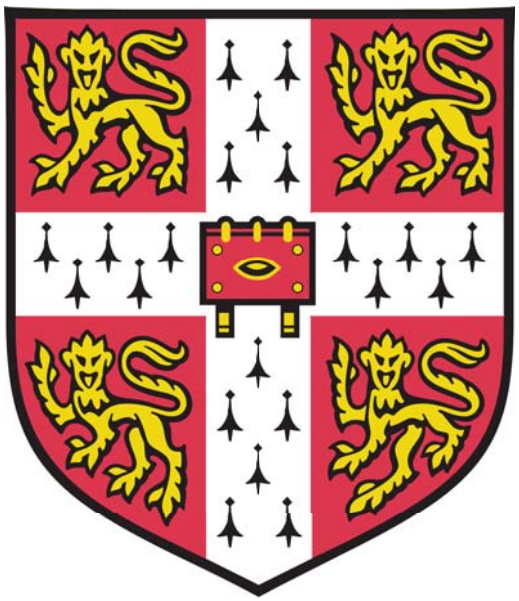
The diversity of disease-causing and environmental *Legionella pneumophila*

Sophia David

Clare College, University of Cambridge,
Wellcome Trust Sanger Institute
& Public Health England

August 2016



This dissertation is submitted for the degree of
Doctor of Philosophy

The diversity of disease-causing and environmental *Legionella pneumophila*

Sophia David

Abstract

Legionella pneumophila is a species of Gram-negative bacteria that survives in natural freshwater and soil habitats. It also now colonises modern, man-made water systems from which humans can become infected, usually *via* inhalation of contaminated aerosols. Infection can result in a severe and potentially fatal pneumonia known as Legionnaires' disease. This thesis uses whole genome sequencing (WGS) of large sample collections of *L. pneumophila*, firstly, to develop our understanding of the evolution and emergence of this important human pathogen. Secondly, it explores how WGS data can be used in a clinical setting for outbreak detection and resolution.

To aid outbreak investigations and surveillance, *L. pneumophila* isolates are currently subdivided into "sequence types" (STs) using sequence-based typing (SBT), a method analogous to multi-locus sequence typing (MLST). Analysis of the SBT database has shown that a large proportion of Legionnaires' disease cases are caused by just a small number of STs, despite much higher diversity being observed in commonly implicated environmental sources of *L. pneumophila*. The first part of this thesis describes the application of whole genome sequencing (WGS) to understand the emergence of five major disease-associated STs (1, 23, 37, 47 and 62) within the context of the *L. pneumophila* species. Phylogenetic analysis showed that all five STs have very limited diversity (excluding recombined regions), they have emerged recently, and have since dispersed rapidly and internationally. The findings support the idea that humans are not "accidentally' infected by any *L. pneumophila* strain that happens to be present in an environmental source, but rather are infected by specific clones that are more efficient at human infection.

Analysis of the five major disease-associated STs revealed that recombination accounts for >95% of diversity in some lineages. The next part of the thesis characterises the dynamics and biological impact of homologous recombination on *L. pneumophila*

evolution. This revealed novel insights into the selection pressures of *L. pneumophila* through the identification of hotspot regions, and provided a greater understanding of the genomic flux within the species.

In addition to its use in studies of bacterial evolution and pathogenicity, WGS also now represents a promising typing tool that could supplement or even replace current methods such as SBT. In the next part of this thesis, several WGS-based methods are evaluated for the epidemiological typing of *L. pneumophila*. A 50-gene core genome multi-locus sequence typing (cgMLST) scheme is proposed as the optimal method for future development since it substantially improves upon the discrimination achieved by SBT whilst maintaining high epidemiological concordance.

The final part of this thesis explores whether WGS can be used in nosocomial investigations to support or refute suspected links between hospital water systems and cases of Legionnaires' disease. We focused on cases involving ST1, which is a major nosocomial-associated strain. Overall, we found that WGS can be used successfully to aid investigations but that deep hospital sampling is required. This is due to the potential co-existence of multiple populations within the hospital water system, the existence of substantial diversity within hospital populations, and the similarity of hospital isolates to local populations.

Declaration

This thesis describes work carried out between May 2013 and August 2016 under the supervision of Professor Julian Parkhill at the Wellcome Trust Sanger Institute and Dr Timothy Harrison at Public Health England. I am a member of Clare College, University of Cambridge.

I hereby declare that this dissertation is my own work and contains nothing that is the outcome of work performed in collaboration except where specifically indicated at the beginning of each chapter.

No part of the dissertation has been submitted for any other qualification and it does not exceed the word limit (60,000) stipulated by the Biological Sciences Degree committee.

Sophia David
August 2016

Acknowledgements

I would firstly like to thank Julian Parkhill and Tim Harrison for giving me the opportunity to perform this work and for always making time for me. They are undoubtedly the most knowledgeable and inspirational supervisors I could have wished for. Simon Harris has also been a wonderful mentor who has provided endless advice throughout my PhD. Furthermore, I am very grateful to Anthony Underwood for his time and patience in teaching me the ropes of bioinformatics at the start of my project, which was completely invaluable, but also for his kind support and advice throughout. Thank you to each of you.

I also wish to thank the numerous colleagues and collaborators who have made this work possible including Massimo Mentasti, Baharak Afshar, Rediat Tewolde, Leonor Sánchez-Busó, Carmen Buchrieser, Christophe Rusniok, Sophie Jarraud and Christophe Ginevra. The sequencing and informatics teams at the Sanger Institute and Public Health England have also provided extensive support. I am grateful to my thesis committee, made up of Sharon Peacock, Carl Anderson and Paul Kellam, who have provided valuable guidance. I also thank all members of the Pathogen Genomics team at the Sanger Institute for their friendship and support, and who have made my PhD so enjoyable. In particular, Kate Baker, Sandra Reuter, Claire Chewapreecha, Josie Bryant and Michelle Toleman have all offered kind support and encouragement for which I am very grateful.

On a personal note, I would like to thank the friends I have made during my time in Cambridge – in particular, Mia Petljak, Neneh Sallah, Pinky Langat, Kirsty Dundas, Jane Patrick, Jess Forbester and James Hadfield - who always brighten up my day and have made my PhD very memorable. I am hugely grateful to my partner, Feyruz, for his wonderful support and understanding, and for making me laugh endlessly. My final thanks are to my family and in particular to my mum, who continues to support me in every way she can and inspires me to work hard.

Table of Contents

List of Figures

List of Tables

Abbreviations

WTSI, Wellcome Trust Sanger Institute

PHE, Public Health England

CDC, Centers for Disease Control and Prevention

ECDC, European Centre for Disease Prevention and Control

sg, serogroup

mAb, monoclonal antibody

LCV, *Legionella* containing vacuole

ER, endoplasmic reticulum

RER, rough endoplasmic reticulum

Dot/Icm, defect in organelle trafficking/intracellular multiplication

BCYE, buffered charcoal yeast extract

BAL, bronchoalveolar lavage

IFA, indirect fluorescent antibody

ELISA/EIA, enzyme-linked immunosorbent assay

DFA, direct fluorescent antibody

VBNC, viable but not culturable

PCR, polymerase chain reaction

SBT, sequence-based typing

PFGE, pulsed field gel electrophoresis

AFLP, amplified fragment length polymorphism

MLST, multi-locus sequence typing

ST, sequence type

ESGLI, European study group for Legionella infections

ELDSNet, European Legionnaires' Disease Surveillance Network

EU, European Union

EEA, European Economic Area

HSE, Health and Safety Executive

CFU, colony forming units

UV, ultra-violet

HGP, Human Genome Project

NGS, next-generation sequencing

SMRT, single-molecule real-time

PacBio, Pacific Biosciences

ZMW, zero-mode wavelength

SNP, single nucleotide polymorphism

BWA, Burrows-Wheeler Aligner

T4SS, type 4 secretion system

MGE, mobile genetic element

cgMLST, core genome multi-locus sequence typing

TE, Tris-EDTA

ENA, European Nucleotide Archive

MCC, maximum clade credibility

MRCA, most recent common ancestor

TMRCA, time to most recent common ancestor

HPD, highest posterior density

MLEE, multi-locus enzyme electrophoresis

LPS, lipopolysaccharide

NCBI, National Center for Biotechnology Information

PRR, pattern recognition receptor

BIGSdb, Bacterial Isolate Genome Sequence Database

T, typability

R, reproducibility

E, epidemiological concordance

D, index of discrimination

S, stability

ESCMID, European Society for Clinical Microbiology and Infectious Diseases

ESGEM, ESCMID Study Group on Epidemiological Markers

rMLST, ribosomal multi-locus sequence typing

QC, quality control

SD, standard deviation

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**).

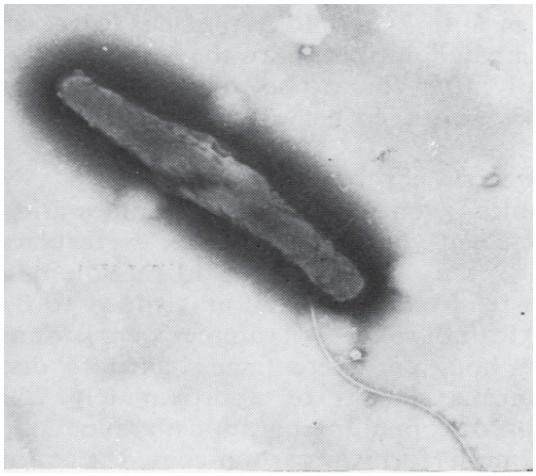


Figure 1.1. Electron microscope image of *L. pneumophila*. The flagellum is negatively stained with 1% phosphotungstic acid. Figure obtained with permission from Rodgers *et al.* (1980).

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,

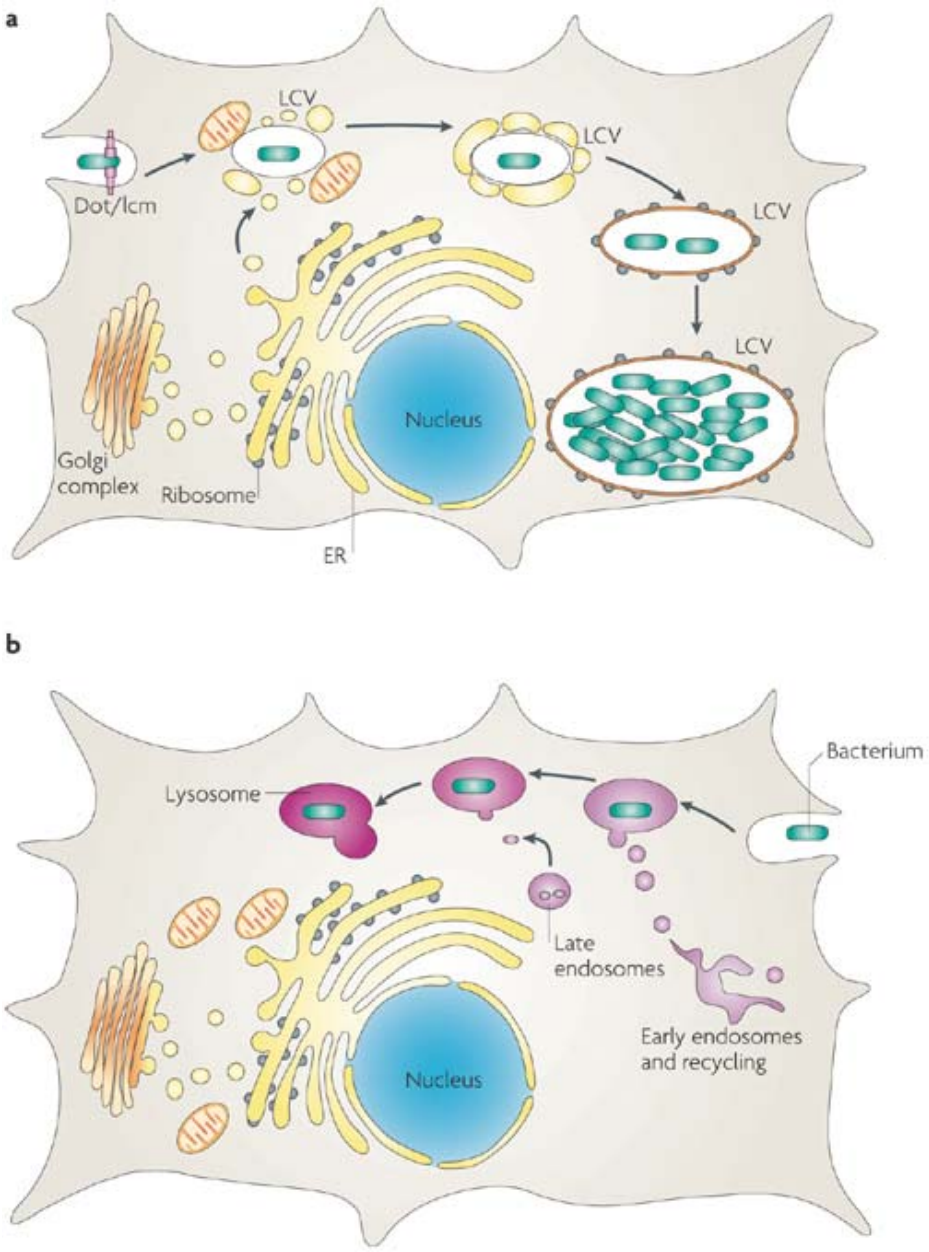


Figure 1.2. The *Legionella*-containing vacuole. a) *L. pneumophila* enters the host cell by phagocytosis, evades delivery to the lysosomal network and immediately establishes a compartment known as the LCV. Endoplasmic reticulum (ER)-derived vesicles and subsequently mitochondria surround the LCV. The 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.

