmental isolates sampled from putative sources. As with the diagnostic methods, a number of methods have been used over the years to "type" *L. pneumophila*, and the most widely used are discussed below. Often methods are used together to further increase the index of discrimination. Currently used in most laboratories are monoclonal antibody (mAb) subgrouping (Helbig *et al.*, 2002) and sequence-based typing (SBT) (Gaia *et al.*, 2003; Gaia *et al.*, 2005; Ratzow *et al.*, 2007; Mentasti *et al.*, 2014).

1.5.1 Pulsed field gel electrophoresis

Pulsed field gel electrophoresis (PFGE) has been a widely used method of *L. pneumophila* subtyping for over 20 years (Ott *et al.*, 1991; Luck *et al.*, 1995; Nguyen *et al.*, 2006). However, its main use today is in discriminating between isolates of other *Legionella spp*. for which an SBT scheme is not available (Akermi *et al.*, 2006; Matsui *et* *al.*, 2010). It uses rare-cutting restriction enzymes to cut the genome into 10-20 fragments that, when separated on a gel, produce distinct banding patterns. These can be easily analysed visually and the method has been shown to have a high index of discrimination. Its main disadvantage, however, is that the method is difficult to standardise and results are not easily exchangeable between different laboratories (Fry *et al.*, 1999).

1.5.2 Amplified fragment length polymorphism

Prior to SBT, amplified fragment length polymorphism (AFLP) was adopted as the international standard for *L. pneumophila* typing (Fry *et al.*, 2002). Genomic DNA is firstly digested with restriction enzymes before adaptor sequences are ligated to the sticky ends of the resulting fragments. A specific subset of the fragments is then amplified using primers complementary to the adaptor sequences, the restriction site sequence and additional bases inside the restriction site fragments. The fragments are separated by gel electrophoresis allowing the comparison of banding patterns. However, as with PFGE, this approach has been difficult to standardise between different laboratories and has now largely been replaced by SBT.

1.5.3 Monoclonal antibody subgrouping

Isolates of *L. pneumophila* sg 1 can be subtyped using panels of monoclonal antibodies (mAb) (Helbig *et al.*, 1997). Depending on the panel of antibodies used, isolates are partitioned into 8 to 10 groups giving this subgrouping only a low index of discrimination. Despite this, the method is cheap and easy, and has proved very useful for quickly excluding environmental isolates unrelated to clinical strains (Luck *et al.*, 2013).

1.5.4 Sequence-based typing

Sequence-based typing (SBT) is analogous to multi-locus sequence typing (MLST) whereby isolates are assigned a “sequence type” (ST) based on the sequence of seven genes (Gaia *et al.*, 2003; Gaia *et al.*, 2005; Ratzow *et al.*, 2007; Mentasti *et al.*, 2014).

However, whereas MLST schemes usually use housekeeping genes, SBT uses a mixture of housekeeping genes and virulence genes in order to achieve a higher index of discrimination. The seven gene targets used in SBT are *flaA*, *pilE*, *asd*, *mip*, *mompS*, *proA* and *neuA*. SBT has a high index of discrimination and, as of 8 July 2016, there are 2190 STs recorded in the European Study Group for *Legionella* Infections (ESGLI) database (http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php). Nested protocols, which involve two rounds of PCR, can also be performed on DNA extracts from clinical samples containing only low amounts of genomic DNA (Ginevra *et al.*, 2009). However, some STs such as ST1 are isolated very frequently (e.g. ST1), and thus the method can lack discriminatory power.

1.6 Epidemiology of Legionnaires' disease

1.6.1 The incidence of Legionnaires' disease

The global incidence of Legionnaires' disease is difficult to measure since, in many parts of the world, Legionnaires' disease is an under-recognised and under-diagnosed disease (Phin *et al.*, 2014). This is due to a number of factors including the difficulty of clinically distinguishing Legionnaires' disease from other pneumonias (Edelstein, 1993). A diagnosis of Legionnaires' disease is reliant on clinicians requesting specific microbiological testing, and it can take several days for results to be returned. However, when a patient is diagnosed with pneumonia, antibiotic treatment is usually started immediately. If antibiotics effective against *Legionella* are used, the patient usually recovers and often no cause of the pneumonia is sought. Another important factor is that the most commonly used diagnostic method, the urinary antigen test, detects only *L. pneumophila* sg 1 (Kashuba & Ballow, 1996). It is therefore probable that many cases of Legionnaires' disease are missed that are caused by other species and serogroups.

However, some countries do have surveillance systems in place for Legionnaires' disease including the USA, Canada, New Zealand, Australia, Japan and Singapore (Phin *et* *al.*, 2014). Additionally, a system called the European Legionnaires' Disease Surveillance Network (ELDSNet), coordinated by ECDC, performs surveillance of Legionnaires' disease in Europe, while many European countries also have their own national systems in place. In 2013, a total of 5,851 cases of Legionnaires' disease were reported by 28 European Union (EU) member states and Norway (ECDC, 2015). However, just six countries (France, Italy, Spain, Germany, the Netherlands and the UK) accounted for 83% cases (ECDC, 2015), reflecting the under-diagnosis and under-reporting of Legionnaires' disease in much of Europe. While the number of reported cases in Europe was increasing for several years, possibly due to improved diagnosis and reporting, increased use of the urine antigen test and improved clinical awareness, the incidence has been quite consistent since 2005 (**Figure 1.5**) (ECDC, 2015). Interestingly, the incidence of Legionnaires' disease shows a seasonal trend, with cases peaking in the late summer to autumn (**Figure 1.6**) (ECDC, 2015). This could be due to warmer, wetter weather and higher humidity at this time of year (Fisman *et al.*, 2005; Ng *et al.*, 2008; Karagiannis *et al.*, 2009; Ricketts *et al.*, 2009).