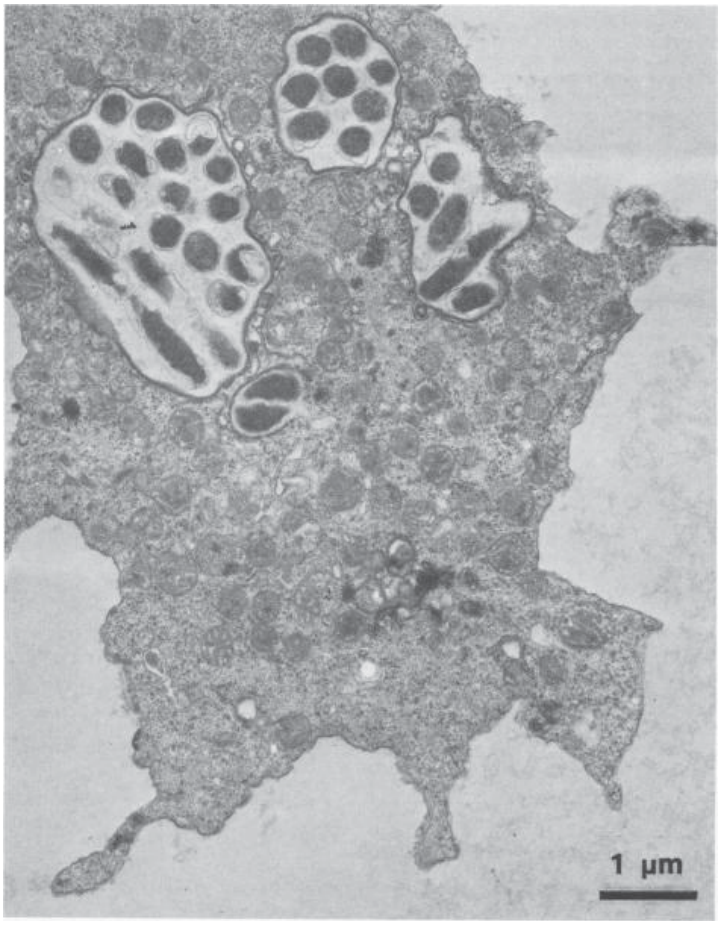


Figure 1.3. Infection of *Acanthamoeba castellani* with *L. pneumophila*. Transmission electron micrograph of *L. pneumophila* contained within LCVs inside *Acanthamoeba castellani* at 48h post-infection. Figure reproduced with permission from Holden *et al.* (1984).

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.

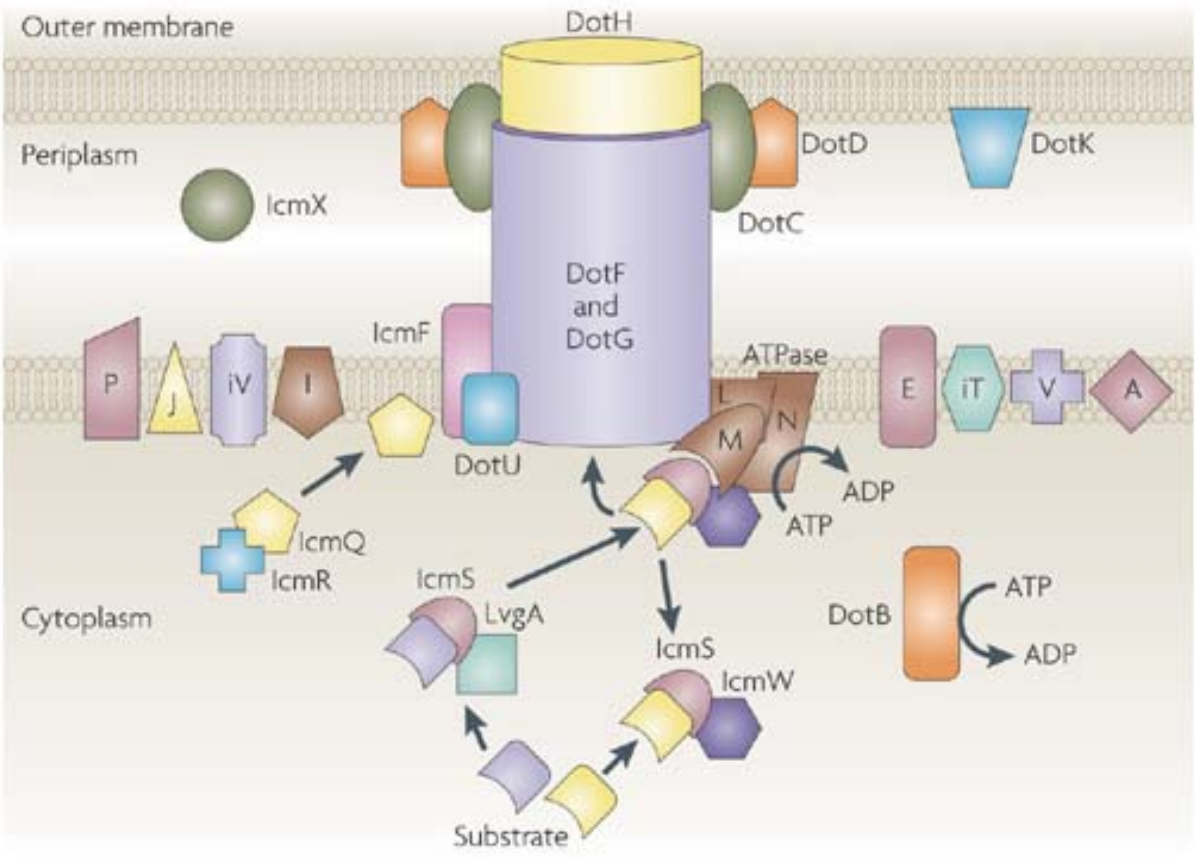


Figure 1.4. Dot/Icm machinery. The components of the Dot/Icm (defect in organelle trafficking/intracellular multiplication) machinery in the bacterial cell envelope. Figure reproduced with permission from Isberg *et al.* (2009).

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).

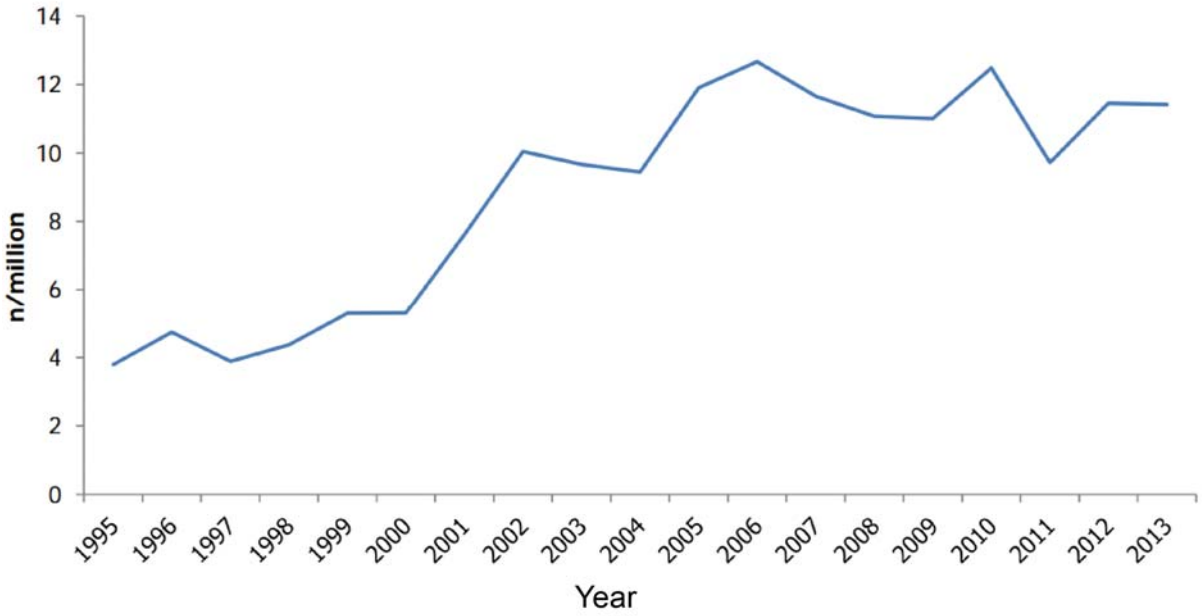


Figure 1.5. Incidence of Legionnaires' disease. The annual notification rates of Legionnaires' disease in the European Union/European Economic Area (EU/EEA) from 1995 to 2013. Figure reproduced with permission from ECDC (2015).

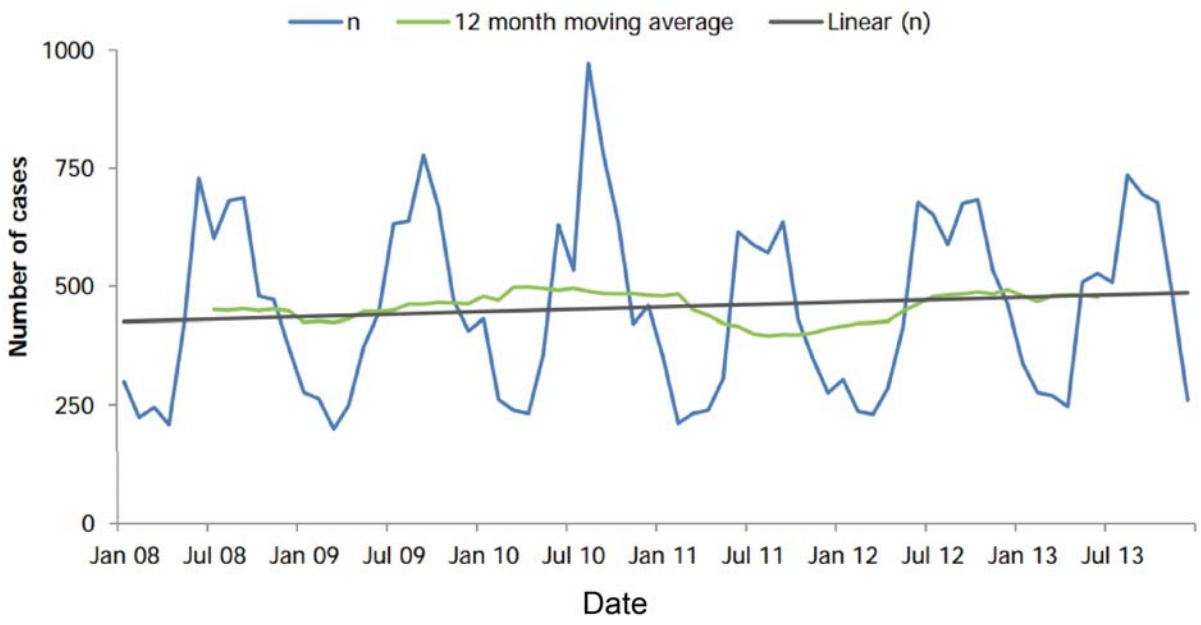


Figure 1.6. Seasonal trend of Legionnaires' disease. Reported cases of Legionnaires' disease by week of onset in the EU/EEA between 2008 and 2013. Figure reproduced with permission from ECDC (2015).

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).

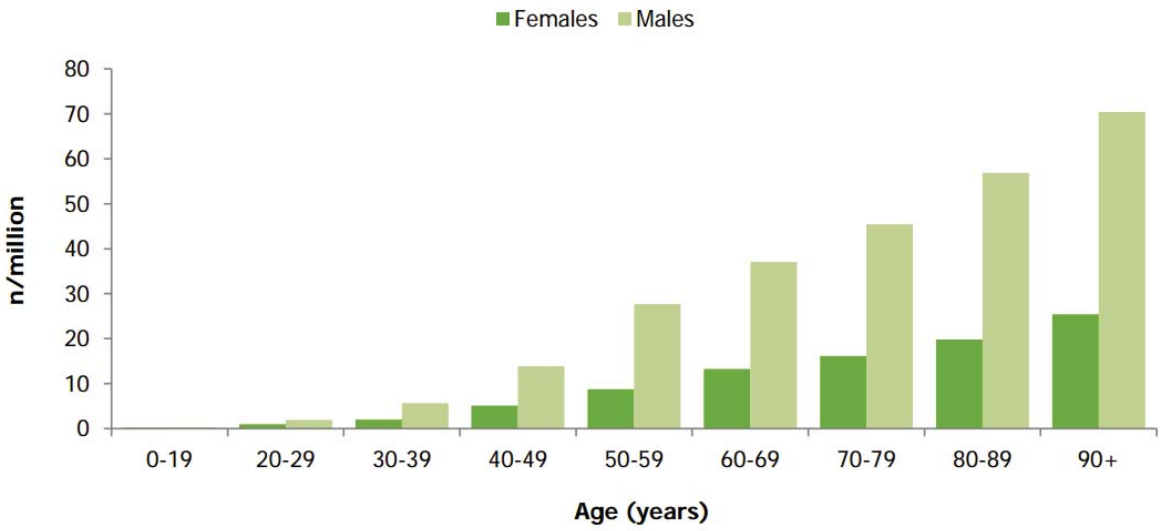


Figure 1.7. Distribution of Legionnaires' disease cases by age and sex. The number of reported cases of Legionnaires' disease per million by gender and age in the EU/EEA in 2013. Figure reproduced with permission from ECDC (2015).

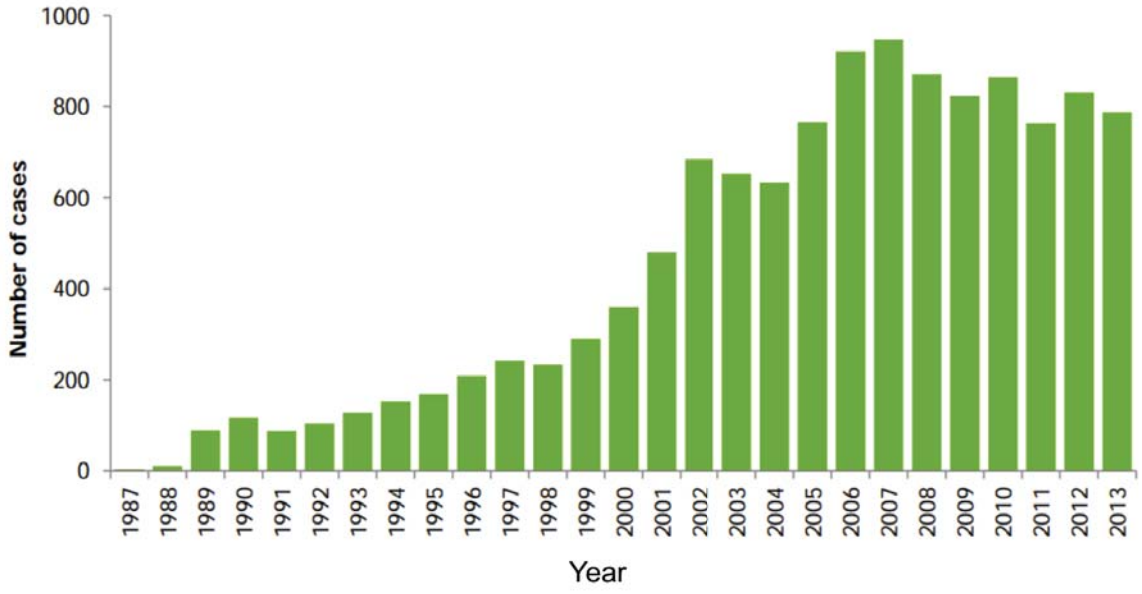


Figure 1.8. Incidence of travel-associated Legionnaires' disease. The number of annually reported cases of travel-associated Legionnaires' disease in EU/EEA member states from 1987 to 2013. Figure reproduced with permission from ECDC (2015).

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.

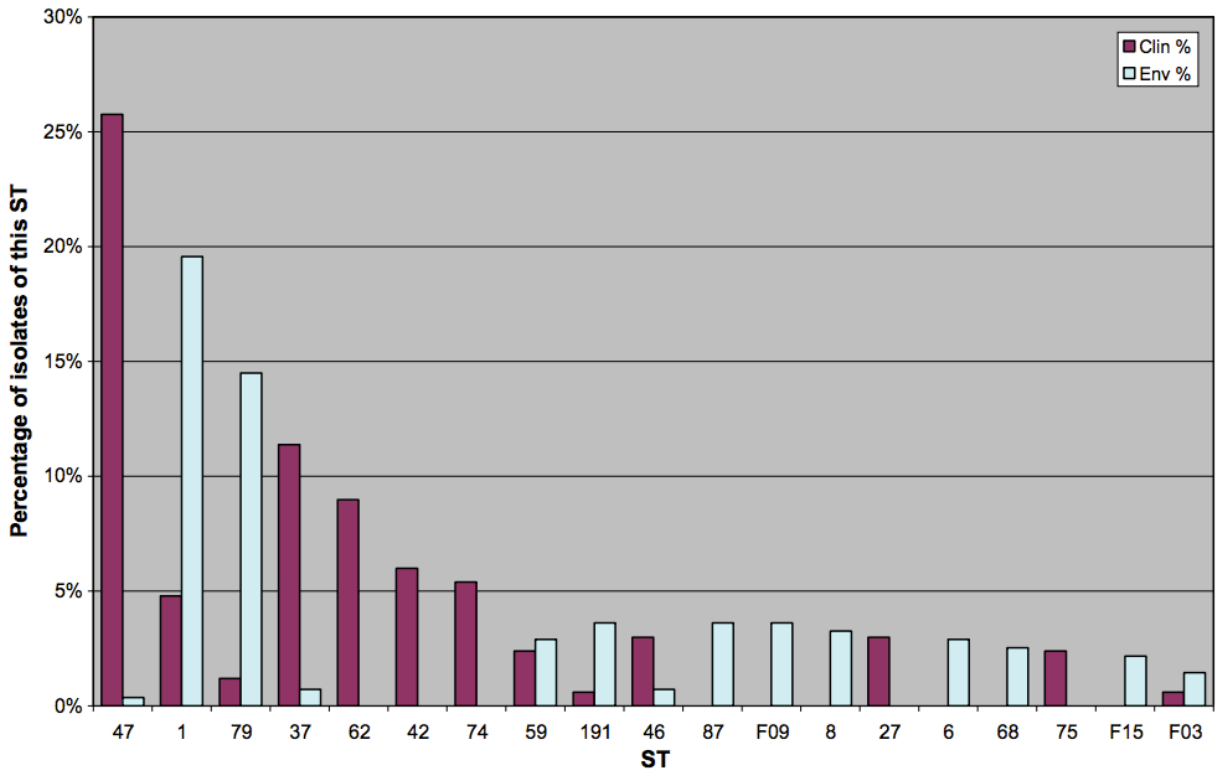


Figure 1.9. Distribution of STs among clinical and environmental isolates (previous page). The prevalence of various STs among community-acquired clinical isolates (*n*=167) and environmental isolates (*n*=276) in England and Wales. Figure reproduced with permission from Harrison *et al.* (2009).

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,

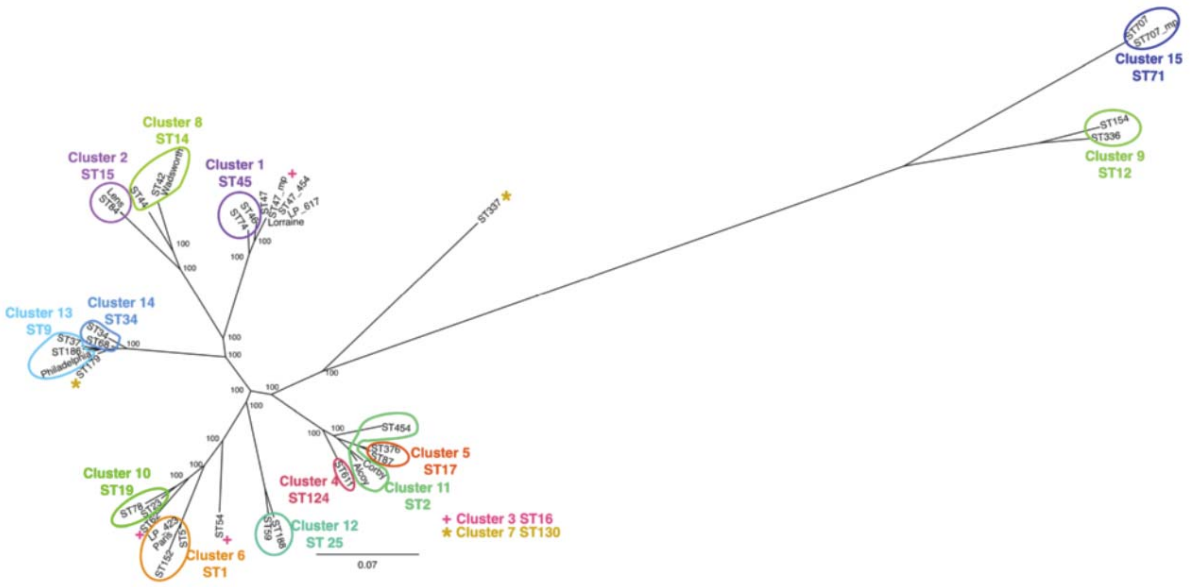


Figure 1.10. Population structure of *L. pneumophila*. A maximum likelihood tree of 36 *L. pneumophila* isolates, based on SNP differences detected by mapping sequence reads to the Corby reference genome. Figure reproduced with permission from Underwood *et al.* (2013).

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.

2. Materials & Methods

This chapter includes methods that were used in several of the results chapters. Many of these methods form part of in-house scripts or pipelines that have been created by the Pathogen Informatics team or members of the Pathogen Genomics group at the WTSI, and these are indicated as such. Methods that are specific to certain analyses are described in the relevant chapter.

2.1 Culture and DNA extraction

All culture and DNA extraction of isolates for this thesis was performed by collaborators. Isolates were grown at 37°C on BCYE agar for 48-72h prior to DNA extraction. DNA was subsequently extracted using either the Wizard (Promega UK, Southampton, UK), PurElute (VH Bio, Gateshead, UK) or DNease Blood & Tissue (Qiagen) kits, according to the manufacturer's instructions. This was eluted in 1x Tris-EDTA (TE) buffer (pH 8.0) and quantified using a Qubit Fluorometer (Life Technologies Ltd, Paisley, UK).