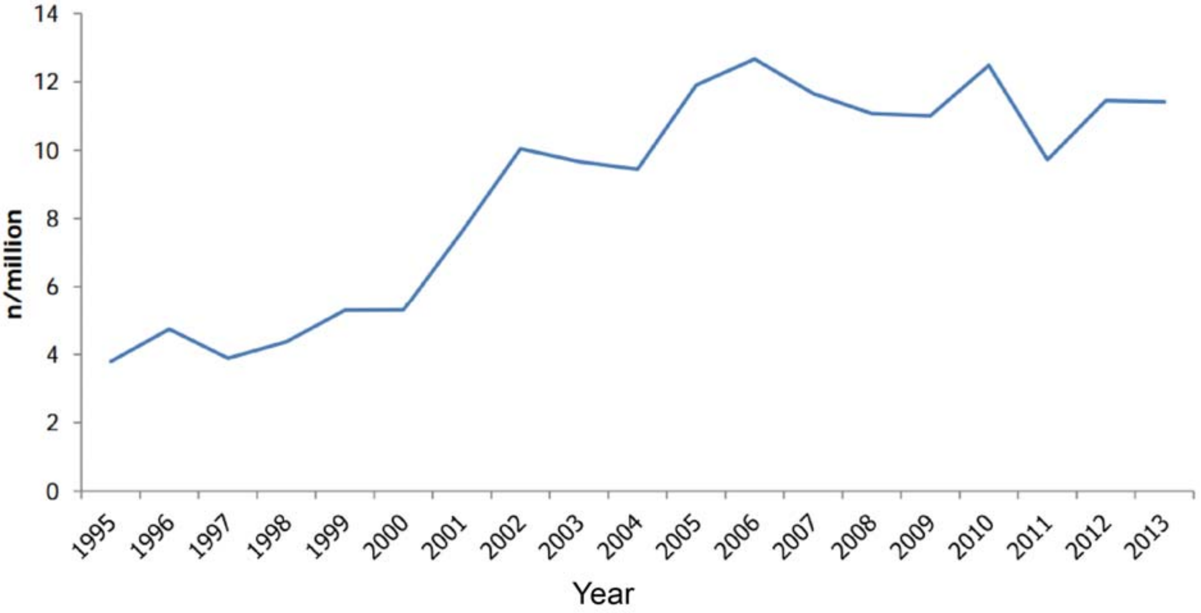

Figure 1.5. Incidence of Legionnaires' disease. The annual notification rates of Legionnaires' disease in the European Union/European Economic Area (EU/EEA) from 1995 to 2013. Figure reproduced with permission from ECDC (2015).

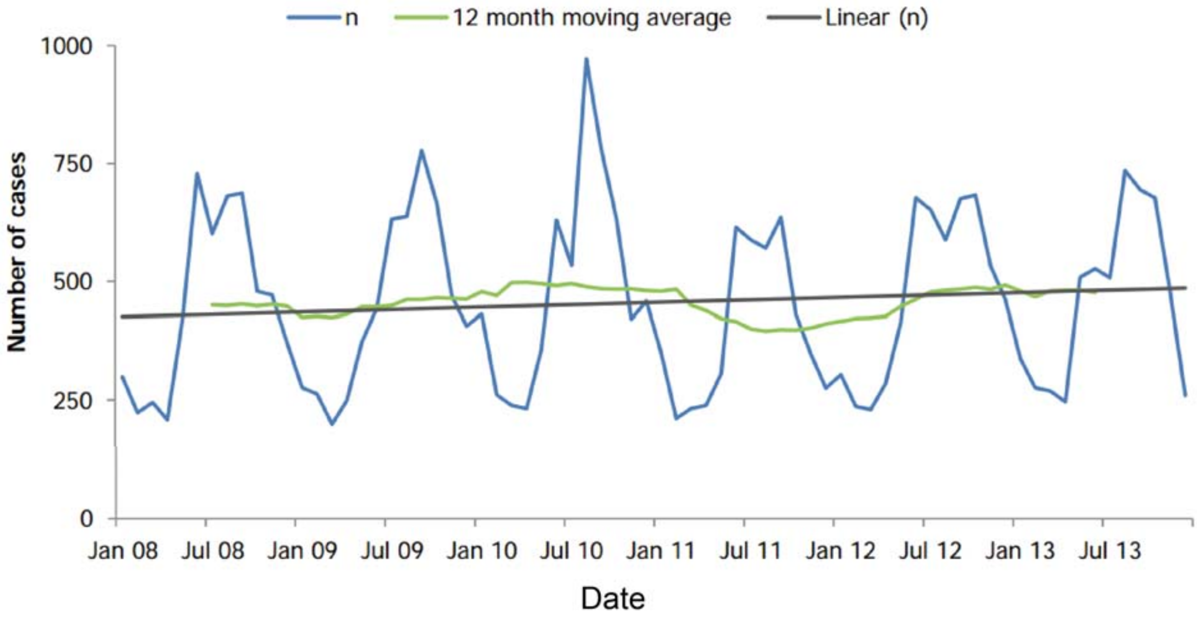

Figure 1.6. Seasonal trend of Legionnaires' disease. Reported cases of Legionnaires' disease by week of onset in the EU/EEA between 2008 and 2013. Figure reproduced with permission from ECDC (2015).

1.6.2 Sporadic cases, clusters and outbreaks

While Legionnaires' disease is sometimes associated with dramatic outbreaks, the majority of cases occur sporadically (Beaute *et al.*, 2013). However, the proportion of cases in clusters is higher in travel-associated cases (~20%) than community-acquired cases (5%) (ECDC, 2015).

In this thesis, a "cluster" refers to the occurrence of two or more cases that are linked in both space (e.g. place of work, hospital) and time (up to six months), but no common source of infection is identified. An "outbreak" is defined by the occurrence of two or more cases closely linked in time (weeks rather than months) and space, and where there is a suspected or proven common source.

1.6.3 Common sources of infection

In sporadic cases of Legionnaires' disease, the environmental source of infection is often not identified. While it is likely that a significant number of sporadic cases are residentially acquired and due to contaminated domestic water systems (Straus *et al.*, 1996), further studies are warranted. However when outbreaks of Legionnaires' disease occur, such as that which occurred at the Philadelphia convention in 1976, they provide the opportunity to identify common sources (O'Loughlin *et al.*, 2007). Outbreaks have been associated with a range of man-made environments including cooling towers, spa pools, decorative fountains, air-scrubbers and hot and cold water systems of large buildings (Shands *et al.*, 1985; Zumla *et al.*, 1988; Hlady *et al.*, 1993; Fields *et al.*, 2002; O'Loughlin *et al.*, 2007; Nygard *et al.*, 2008; Coetzee *et al.*, 2012; Silk *et al.*, 2012; Bennett *et al.*, 2014). In such environments, warm and/or stagnant water and biofilms can promote the replication and growth of *Legionella spp*. including *L. pneumophila*. Natural environmental sources have only been implicated in disease rarely although, increasingly, cases associated with hot springs are being reported (Ito *et al.*, 2002; Lin *et al.*, 2007).

1.6.4 Transmission

The inhalation of contaminated aerosols is thought to be the primary route of *L. pneumophila* infection and has been implicated in the vast majority of disease cases (Muder *et al.*, 1986). In most well described outbreaks, patients have come into close contact with the putative source although there have also been studies implicating the dissemination of *Legionella* from cooling towers over large distances (up to several kilometres) in a small number of disease cases (Addiss *et al.*, 1989; Nguyen *et al.*, 2006).

It has been proposed that aspiration or ingestion of contaminated water may also play a role in the acquisition of some infections (Yu, 1993; Venezia *et al.*, 1994) although such cases are probably rare. For example, the aspiration of nasogastric feedings diluted with contaminated tap water was speculated to be responsible for two cases of nosocomial Legionnaires' disease (Venezia *et al.*, 1994) and a case has also been linked to aspiration of ice from an ice-making machine in a hospital (Bencini *et al.*, 2005). A number of extrapulmonary infections have also been associated with direct topical exposure to contaminated tap water (Lowry & Tompkins, 1993).