2.2 Whole genome sequencing

All processing and sequencing of genomic DNA was performed by the core sequencing facilities at either the WTSI or PHE, unless stated otherwise in the relevant results chapters. Paired end libraries were created by these teams as described in previous publications (Quail *et al.*, 2012; Dallman *et al.*, 2014) and most samples were sequenced using the Illumina HiSeq platform and paired-end reads of 100 bases. Any deviations from this are also described in the relevant results chapters.

2.3 *De novo* assembly of Illumina sequence data

All assemblies were produced from the Illumina data using a pipeline developed by the Pathogen Informatics team at the WTSI. This firstly uses Velvet Optimiser (http://bioinformatics.net.au/software.velvetoptimiser.shtml) to determine the optimal kmer size to use before using Velvet to produce the assembly (Zerbino & Birney, 2008). The assembly was further improved using SSPACE (Boetzer *et al.*, 2011) to scaffold the contigs of the assembly and GapFiller (Boetzer & Pirovano, 2012) to close gaps of 1 or more nucleotides.

2.4 Control for sample mix-up through determination of sequence type

The sequence type (ST) of each isolate was derived from the *de novo* assembly using an in-house script at the WTSI. This was compared with the ST that the isolate had previously been designated using the standard laboratory protocol for sequence-based typing (SBT), to help verify that the sample had not been involved in a mix-up during the culture, DNA extraction or sequencing procedures (http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php).

2.5 Mapping of Illumina sequence data

Illumina sequence reads (in fastq format) were mapped to different reference genomes using SMALT v0.7.4 (http://www.sanger.ac.uk/science/tools/smalt-0) or BWA-MEM (Li & Durbin, 2009) (see results chapters for details). In either case, an in-house pipeline at the WTSI was used to call bases and identify SNPs using SAMtools (Li *et al.*, 2009), mpileup and BCFtools. Various filters were applied to ensure high accuracy base calling (**Table 2.1**). Any positions that did not pass the filters were called as "N" in the alignment. Additionally, any reads that mapped to more than one region equally well were discarded.

Table 2.1. Filters that were applied to the mapping and base calling of Illumina sequence data against a reference genome.

Filtering criteria	Threshold
Minimum base quality	50
Minimum mapping quality	30 (SMALT); 20 (BWA-MEM)
Minimum number of high quality reads matching base	4 (SMALT); 8 (BWA-MEM)
Minimum number of high quality reads on each strand matching base	2 (SMALT); 3 (BWA-MEM)
Minimum proportion of high quality mapped reads matching base	0.75 (SMALT); 0.8 (BWA-MEM)
Allele frequency	Within 0.5 of 1 (for a SNP) or 0 (for a non-variant)
Minimum strand bias p-value	0.001
Minimum mapping quality bias p-value	0.001
Minimum tail distance bias p-value	0.001

2.6 Phylogenetic analysis

Maximum likelihood phylogenetic trees were constructed based on variable positions within the core genome alignment using RAxML v7.0.4 (Stamatakis, 2006), usually after the removal of recombined regions where recombination detection was possible (see individual results chapters). The GTR+GAMMA method for among site rate variation was used and 100 bootstrap replicates were performed to assess support for nodes unless specified otherwise. In order to scale the branch lengths by the number of SNPs, SNPs were reconstructed onto the phylogeny using accelerated transformation parsimony (Farris, 1970), performed with a script written by Dr Simon R. Harris. Phylogenetic trees were visualised using Figtree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

2.7 Statistical analyses and figures

Statistical analyses were performed using R version 3.0.0 (R Core Team, 2013). Figures were also generated in R and using Adobe Illustrator CS5.

3. Recent emergence of five major disease-associated STs

Declaration of work contributions

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

Publication

The following work has been published:

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

3.1 Introduction

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

Investigation of clinical *L. pneumophila* isolates by various typing methods has revealed that some types cause human infection far more commonly than others. For example, while there are 16 serogroups (sg) currently described, sg 1 was responsible for 83% culture-confirmed Legionnaires' disease cases attributed to *L. pneumophila* in Europe in 2013 (ECDC, 2015). Furthermore, as of 8 July 2016, 2191 STs have been reported to the ESGLI SBT database but a relatively small proportion has been commonly associated with disease. Indeed, analysis of all clinical isolates submitted to the ESGLI SBT database (n=6116) prior to April 2015 found that isolates belonging to just five STs (1, 23, 37, 47, 62) accounted for over 40% clinical isolates submitted to the SBT database from Europe (**Figure 3.1A**). There is no evidence that the high proportion of isolates found in clinical samples belonging to these five STs is a result of laboratory artefacts such as an increased growth of these STs in culture compared with other STs. Data from 2009 to 2014 obtained by SBT on clinical isolates (n=1762) and nested-PCR-based SBT (NP-SBT) performed directly from respiratory samples from patients (n=99) confirmed a similar distribution of these STs among culture-proven and culture-negative but NP-SBT positive patients in France.

One of these five STs, ST1, has been described as a leading cause of Legionnaires' disease from numerous countries worldwide including Canada (Tijet *et al.*, 2010), Japan (Amemura-Maekawa *et al.*, 2010), France (Ginevra *et al.*, 2012), Belgium (Vekens *et al.*, 2012) and Israel (Moran-Gilad *et al.*, 2014). ST1 isolates have been reported to the SBT database from all continents that actively report *L. pneumophila* isolates, and comprise 11.0% of the total including 19.1% of isolates from North America (**Figure 3.1A**). However, several studies have found that ST1 isolates are also found commonly in

environmental samples (Kozak-Muiznieks *et al.*, 2014; Amemura *et al.*, 2012). A study of 443 isolates (including 167 clinical and 276 environmental isolates) obtained between 2000 and 2008 in England and Wales showed that ST1 isolates are more prevalent in environmental samples than clinical samples (Harrison *et al.*, 2009), a finding that was mirrored in the analysis of clinical (*n*=6116) and environmental (*n*=2826) isolates submitted to the SBT database (**Figure 3.1B**).

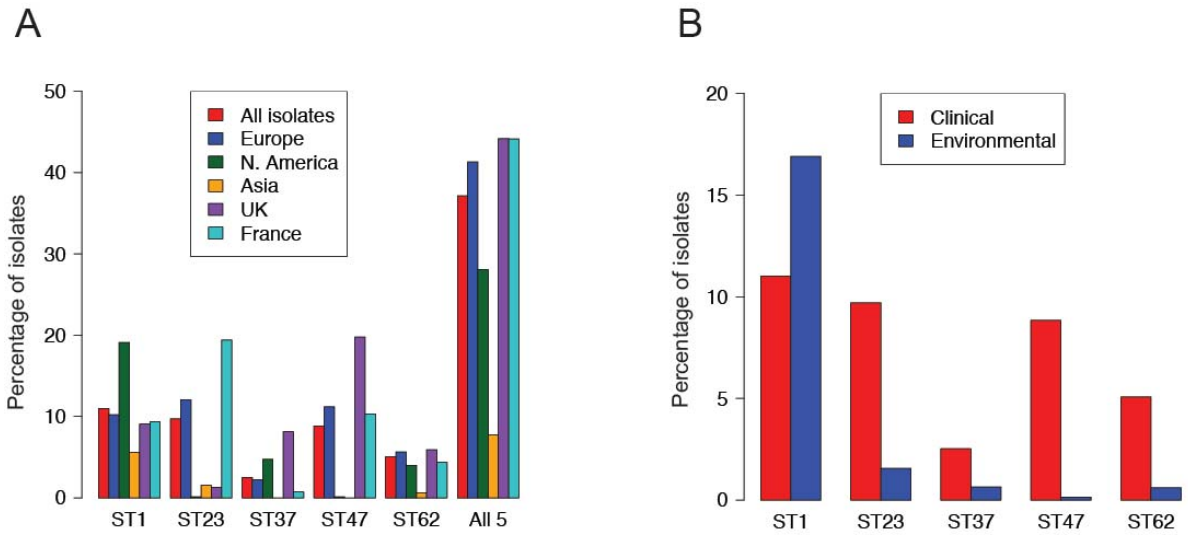


Figure 3.1. Geographical distribution of STs and their prevalence in clinical and environmental samples. A) The percentage of clinical isolates submitted to the ESGLI SBT database from different geographical regions that belong to STs 1, 23, 37, 47 and 62. These data are based on a total of 6116 epidemiologically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

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

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

3.2 Materials & Methods

3.2.1 Bacterial isolates

A total of 364 *L. pneumophila* isolates, of which 35 are previously published and 329 are newly sequenced, were used in this thesis chapter. The previously published isolates include those belonging to 32 STs (**Appendix Table 1**), which were selected as representatives of the known species diversity (Underwood *et al.*, 2013). 337 isolates,

including 327 that are newly sequenced, belong to one of the five major disease-associated STs and include 71 ST1 (or "ST1-derived"), 37 ST23, 72 ST37, 122 ST47 and 35 ST62 isolates (**Appendix Tables 1 & 2**). ST1-derived isolates belong to other STs that are nested within, and thus evolved from, ST1 isolates. All newly sequenced isolates are from the culture collections at PHE, UK or the National Reference Center of *Legionella*, France. Culture and DNA extraction of all isolates was performed as described in *Chapter 2 (Materials & Methods)*.

3.2.2 Whole genome sequencing

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

3.2.3 Mapping of sequence reads and phylogenetic analysis

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

3.2.4 Time-dependent phylogenetic reconstruction

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

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

The roots of the ST1, ST23, ST47 and ST62 maximum likelihood trees (constructed after recombination removal) were established using outgroup isolates. Each tree was subsequently constructed without the outgroup and rooted appropriately. Using the evolutionary rates estimated for the ST37 lineage and the previously published ST578 lineage (Sanchez-Buso *et al.*, 2014) with BEAST, the approximate length of time that it would have taken for the diversity to be acquired in each of the four lineages was estimated. Firstly, accelerated transformation parsimony was used to scale the branches of each phylogenetic tree by the number of SNPs that had occurred (see *Chapter 2 (Materials & Methods)*). The number of SNPs on each branch was then scaled up by the proportion of the genome that had been removed due to recombination, in order to

account for any additional SNPs that may have occurred by *de novo* mutation on top of the recombined regions. Using the two estimated substitution rates and the known sampling dates of all isolates, each root-to-tip distance in the phylogenetic tree was used to calculate the length of time it would have taken for each isolate to have evolved from the common ancestor of the lineage. An estimated age of the tree root was inferred from the mean of these values.

3.2.6 Gene content analysis

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

3.2.7 Searching for evidence of positive selection using CodeML

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

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

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

3.2.9 Identification of recombination donors

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

3.3 Results

3.3.1 Independent emergence of the five STs

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

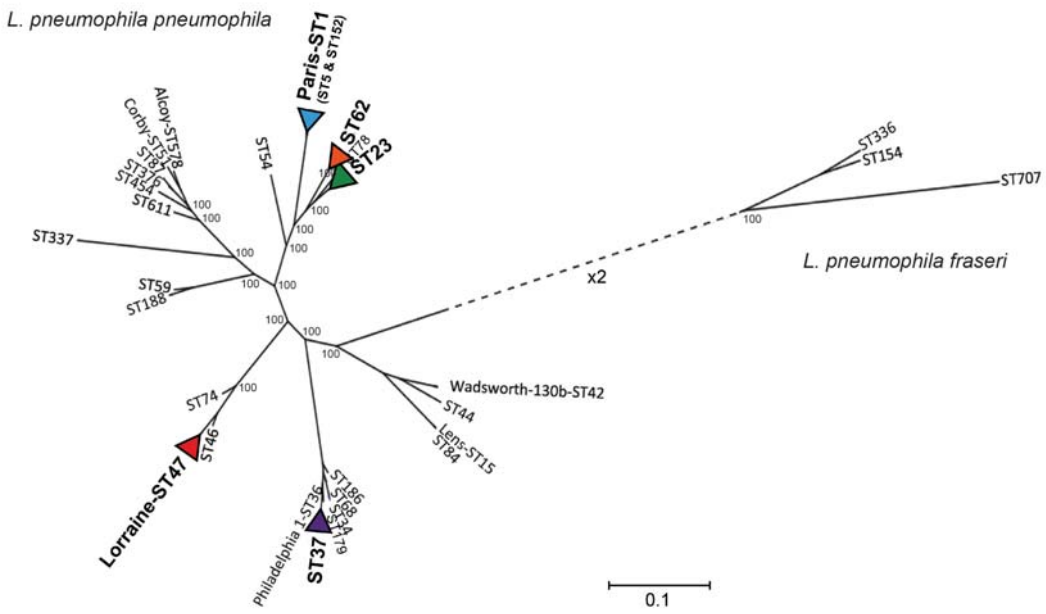


Figure 3.2. Population structure of *L. pneumophila* highlighting five major disease-associated STs of interest (previous page). A maximum likelihood tree of 32 *L. pneumophila* isolates that represent the known species diversity. The five major disease-associated STs, which are highlighted by coloured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

3.3.2 Investigation of the diversity within the five STs

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

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

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

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

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

In contrast to ST47, the total numbers of SNPs detected within STs 1, 23, 37 and 62 were substantially higher (**Table 3.1**). The highest number of SNPs was observed in the globally dispersed ST1 lineage, which has 48,655 SNPs between 71 isolates, and a

maximum difference between any pair of 15,227. The second highest number of SNPs was observed in the ST62 lineage (*n*=33,200), followed by ST23 (*n*=26,945) and ST37 (*n*=14,829). However, visualisation of the SNP distributions in these four lineages using SynTView (**Figure 3.3B-E**) showed that many of the SNPs are found in very close proximity to others, suggesting the occurrence of recombination events.

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

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

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

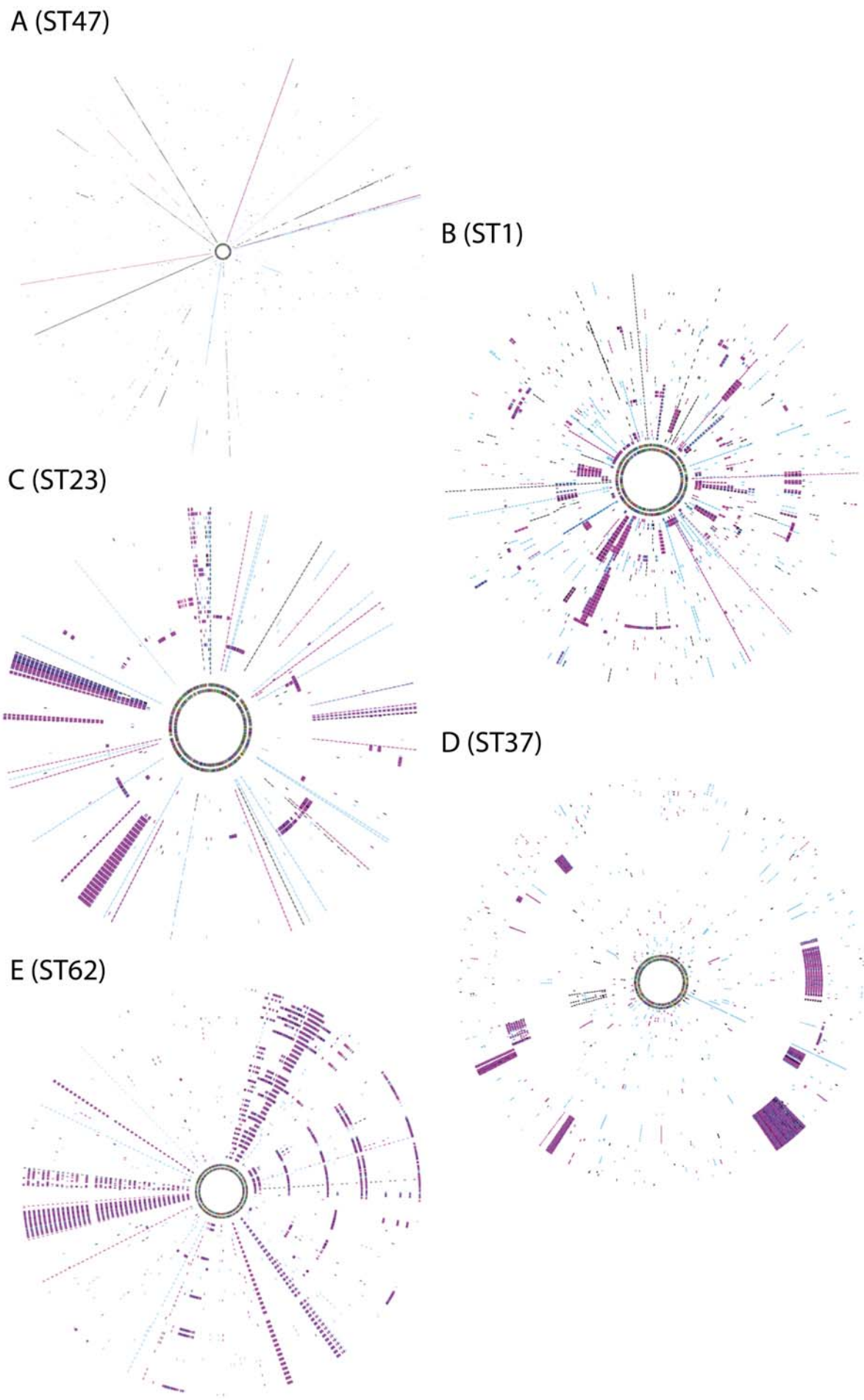

A (ST47)
B (ST1)
C (ST23)
D (ST37)
E (ST62)

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

3.3.3 Dating the emergence of the five STs

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

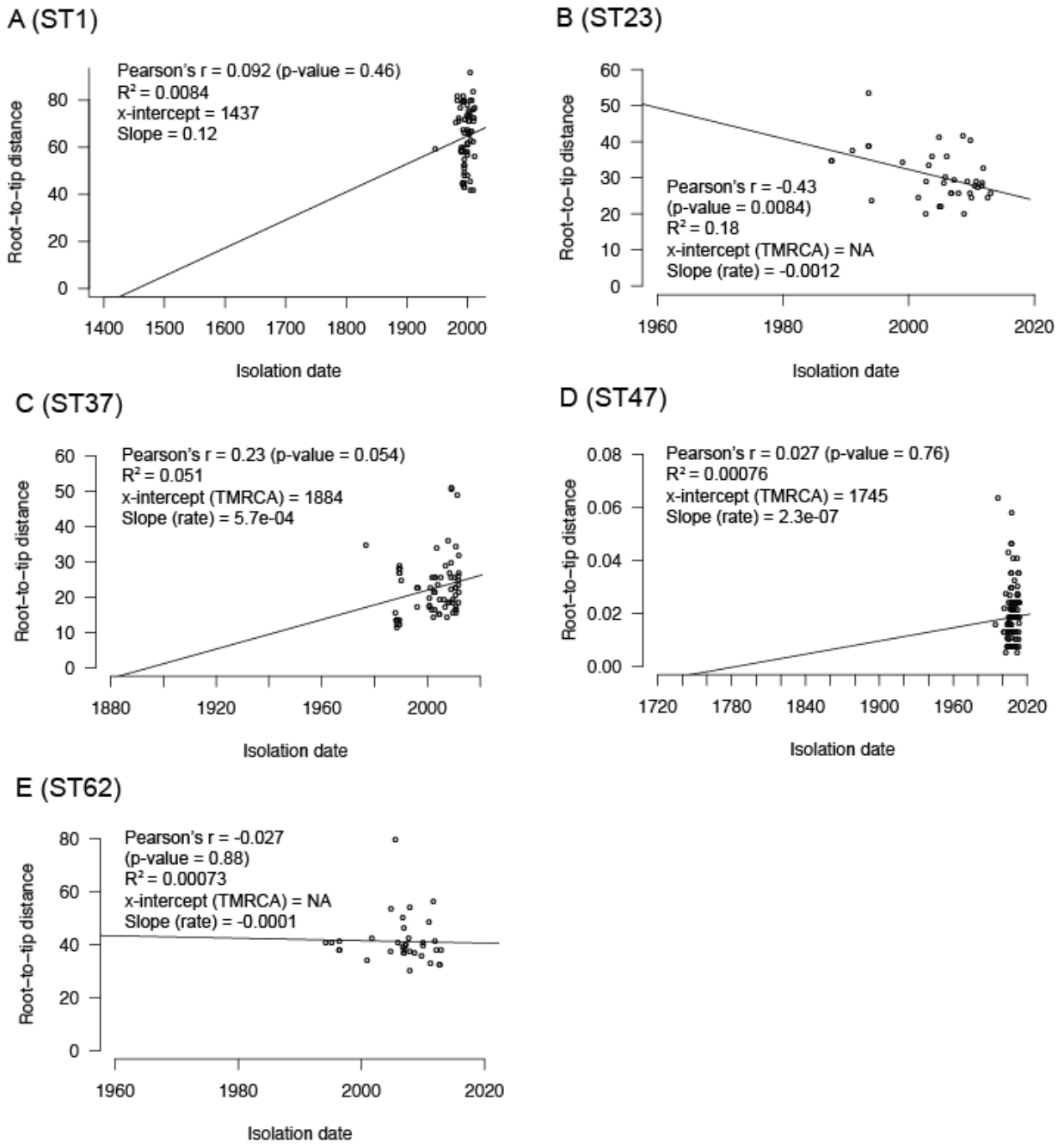


Figure 3.4. Linear regression analyses of root-to-tip distances against sampling date in each of the five STs. Each regression was performed after the removal of SNPs in recombined regions. The correlation coefficient (Pearson's r) is weak in all lineages although slightly higher in the ST37 lineage.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

Given the results of the linear regression analyses, we next attempted to date only the ST37 lineage using an alternative Bayesian coalescent method implemented with BEAST software (Drummond *et al.*, 2012). This allows a relaxed molecular clock (i.e. a

substitution rate that varies between tree branches) to be incorporated into the model, which was predicted to provide a better fit to the ST37 data. Indeed, after testing and comparing a range of model parameters (see *Materials & Methods*), a model that uses an exponential relaxed substitution rate and a Bayesian skyline (variable) population size was found to converge and have the best fit to the data. The model predicted the median age of emergence of the ST37 lineage to be 1979 (95% highest posterior density (HPD) intervals: 1968 to 1985) (**Figure 3.5**).