Finally, a probable case of person-to-person transmission has also recently been reported between a mother and son (Correia *et al.*, 2016). The son was part of a cluster in Vila Franca de Xira, Portugal and, after becoming infected, travelled approximately 300km to stay with his mother. The son had very severe respiratory symptoms including an intense cough and was looked after by his mother for 8 hours in a small, non-ventilated room before being admitted to hospital. Approximately one week later, the mother was admitted to hospital with septic shock due to pneumonia. Both patients tested positive for *L. pneumophila* sg 1 and whole genome sequencing (WGS) revealed that there were no nucleotide differences between isolates from the two patients.

1.6.5 Host risk factors

Not everyone is equally susceptible to Legionnaires' disease and there are many risk factors that predispose individuals to disease. These include older age (being of 50 years or older) and gender (see **Figure 1.7**) as well as smoking, alcohol misuse, chronic cardiovascular or respiratory disease, diabetes, renal disease, cancer and immunosuppression (Rosmini *et al.*, 1984; Marston *et al.*, 1994; Den Boer *et al.*, 2006).

1.6.6 Travel-associated Legionnaires' disease

In 2013, 19% of Legionnaires' disease cases in Europe were associated with travel, 83% of which involved hotels (ECDC, 2015). Cruise ships have also been associated with Legionnaires' disease cases and accounted for 2% of travel-associated cases in 2013 (ECDC, 2015). Since travel-associated Legionnaires' disease is not usually diagnosed until the patient is back in their home country, an international response is often required. For this reason, ECDC set up the European surveillance system, ELDSNet, to investigate travel-associated cases with the aim of identifying the source and initiating public health action. Since the establishment of a European surveillance system in 1987, the number of reported travel-associated cases has increased dramatically (**Figure 1.8**) (ECDC, 2015).

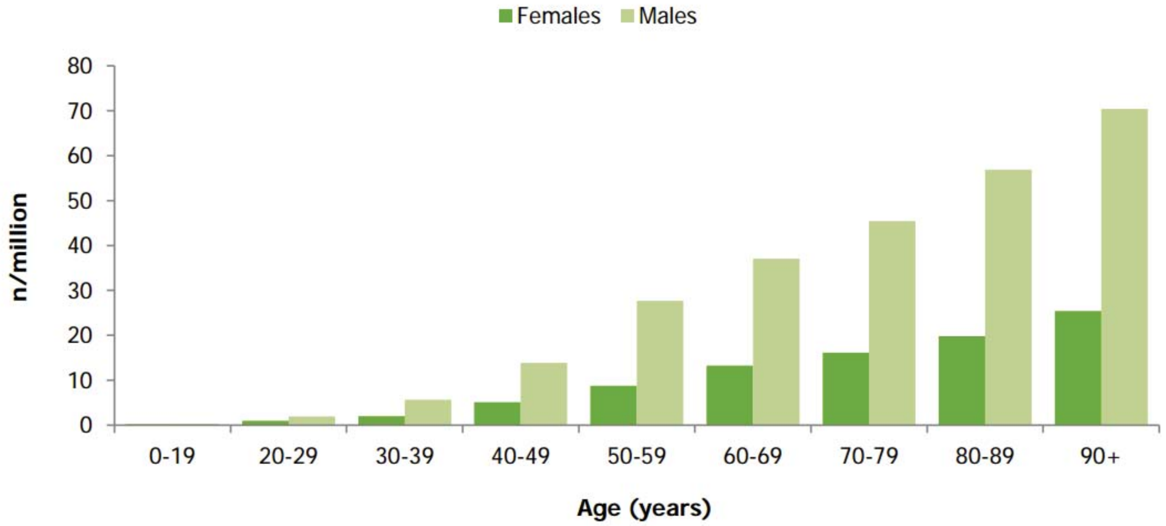

Figure 1.7. Distribution of Legionnaires' disease cases by age and sex. The number of reported cases of Legionnaires' disease per million by gender and age in the EU/EEA in 2013. Figure reproduced with permission from ECDC (2015).

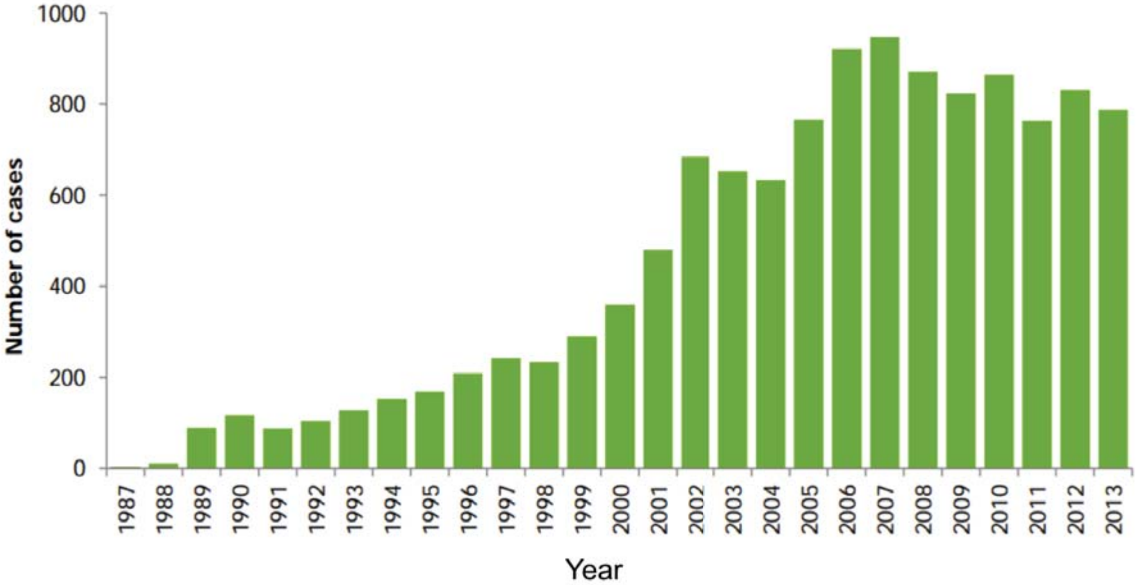

Figure 1.8. Incidence of travel-associated Legionnaires' disease. The number of annually reported cases of travel-associated Legionnaires' disease in EU/EEA member states from 1987 to 2013. Figure reproduced with permission from ECDC (2015).