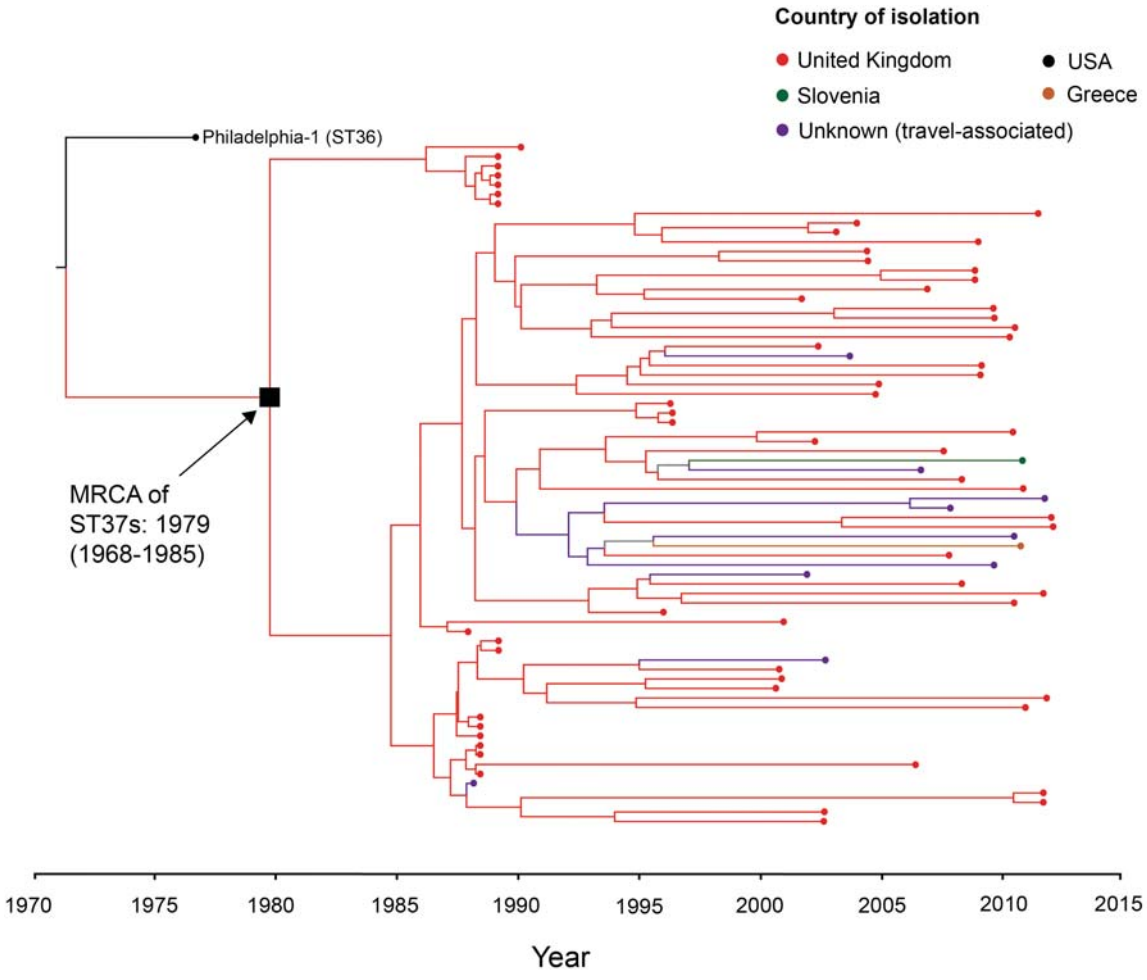


Figure 3.5. Time-dependent phylogenetic reconstruction of the ST37 lineage inferred using a Bayesian coalescent model in BEAST. The Philadelphia-1 isolate (ST36) was included in the analysis as an outgroup. The MRCA of the 72 ST37 isolates is labelled with the median estimated date and the 95% HPD intervals.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

The oldest known isolate of ST37 is from 1982 and within the time interval predicted by BEAST, which makes the dating estimations seem plausible. An evolutionary rate of 2.07×10^{-7} SNPs/site/year (95% HPD interval: 1.69×10^{-7}-2.44×10^{-7}) was estimated, which is slightly higher than the rate predicted for the *L. pneumophila* ST578 lineage (1.39×10^{-7}) (Sanchez-Buso *et al.*, 2014).

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

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

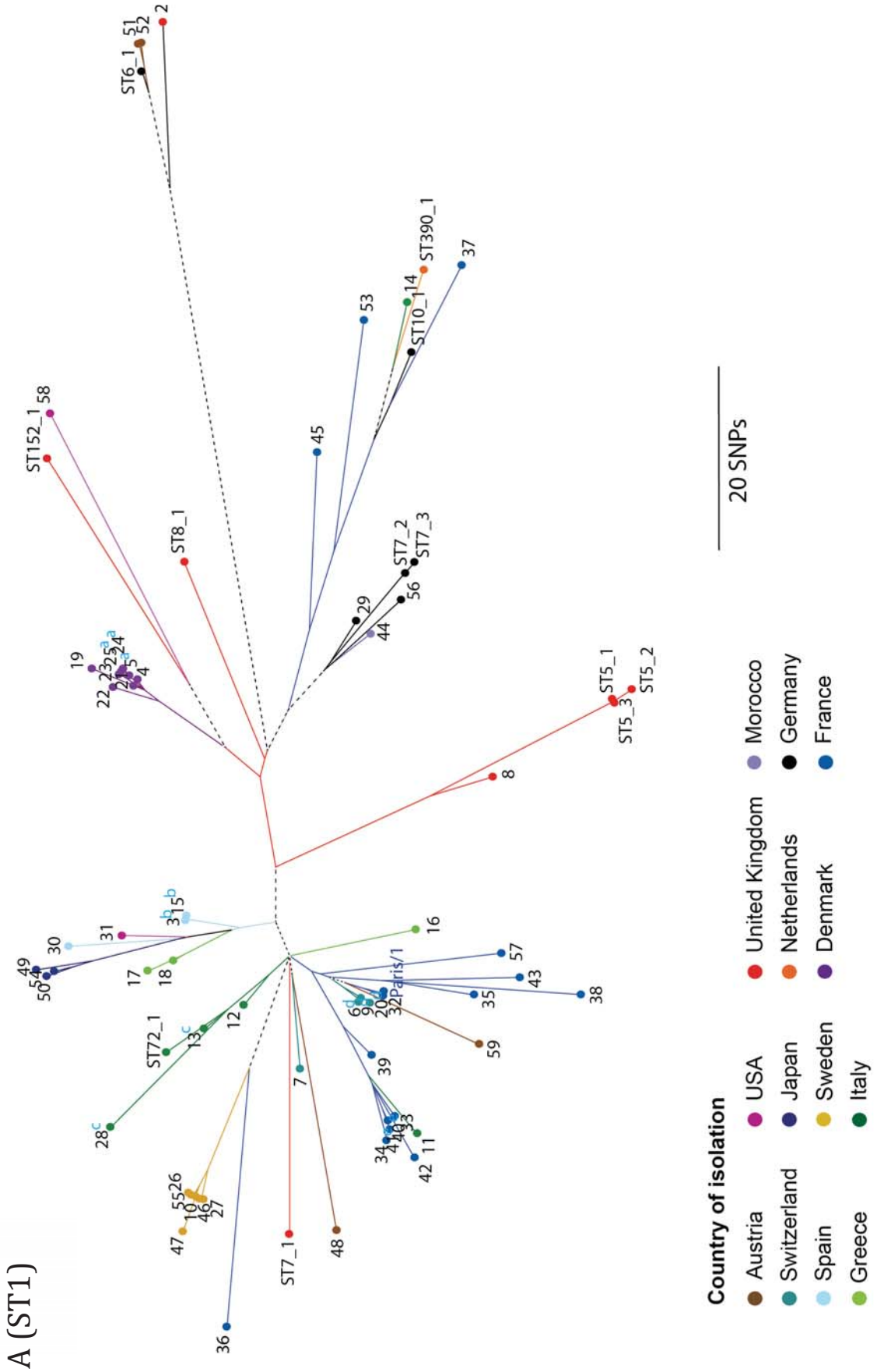
3.3.4 Analysis of the spread of the disease-associated STs

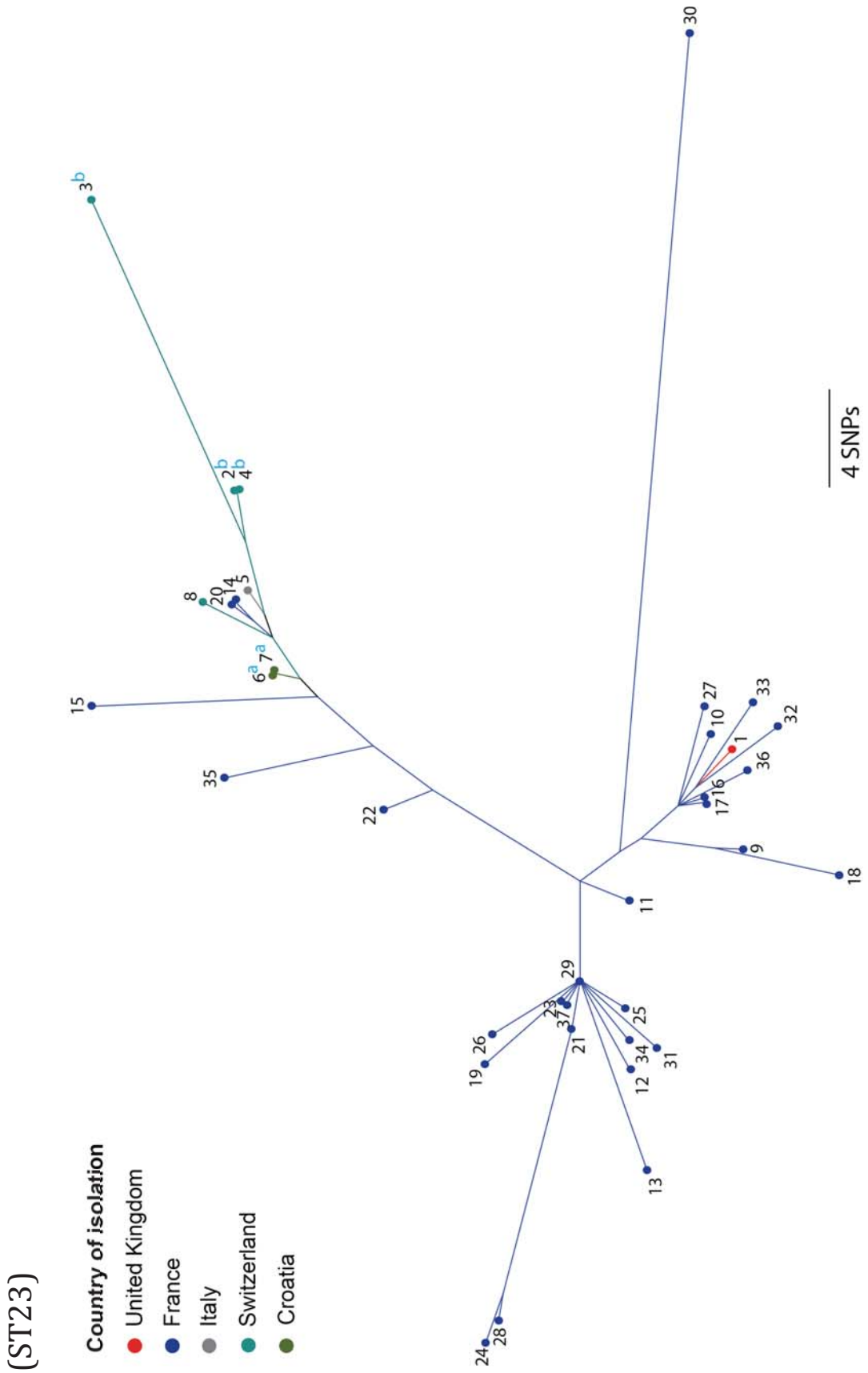
The phylogeographic structure of each of the five disease-associated STs was next analysed using phylogenetic trees constructed using only vertically inherited SNPs. Interestingly, a maximum likelihood tree of the globally dispersed ST1 lineage shows that isolates recovered in the same country do not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