1.6.7 The distribution of *L. pneumophila* subtypes in clinical disease

As of 8 July 2016, clinical isolates (*n*=7181) submitted to the ESGLI SBT database comprise 1171 STs while environmental isolates (*n*=3631) comprise a total of 1324 STs. This indicates more diversity is found within environmental isolates than clinical isolates and that the distribution of clinical STs does not simply mirror what is found in the environment. It also suggests that important differences in virulence may exist between strains. A number of studies have also echoed these observations. For example, a study of 443 environmental and community-acquired clinical isolates obtained in England and Wales from 2000 to 2008 showed that almost 50% of clinical cases were attributed to just three STs (ST37, ST47 and ST62), which were found in the environment very rarely (**Figure 1.9**) (Harrison *et al.*, 2009). Conversely, STs that were found commonly in the environment (e.g. ST1 and ST79) caused disease less frequently than expected given their environmental prevalence. These findings suggest that knowing which particular strains are present in a system could be an important factor in weighing up the risk of *L. pneumophila* infection.

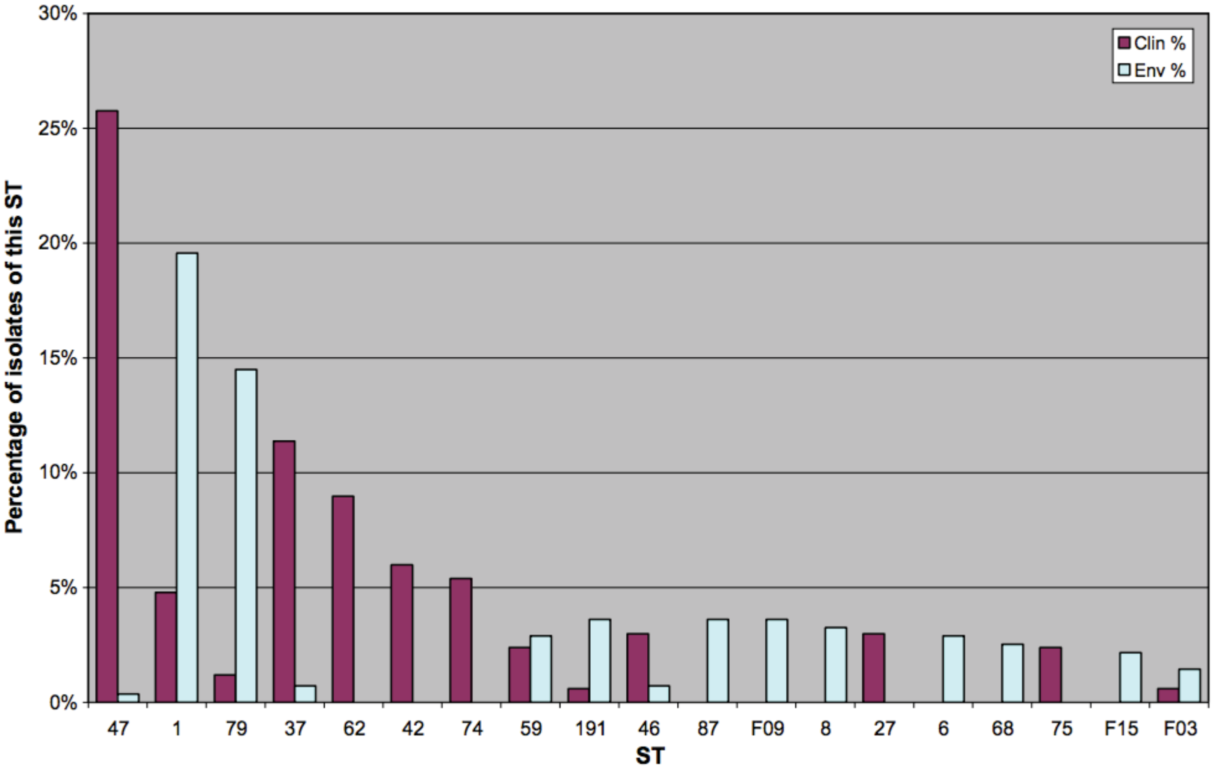

Figure 1.9. Distribution of STs among clinical and environmental isolates (previous page). The prevalence of various STs among community-acquired clinical isolates (*n*=167) and environmental isolates (*n*=276) in England and Wales. Figure reproduced with permission from Harrison *et al.* (2009).

The distribution of STs in clinical isolates also varies geographically. Various studies analyzing the diversity of clinical *L. pneumophila* isolates have shown that there are worldwide-distributed strains (e.g. ST1) but also strains unique to certain regions (e.g. ST47 to northern Europe; ST211 to Ontario, Canada) (Harrison *et al.*, 2005; Borchardt *et al.*, 2008; Tijet *et al.*, 2010).

1.7 The environmental distribution of *L. pneumophila*

1.7.1 *L. pneumophila* in the natural environment

Shortly after the discovery of *L. pneumophila* in 1976, the bacterium was detected in almost all of the 267 freshwater sites, including lakes, rivers and wet soil, investigated in the USA (Fliermans *et al.*, 1979; Fliermans *et al.*, 1981). A number of studies have since confirmed the presence of *L. pneumophila* in freshwater environments globally (Joly *et al.*, 1984; Ortiz-Roque & Hazen, 1987; Pastoris *et al.*, 1989; Verissimo *et al.*, 1991; Lawrence *et al.*, 1999). The bacterium has also been found in marine and estuarine environments (Ortiz-Roque & Hazen, 1987; Palmer *et al.*, 1993; Heller *et al.*, 1998) and soil (Wallis & Robinson, 2005; van Heijnsbergen *et al.*, 2014).

In aquatic environments, *L. pneumophila* can exist in a range of forms including as an intracellular parasite of protozoa, a free-living bacterium or a constituent of biofilms (Marrao *et al.*, 1993; Hay *et al.*, 1995; Fields, 1996; Atlas, 1999; Desai *et al.*, 1999; Murga *et al.*, 2001), although protozoal infection is required for replication (Abu Kwaik *et al.*, 1998). *L. pneumophila* has also been shown to enter into a VBNC form in low-nutrient environments (Steinert *et al.*, 1997) including after the application of biocide treatments

(Garcia *et al.*, 2007; Alleron *et al.*, 2008). In this form, *L. pneumophila* cannot be grown on standard growth media but retains cellular integrity and metabolic activity (Ducret *et al.*, 2014). It has been shown that *L. pneumophila* in this form (as induced by heat-treatment) is not infectious for human cell lines but can be resuscitated to an infectious form by addition of *Acanthamoeba polyphaga* (Epalle *et al.*, 2015). Overall, the abundance of *L. pneumophila* is probably at least partly explained by its ability to survive in extreme ranges of environmental conditions including temperatures ranging from 4-63°C (Fliermans *et al.*, 1981; Wadowsky *et al.*, 1985; Heller *et al.*, 1998; Atlas, 1999). The association of *L. pneumophila* with biofilms also enhances its resistance to biocides (Green, 1993; Kim *et al.*, 2002).

1.7.2 The colonisation of man-made water systems by *L. pneumophila*

The emergence of Legionnaires' disease in the 20th century is most likely due to the colonisation of artificial water systems by *L. pneumophila* (Fields *et al.*, 2002). It is from these man-made environments that people usually become infected with the bacterium although infection from natural hot springs is increasingly being recognised (Ito *et al.*, 2002; Lin *et al.*, 2007). The colonisation of water systems by *L. pneumophila* likely depends on a number of factors including temperature, sediment accumulation and the presence of other microflora (Stout *et al.*, 1985).

Hot and cold water systems of large buildings such as hospitals and hotels are particularly at risk of *Legionella* colonisation. Such systems comprise a complex pipe network with a large number of outlets, and it can be difficult to maintain sufficient water temperatures throughout the system to successfully control *Legionella* (Orsi *et al.*, 2014). Pipes can also be prone to the accumulation of biofilms and stagnant water, particularly where dead ends exist in the network. Several studies of hotel water systems in Europe have shown that *Legionella* colonisation is common and affects 27-75% hotels (Alexiou *et al.*, 1989; Leoni *et al.*, 2005; Borella *et al.*, 2005). Contaminated hospital water systems have been linked to a number of nosocomial outbreaks of Legionnaires' disease (Cordes *et al.*, 1981; Arnow *et al.*, 1982; Graman *et al.*, 1997). *Legionella* has also been isolated from private residences and one study of apartment buildings in Finland showed that shower water contained the highest concentration of *Legionella* of any household outlet (Zacheus & Martikainen, 1994).

A large proportion of Legionnaires' disease outbreaks are also associated with cooling towers. Generally, cooling towers linked to disease cases are associated with poorly maintained systems, a lack of control measures and untrained personnel (Mouchtouri *et al.*, 2010).

Finally, a study of the bacterial content of drinking water showed that *L. pneumophila* is present in 3-33% of drinking water samples, and proposed that drinking water could therefore represent another important source of infection (Rusin *et al.*, 1997). However, a more recent metagenome-based study of the microbiome of drinking water in the United States showed that just 0.31% of annotated proteins present in free-chlorine-treatment drinking water samples were assigned to the *Legionella* genus, and only 0.09% in monochloramine-treated drinking water (Gomez-Alvarez *et al.*, 2012). Another study that performed 16S rRNA sequencing on pre-treated and treated drinking water samples in China also found that <2% rRNA reads belong to the family Legionellaceae (Chao *et al.*, 2013). Indeed, relatively few cases of Legionnaires' disease associated with drinking water have been reported and the majority of these have been nosocomial (Kool *et al.*, 1999).

1.7.3 The control of *L. pneumophila* in man-made water systems

The control of *Legionella* in artificial water systems is crucial to prevent cases of legionellosis. It is recognised that total eradication of *Legionella* from some water systems is very difficult (Marchesi *et al*., 2010) and thus the focus is on controlling the bacteria so that they are present at only very low concentrations. In some countries, including the UK, employers and those responsible for public premises are required to adhere to measures aimed at controlling *Legionella* in water systems. Since legionellosis is believed to be preventable given adequate implementation of control measures, companies and individuals are liable to be sued in the event of disease cases.

The primary method used to control *Legionella* is the regulation of water temperature (Muraca *et al.*, 1990). *Legionella* generally replicates between 20-45°C, and thus the storage and distribution of water within this temperature range should be avoided. Cold water should be stored and distributed at 20°C or lower, and hot water should be stored at 60°C and distributed at a minimum of 50°C (HSE, 2013).

In the UK, it is recommended that high-risk systems such as cooling towers, evaporative condensers and spa pools be tested for *Legionella* at least quarterly (HSE, 2013). Personnel responsible for hot and cold water systems are required to assess the risks of their system and routine microbiological testing may be required. While there is no known safe level of *Legionella,* some studies have shown a significantly increased risk of disease when concentrations exceed 10^3-10^4 colony forming units/litre (CFU/L) in hot and cold water distribution systems (Rota *et al.*, 2004; O'Loughlin *et al.*, 2007). In the UK, counts of >100 CFU/L in piped water systems warrant a review of the control measures and possible disinfection (HSE, 2013).

A number of disinfection methods have been used with varying success to decontaminate water systems. A heat-flushing method is sometimes used as a short-term measure in outbreak situations although its effects are only temporary (Zacheus & Martikainen, 1996). Other methods include copper-silver ionization, chlorine dioxide, monochloramine, point-of-use filtration and ultra-violet (UV) light (Yu *et al.*, 1993; Lin *et al.*, 2011).

1.8 Whole genome sequencing technologies

1.8.1 The history of sequencing

Methods to sequence DNA were pioneered by Frederick Sanger and his colleagues in the 1970s using chain termination technology (Sanger & Coulson, 1975). Using this approach, known as "Sanger sequencing", they sequenced the first DNA genome, that of bacteriophage Φ-X174 which has just 5386 base pairs (Sanger *et al.*, 1977). Maxam and

Gilbert also devised a method based on chemical modification of DNA and subsequent cleavage at specific bases (Maxam & Gilbert, 1977) and initially this was more popular than the Sanger method due to its lack of requirement for a cloning step. However, it was the chain termination method that became the gold standard for the next three decades, due to its high efficiency and low use of toxic materials compared with the Maxam-Gilbert method.

In the 1990s the Sanger method, coupled with a "shot-gun sequencing" approach, was used to sequence a number of landmark genomes including the first bacterial genome, *Haemophilus influenza*, in 1995 (Fleischmann *et al.*, 1995), the first eukaryotic genome, *Saccharomyces cerevisiae*, in 1996 (Goffeau *et al.*, 1996), and the first animal genome, *Caenorhabditis elegans*, in 1998 (the *C. elegans* Sequencing Consortium, 1998). Shot-gun sequencing involves Sanger sequencing many overlapping DNA fragments and using computational methods to assemble overlapping fragments into contigs (Green, 2001).

The original Sanger method, although with dramatically improved fluorescently labelled terminators and automated laser detectors, also facilitated the sequencing of the human genome (International Human Genome Sequencing Consortium, 2001). This was a huge international endeavour requiring over ten years, the efforts of hundreds of scientists, and a cost of $3.8 billion (Tripp & Grueber, 2011). The completion of the human genome show-cased to the scientific community the enormous opportunities offered by genome sequencing. However, it was also clear from the tremendous resources used by the Human Genome Project (HGP) that quicker, cheaper and more high-throughput technologies were required if genome sequencing was to become somewhat routine. Thus the completion of the HGP provided the stimulus for development of a new wave of more sophisticated methods known as "next-generation sequencing" (NGS) technologies.