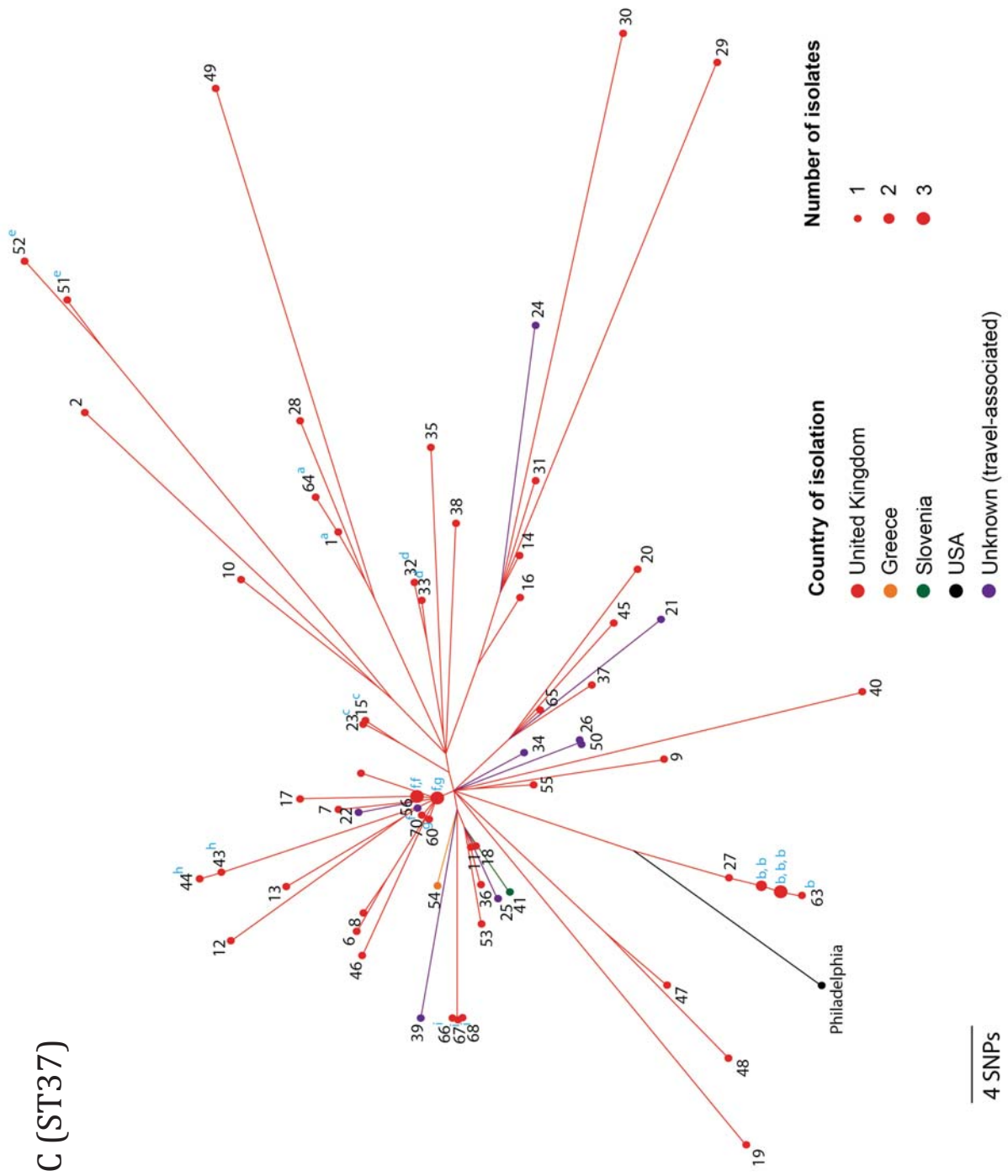
suggests that ST1 isolates have been spread multiple times between different countries and even across continents.

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

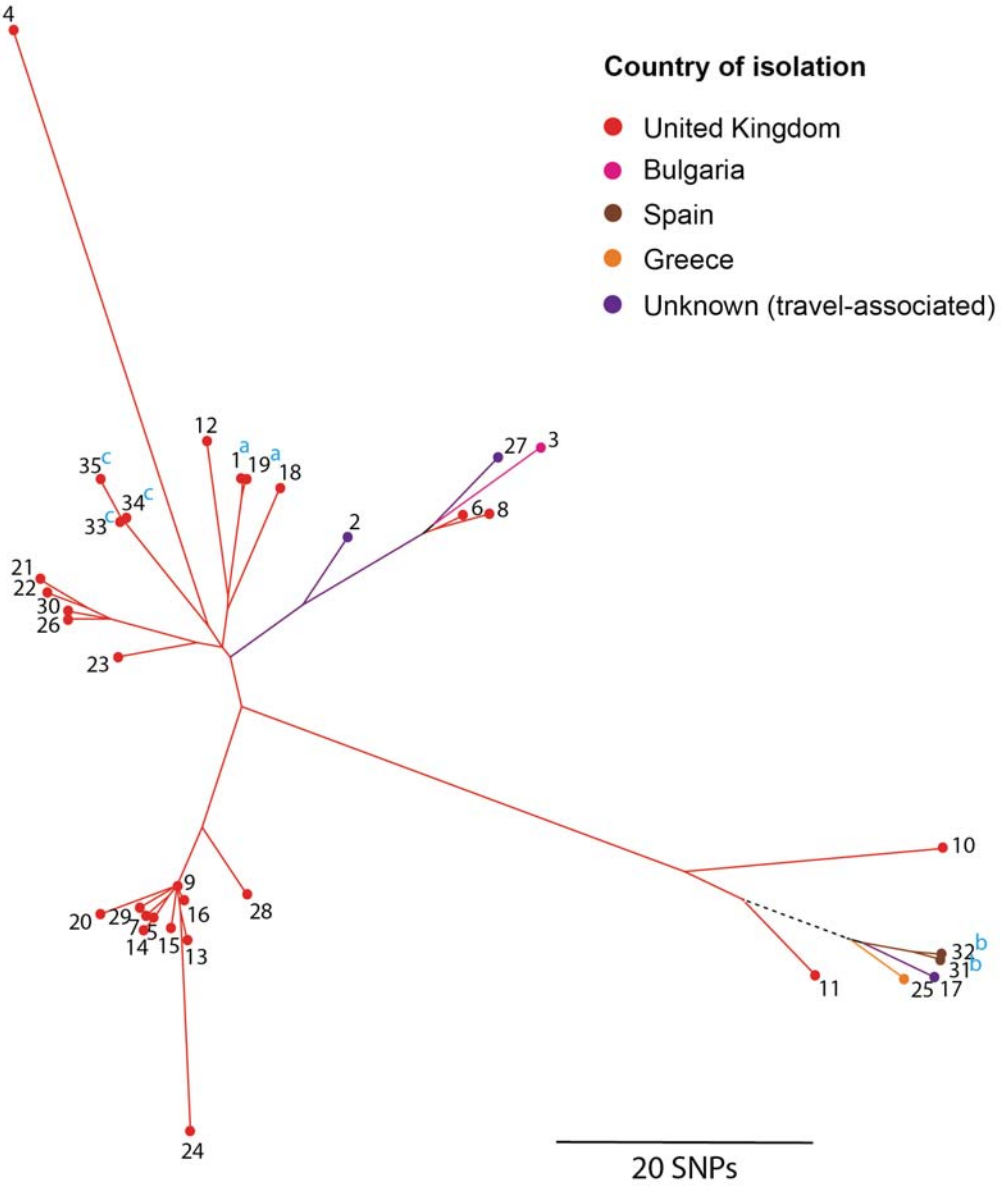
The phylogenetic tree of 122 ST47 isolates shows that isolates mostly cluster by the country of origin (UK or France), although the two clusters are separated by only two SNPs and with low bootstrap support due to the low number of SNPs involved (**Figure 3.6E**). However, there are isolates recovered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


A (ST1)
20 SNPs
Country of isolation
Austria
USA
United Kingdom
Morocco
Switzerland
Japan
Netherlands
Germany
Spain
Sweden
Denmark
France
Greece
Italy
ST6_1
ST152_1
ST8_1
ST390_1
ST10_1
ST7_2
ST7_3
ST5_1
ST5_3
ST5_2
ST72_1
ST7_1
Paris/1


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


C (ST37)
Number of isolates
1
2
3
Country of isolation
United Kingdom
Greece
Slovenia
USA
Unknown (travel-associated)
Philadelphia
4 SNPs

D (ST62)


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

E (ST47)

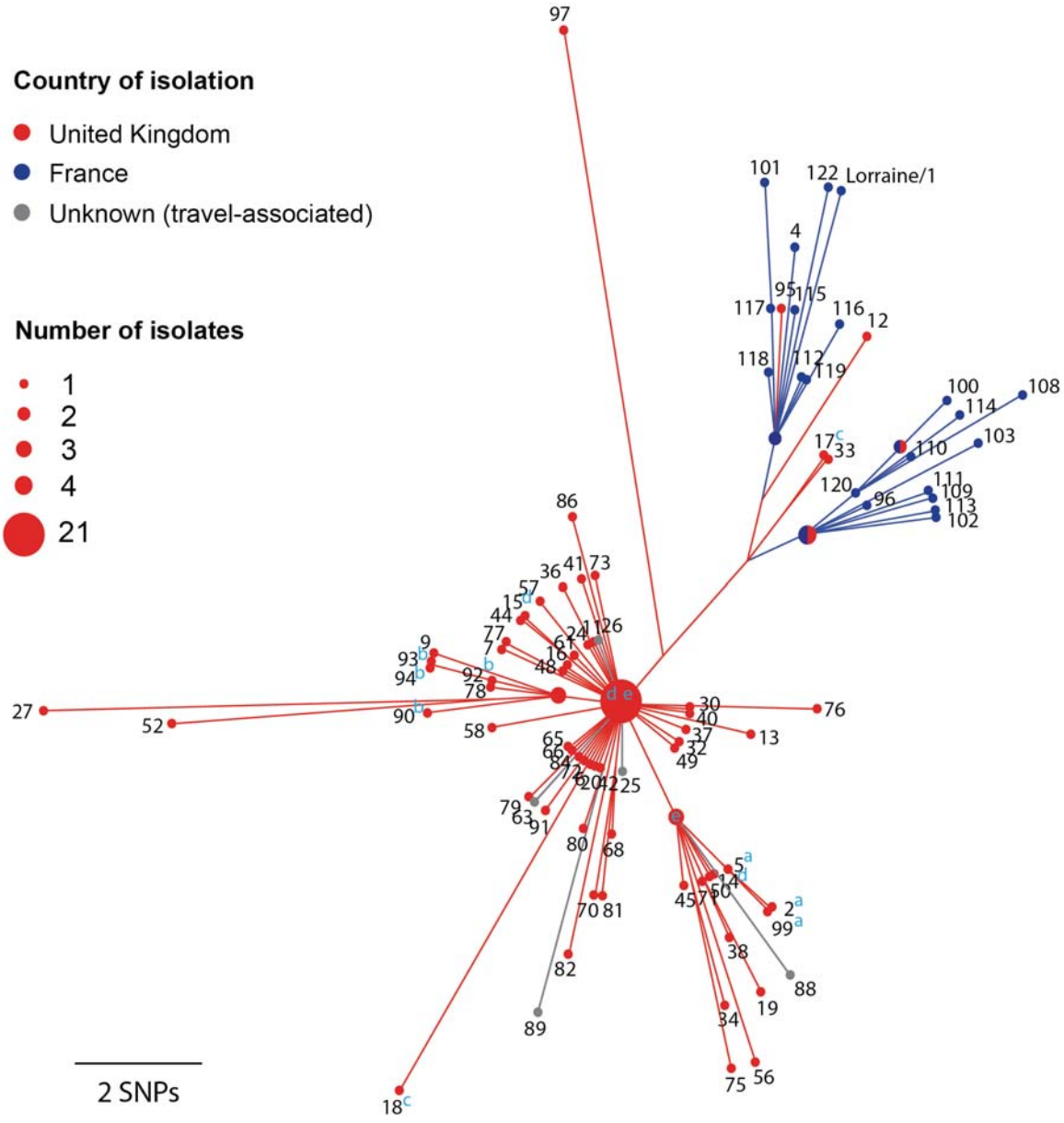


Figure 3.6. Maximum likelihood trees of the ST1 (A), ST23 (B), ST37 (C), ST62 (D) and ST47 (E) lineages. The branch lengths are scaled by the number of SNPs. Isolates are coloured according to the country in which they were recovered and branches are similarly coloured to indicate the origin of descendant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

3.3.5 Evidence of convergent evolution

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

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

Next, homoplasic SNPs that have occurred independently on the branches of the species tree that lead to STs 1, 23, 37, 47 and 62 were searched for. No SNPs were found to have

occurred independently on all five branches although seven occurred on four of the branches, one of which is a non-synonymous change (**Table 3.3**). This SNP is in *LPC_2413* (Corby)/*lpp0942* (Paris), which encodes a diguanylate kinase with a GGDEF domain. This gene is reported to be strongly induced during the transmissive phase of infection (Bruggemann *et al.*, 2006; Weissenmayer *et al.*, 2011). A further 38 SNPs also occurred on three of the five branches, 12 of which are non-synonymous changes (**Table 3.3**). Future studies will be required to determine whether any of these SNPs affect the propensity of *L. pneumophila* to cause disease.