1.8.2 Second generation sequencing technologies

In the last decade, a variety of second generation or NGS technologies have been developed. These have dramatically reduced the costs and time of genome sequencing and allowed massively parallel analysis (Shendure & Ji, 2008). A major advantage of the new technologies is that they do not require bacterial cloning of DNA fragments and instead rely on library preparation in a cell-free system (van Dijk *et al.*, 2014). They also facilitate the sequencing of up to millions of DNA fragments in parallel and, crucially, sequence every base multiple times reducing the number of errors in the final consensus sequence.

The first commercially available NGS technology was a pyrosequencing method, released by 454 Life Sciences (now Roche) in 2005 (Margulies *et al.*, 2005). Rather than using chain termination with dideoxynucleotides as in Sanger sequencing, pyrosequencing relies on the detection of pyrophosphate released during base incorporation. Originally, the method produced approximately 200,000 reads (~20Mb) of 110 base pairs (bp) (van Dijk *et al.*, 2014). In 2007, new technologies were released by Solexa (now Illumina) and Applied Biosystems (now Life Technologies) which produced many more reads than 454 although the reads generated were just 35bp long (Valouev *et al.*, 2008; van Dijk *et al.*, 2014). Subsequently, in 2010, Ion Torrent (now Life Technologies) released a system called the Personal Genome Machine (PGM). This uses similar technology to 454 sequencing but relies on proton, rather than pyrophosphate, release during nucleotide incorporation and furthermore uses semiconductor technology rather than imaging methods for detection. Overall, the system provided higher speed, and a smaller and more affordable sequencer than the previously released methods.

In recent years, there has been enormous competition amongst NGS developers that has contributed to rapidly improving technologies and plummeting sequencing costs for scientists. Within a decade, the per-base cost of DNA sequencing decreased by approximately 100,000-fold, a rate far outpacing the technological advance seen in the semiconductor industry as described by Moore's law (Lander, 2011). Consequently, NGS platforms are now widely available to even small research laboratories. Illumina is currently the leading NGS platform, offering the highest throughput and lowest per-base cost (Liu *et al.*, 2012). Almost all WGS data produced for this project has been generated using the Illumina HiSeq platform.

1.8.3 Third generation sequencing technologies

In recent years, a new third generation of sequencing technologies has been emerging which promises faster run times, higher throughput, a requirement for only a small amount of starting DNA, lower cost and longer reads (Schadt *et al.*, 2010). While not all of these criteria have been met yet, the two notable technologies that are now available are the single-molecule real-time (SMRT) sequencing method of Pacific Biosciences (PacBio) and nanopore sequencing using the MinION sequencing device produced by Oxford Nanopore, both of which are capable of producing long reads. Released in 2011, the PacBio RS was the first long-read sequencer commercially available and works by using zero-mode waveguide (ZMW) nanostructure arrays to observe base incorporations into a growing DNA strand (Eid *et al.*, 2009). The sequencing technique also provides information on base modifications such as methylation. However, the PacBio system has a high capital cost as well as a higher cost per base (Quail *et al.*, 2012; Rhoads & Au, 2015), limiting its current usage to a few sequencing centres. Meanwhile, the MinION, released in 2014, is the first device to use nanopore sequencing, and has the major advantage of being portable and easy to use. It also has a low capital cost, can be run on a standard internet-connected laptop using USB connectivity and allows real-time analysis while data is being generated. These advantages are likely to facilitate its uptake by many laboratories in the public health setting (Judge *et al.*, 2015). Importantly, both PacBio and MinION sequencing technologies are capable of producing long reads in the order of tens of kilobases (Laver *et al.*, 2015; Koren & Phillippy, 2015), in contrast to the maximum paired-end read length of 250bp provided by the Illumina HiSeq 2500. Since the reads are usually longer than repetitive regions, sequence assembly is considerably simplified and has been shown to result in a single contiguous sequence for many bacterial genomes (Koren *et al.*, 2013; Loman *et al.*, 2015). However, both technologies are currently hindered by high error rates (Quail *et al.*, 2012; Laver *et al.*, 2015) and so far have often been used in conjunction with more accurate Illumina sequencing data to counter this problem (Laver *et al.*, 2015).

1.8.4 Bioinformatic advances

The advent of second and third generation sequencing technologies has required the development of a significant number of new bioinformatics tools for data analysis. In particular, the switch to short DNA reads with second-generation tools from the original long reads (>500 base pairs) generated by Sanger sequencing, and subsequently the production of long, error-prone reads generated by third generation technologies, required new algorithms. The enormous increase in sequencing throughput also mean that tools were required to process vast amounts of data, often many terabytes in a single experiment.

The applications of second generation sequencing have mainly focused on mapping short reads to existing complete genome sequences and calling single nucleotide polymorphisms (SNPs) against the reference, a process that is used extensively in this thesis. Various alignment software has been developed including Bowtie, Burrows-Wheeler Aligner (BWA) and SMALT, each of which has advantages and disadvantages evaluated in several reviews (Ruffalo *et al.*, 2011; Hatem *et al.*, 2013; Shang *et al.*, 2014). The use of paired-end reads (the result of sequencing both ends of a DNA molecule) can increase the accuracy of mapping since the approximate distance between the two ends is known. Generally, high coverage enables variants to be called at high accuracy. A challenge in mapping short reads is that some may match many different regions of the genome and thus there is usually a subset of reads that cannot be mapped. This can be a particular problem for repetitive regions. However, the longer reads produced by third-generation sequencing technologies are now helping to resolve these problems. Indels can also cause difficulties for alignment tools, some of which allow the insertion or deletion of nucleotides and some of which do not. Indels can result in the calling of both false positive and false negatives SNPs.

Another important application of second and third generation sequencing has been the generation of *de novo* genome assemblies. This is particularly valuable for capturing variants such as insertions, deletions, rearrangements and mobile genetic elements (MGEs) that are not present in the reference genome and would not be detected using a mapping approach. Various assembly software for short-read data has been developed, the most commonly used of which is Velvet (Zerbino & Birney, 2008). However, short reads make it difficult to resolve repetitive sequences and can result in fragmented assemblies (Pop & Salzberg, 2008). Meanwhile, third generation sequencing technologies such as SMRT sequencing and nanopore sequencing now facilitate the assembly of long reads into a very small number of contigs, or even a single contig. This has enabled the production of new complete and circularised reference sequences, and indeed has been used in this thesis for generating reference genomes of several important *L. pneumophila* STs.

1.8.5 Applications of bacterial WGS

Second and third generation sequencing technologies have been widely used to study the evolution and spread of bacterial pathogens through the sequencing of hundreds and even thousands of isolates. Applications have ranged from tracking the transcontinental spread of pathogens (Harris *et al.*, 2010; Beres *et al.*, 2010; Mutreja *et al.*, 2011), the spread of pathogens through communities (Mellmann *et al.*, 2011; Gardy *et al.*, 2011) and hospitals (Lewis *et al.*, 2010; Koeser *et al.*, 2012; Bryant *et al.*, 2013) and identifying person-to-person transmission events (Harris *et al.*, 2010; Bryant *et al.*, 2013; Bosch *et al.*, 2013; Luo *et al.*, 2014). Several WGS studies have also begun to elucidate the implication of clinical interventions, such as antibiotics and vaccines, on bacterial evolution. For example, one study of *Streptococcus pneumoniae* isolates discovered that, after the introduction of the conjugate polysaccharide vaccine, there was a population shift as vaccine-escape isolates emerged (Croucher *et al.*, 2011). Interestingly, the change in population could be traced to the occurrence of capsule-switching events. A further study of over 3000 *S. pneumoniae* isolates subsequently demonstrated that loci associated with antibiotic resistance undergo recombination events more frequently, resulting in rapid spread of resistance through the bacterial population (Chewapreecha *et al.*, 2014).

The sharp decreases in both cost and turn-around time, facilitated by the emergence of NGS technologies, now makes WGS a viable option in public health reference laboratories (Bertelli & Greub, 2013). Applications include pathogen surveillance, antibiotic susceptibility testing as well as typing in outbreak scenarios (Didelot *et al.*, 2012; Kwong *et al.*, 2015). Indeed, in some public health laboratories, WGS now costs less than traditional typing methods, including SBT of *L. pneumophila*, and also yields considerably more information. The major challenge to implementation of WGS-based bacterial typing methods is now posed by the need for scalable and portable classification schemes, as well as specialist computing infrastructure and bioinformatics expertise.

1.9 Application of WGS to *L. pneumophila*

The first genome of *L. pneumophila* was sequenced in 2004 and was that of a clinical isolate, Philadelphia-1, from the original Philadelphia outbreak (Chien *et al.*, 2004). Together with subsequent genomes, these facilitated transcriptional studies using microarrays whereby every mRNA encoded by the genome could be quantified (Bruggemann *et al.*, 2006). These were powerful studies that offered significant insight into the interaction of *L. pneumophila* with its eukaryotic host cell. Later, the advent of NGS allowed much larger numbers of genomes to be sequenced (Underwood *et al.*, 2013; Sanchez-Buso *et al.*, 2014), facilitating the study of *L. pneumophila* diversity and evolution. RNA sequencing also took over from microarrays as the primary tool for studying *L. pneumophila* transcriptomics (Weissenmayer *et al.*, 2011; Sahr *et al.*, 2012). More recently, WGS of *L. pneumophila* has become an important tool for investigating outbreaks of Legionnaires' disease (Reuter *et al.*, 2013; Levesque *et al.*, 2014; Graham *et al.*, 2014; Sanchez-Buso *et al.*, 2016). Discussed below are the general features of the *L. pneumophila* genome, insights into the population structure, diversity and evolution gained from genome sequencing and the application of WGS to *L. pneumophila* outbreak typing.

1.9.1 The structure and features of the *L. pneumophila* genome

The *L. pneumophila* genome is approximately 3.4Mb, contains about 3000 protein-coding genes and has a GC content of 38%. Some isolates have plasmids and these vary significantly in size from that of the Paris strain (132kb) to that of Lens (60kb). Several studies also described chromosomal regions that can be excised and maintained as plasmids (Cazalet *et al.*, 2004; Chien *et al.*, 2004). The content of these mobile elements is variable but have been shown to contain the Lvh type 4 secretion system (T4SS) as well as two new T4SSs (*tra*/*trb*) first described in the *L. pneumophila* Corby genome (Glockner *et al.*, 2008).

A major finding from the initial sequencing of *L. pneumophila* genomes was an unexpectedly large number of proteins with high similarity to eukaryotic proteins or containing eukaryotic domains (Cazalet *et al.*, 2004). Many of these are now known to be secreted by the Dot/Icm secretion system and are involved in manipulating host cell processes to allow intracellular replication (Hubber & Roy, 2010). They have likely arisen through horizontal gene transfer from eukaryotic hosts (de Felipe *et al.*, 2005; Lurie-Weinberger *et al.*, 2010).

Comparisons of multiple *L. pneumophila* isolates have shown that this species possesses remarkable plasticity in its genome content. Strains differ widely in their content of MGEs, plasmids, and even in their repertoire of Dot/Icm effectors (Gomez-Valero *et al.*, 2011). The dynamic nature of *L. pneumophila* genomes can be attributed to the occurrence of recombination and horizontal gene transfer events. Indeed, a comparison of just six *L. pneumophila* isolates suggested that large chromosomal fragments of over 200kb are exchanged horizontally between strains (Gomez-Valero *et al.*, 2011).

1.9.2 The population structure, diversity and evolution of *L. pneumophila*

A collection of 36 *L. pneumophila* isolates, thought to represent most of the known species diversity, were sequenced providing the first detailed snapshot of the population structure (Underwood *et al.*, 2013). A phylogenetic tree based on SNP differences showed that the isolates were split up into distinct clusters often separated by very long branches (**Figure 1.10**). 2172 genes were conserved across all isolates, representing about 70% of the genes in each genome. The remaining "accessory" genes are made up of a wide range of genes, but include large numbers of genes involved in protein transport or secretion, and many involved in mobilising DNA (Underwood *et al.*, 2013).

This study also suggested that different STs of *L. pneumophila* seem to contain highly variable levels of diversity (Underwood *et al.*, 2013). For example, three ST47 isolates,

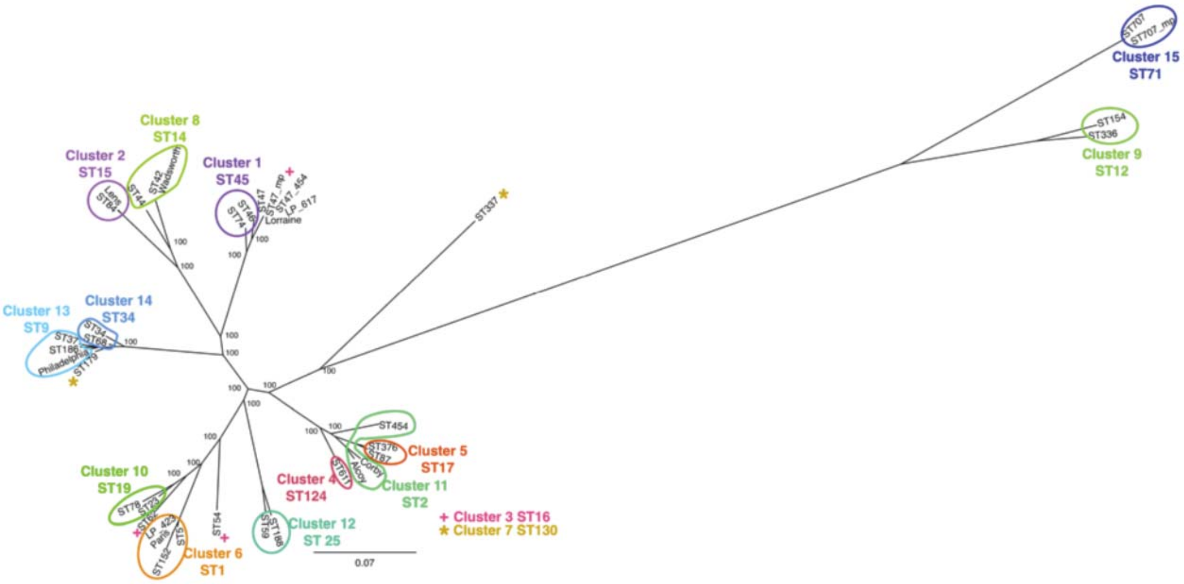

Figure 1.10. Population structure of *L. pneumophila*. A maximum likelihood tree of 36 *L. pneumophila* isolates, based on SNP differences detected by mapping sequence reads to the Corby reference genome. Figure reproduced with permission from Underwood *et al.* (2013).

two from the UK and one from France, which were each isolated in a different year between 2003 and 2006 are a maximum of four SNPs apart. However, two ST1 isolates, one from France and one from the UK, which were isolated two years apart, are distinguished by 280 SNPs. While more data is required to confirm these observations, the authors suggested a number of interesting evolutionary scenarios that could explain this data. One possibility is that some STs simply emerged earlier than others and there has been more time for diversification by genetic drift. Alternatively, it could be that only a small subset of ST47 isolates are able to cause human disease and thus a large amount of diversity goes unnoticed. Differences in recombination frequencies across STs could also account for differences in diversity since recombination events have the potential to bring in many SNPs quickly.

Indeed, various genomic studies have now highlighted the importance of recombination in *L. pneumophila* evolution (Gomez-Valero *et al.*, 2011; Underwood *et al.*, 2013; Sanchez-Buso *et al.*, 2014). In particular, a study of 46 isolates from a single ST (ST578) showed that 98% of SNPs were contained within recombined regions (Sanchez-Buso *et al.*, 2014). These regions were an average of 35.7kb although the largest was 141kb

(Sanchez-Buso *et al.*, 2014). It is likely that recombination aids rapid adaptation to new hosts and environments.

1.9.3 WGS in outbreak investigations

The feasibility of using WGS to discriminate outbreak isolates from concurrent non-outbreak isolates during outbreak investigations of Legionnaires' disease was first demonstrated in a retrospective study (Reuter *et al.*, 2013). Two clinical and three environmental isolates were found to cluster very closely (<15 SNPs were found between the five isolates) and thus considered to be the outbreak isolates. Meanwhile, a third patient could be excluded along with two more environmental isolates based on the large number of SNP differences observed between these and the putative outbreak isolates. These observations were consistent with the conclusions made from the original investigation using epidemiological information and SBT data (Reuter *et al.*, 2013). Further studies have since successfully used WGS to investigate outbreaks (Levesque *et al.*, 2014; Graham *et al.*, 2014; McAdam *et al.*, 2014; Moran-Gilad *et al.*, 2015; Sanchez-Buso *et al.*, 2016). Notably, WGS was applied to a cluster of Legionnaires' disease cases in Edinburgh in 2012 in which no environmental source was found (McAdam *et al.*, 2014). The authors discovered that, despite the clinical isolates belonging to the same uncommon ST, ST191, they could be divided into distinct subtypes based on WGS. They hypothesised that the ST191 isolates had likely diversified in the environment for several years prior to the outbreak accounting for the mutation, recombination and horizontal gene transfer events observed between the clinical isolates.

While most studies of *L. pneumophila* outbreaks have used mapping of read data against a reference genome followed by analysis of SNP variation, Moran-Gilad *et al*. (2015) tested a scaled-up MLST approach known as core genome MLST (cgMLST), utilising 1521 core genes, rather than the usual seven in traditional MLST. The main advantage of MLST is its ease of standardisation and portability, since each isolate can be assigned a "type" based on their combination of alleles, which is either the same or different to that of other isolates. By extracting the gene sequences from the *de novo* assemblies, the authors compared isolates based on the number of allele differences, rather than the number of SNPs. While the study showed that epidemiologically related and unrelated isolates could be readily distinguished by their core gene profiles, some allele differences were seen among known related isolates due to a small number of SNPs in the core genes. The authors therefore suggested that a threshold of four allele differences could be used in defining a "type". However, the use of a threshold would also require a clustering algorithm that would likely need to be re-run each time isolates are typed, reducing the simplicity and scalability of this method.

Apart from the mapping and cgMLST approaches, there are also various whole-genome based typing methods that have been applied to other bacteria. These include a comparison of the k-mer (a short DNA sequence of k nucleotides in length) content of isolates and analysis of the pan-genome content (Leekitcharoenphon *et al.*, 2014). As more laboratories start to use WGS typing approaches in outbreak investigations, it is important to evaluate the advantages and disadvantages of different approaches and develop standardised methods for each species. Development of an optimal WGS-based typing method for *L. pneumophila* will require an in-depth understanding of the diversity in *L. pneumophila* populations at all levels, ranging from the individual patient to the global population structure, in order to achieve an appropriate balance between the need for higher discrimination between isolates and epidemiological concordance.

1.10 Thesis outline

The overall aims of the project are:

1) To investigate the diversity and evolution of *L. pneumophila* using WGS, in order to improve our understanding of how this environmental bacterium has emerged as an important human pathogen.
2) To explore how WGS can be used in a clinical setting to aid outbreak detection and resolution.

Specifically, the first results chapter investigates the diversity and emergence of five major disease-associated STs (1, 23, 37, 47, 62), which together account for almost half of all Legionnaires' disease cases in Europe. As these five lineages have emerged independently from within a diverse species, the chapter also explores whether there are signs of convergent evolution that could explain their predominance in human disease.

The second results chapter explores the dynamics of homologous recombination within the major disease-associated STs, a process that is found to be a significant contributor to *L. pneumophila* diversity in the first results chapter as well as in other studies. It investigates whether there are "hotspots" of homologous recombination within the genome that could provide novel insights into the selection pressures of this bacterium. By predicting potential donor lineages of recombined regions, the chapter also investigates the extent to which homologous recombination occurs within and between within major lineages of *L. pneumophila*.

The third results chapter of this thesis evaluates a number of WGS-based methods for the epidemiological typing of *L. pneumophila*. Using published guidelines and a test population used for evaluating previous *L. pneumophila* typing schemes, the chapter compares their performance to current gold standard methods, and proposes the most suitable methodology for future development.

Finally, the last results chapter investigates whether WGS can be used in nosocomial investigations to support or refute suspected links between hospital water systems and cases of Legionnaires' disease.