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

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

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

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

A final approach used to search for evidence of convergent evolution between the five disease-associated STs was to identify genes with a higher than expected nucleotide similarity between the five STs compared with the rest of the species representatives. This is a potentially more powerful approach that takes into account all evolution that has occurred during the formation of the five STs rather than relying on signals of selection on the individual, sometimes short, branches leading to each of the lineages. First, a total of 1888 genes that are present in all 32 species representatives were identified excluding three isolates belonging to the *L. pneumophila fraseri* subspecies (ST154, ST336 and ST707), which were omitted from this analysis. For each of the 1888 genes, an alignment was created using one representative isolate from each of the five disease-associated STs, and excluding all other species representatives. The nucleotide

diversity, a value first described by Nei and Li (1979), was calculated for each of the alignments containing the five isolates. Interestingly, many genes were found to possess very low nucleotide diversity values (meaning they are highly similar) and some genes are indeed identical between the five representative isolates (i.e. the nucleotide diversity is 0) (**Figure 3.7**). Most of these localise to a large region about a quarter of the way along the genome.

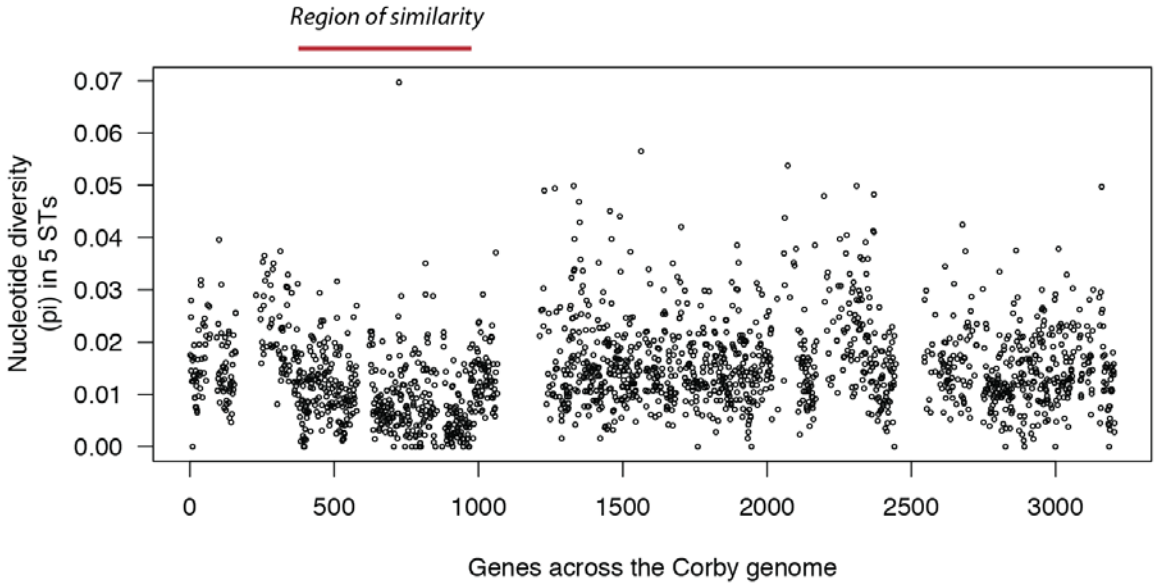


Figure 3.7. Nucleotide diversity of the five STs across the genome. Nucleotide diversity values were calculated for each of the 1888 core genes (i.e. those present in all isolates excluding ST154, ST336 and ST707) using an alignment containing a single representative isolate from each of the five disease-associated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

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

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

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

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

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

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

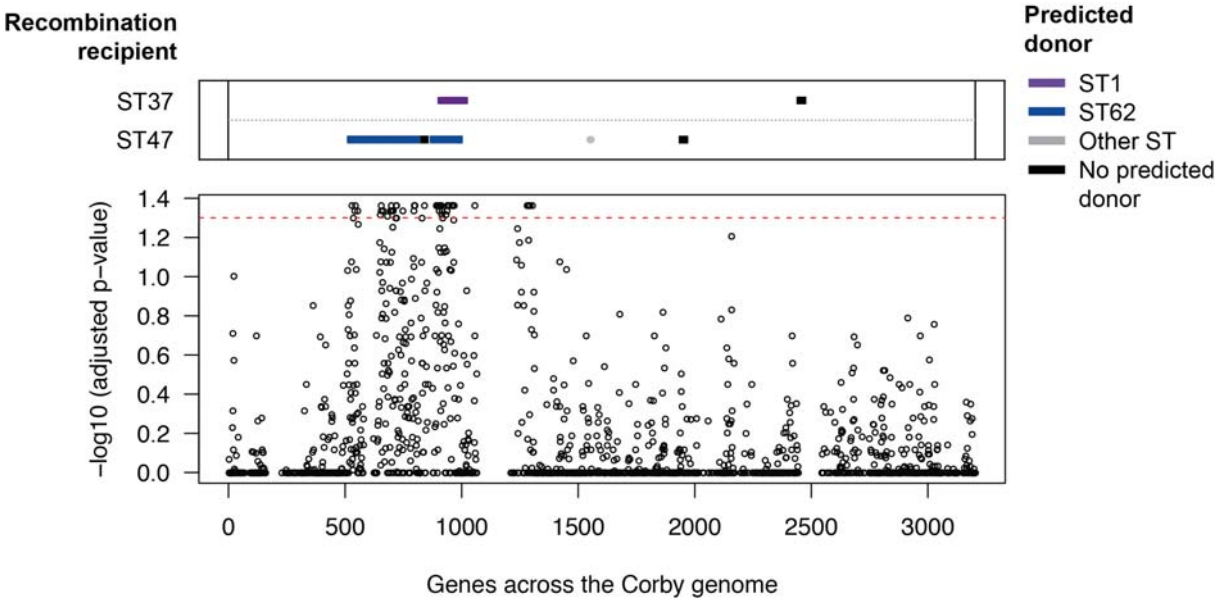


Figure 3.8. Similarity of genes across the five STs and recombination events that have occurred on the branches leading to STs 37 and 47. The bottom plot shows log-transformed p-values derived from testing whether representative isolates from STs 1, 23, 37, 47 and 62 have lower than expected nucleotide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

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

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

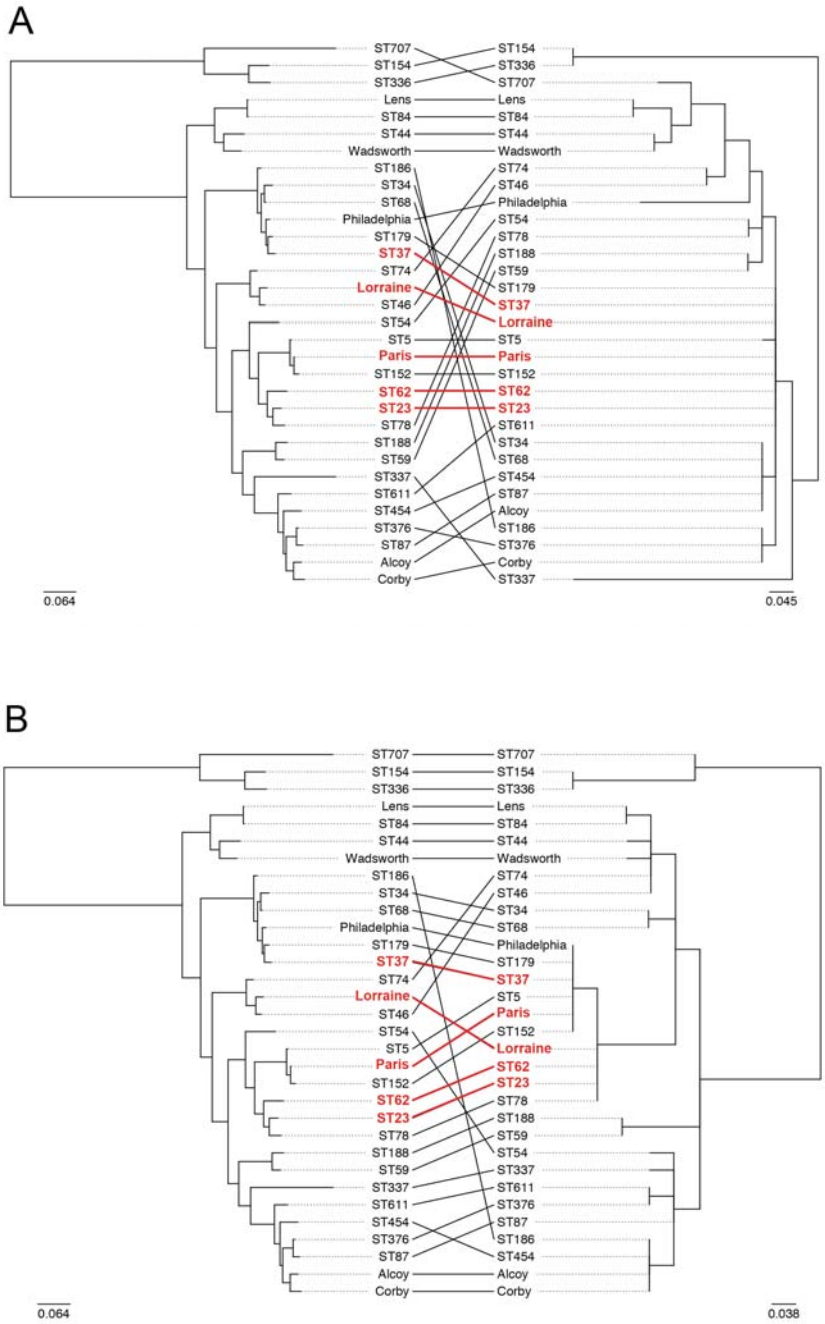


Figure 3.9. Tanglegrams comprising maximum likelihood trees of 32 STs of *L. pneumophila* that are representative of the known species diversity (previous page). The tree on the left hand side of each tanglegram was constructed using the whole genome alignment, generated by mapping sequence reads to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

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

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

between the major disease-associated STs, resulting in a common pool of allelic variants that may be affecting their propensity to cause human disease.

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

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

3.4 Discussion

While more than 2000 STs of *L. pneumophila* have now been reported, analysis of the SBT database demonstrated that just five STs (1, 23, 37, 47 and 62) accounted for over 40% of European isolates submitted prior to April 2015. Four of these STs (23, 37, 47 and 62) are also found very rarely in environmental sources, suggesting that their predominance in human infections may be even more pronounced. This thesis chapter aimed to understand the emergence and diversity of these five STs within the context of

the species and explore whether they possess common genomic features that could be related to their increased propensity to cause disease. A total of 364 *L. pneumophila* genomes were studied here, including 329 that were newly sequenced, which constituted the largest genome collection of this species sequenced and analysed to date.

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

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

Time-dependent phylogenetic analysis of the ST37 lineage predicted the substitution rate to be 2.07×10^{-7} SNPs/site/year (0.71 SNPs/genome/year), which is slightly higher than the previously estimated rate for the ST578 lineage of 1.39×10^{-7} SNPs/site/year (0.49 SNPs/genome/year) (Sanchez-Buso *et al.*, 2014). Both of these estimates are relatively low in comparison with many bacteria but similar to that of *Mycobacterium tuberculosis*, a notoriously slow-evolving pathogen (Ford *et al.*, 2013). Further support for a low substitution rate in *L. pneumophila* comes from the fact that only 20 vertically inherited SNPs were detected between the OLDA1 isolate, recovered in 1947, and another ST1 isolate, recovered in 1995. Similarly, no SNPs were identified amongst 21 ST47 isolates recovered between 2003 and 2012. Furthermore, linear regression analysis of the root-to-tip distances against sampling time in each of the five disease-associated lineages showed an extremely poor correlation in all lineages except ST37. This observation, together with the low estimates for the evolutionary rate, suggests that *L. pneumophila* may undergo periods of dormancy in which no replication occurs.

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

including those that cause a large proportion of human disease as well as those that are rarely implicated in disease. It could be extremely interesting if major disease-associated STs were transmitted more frequently across long distances (e.g. across countries and continents) than STs that are rarely implicated in disease. While this could suggest that disease-associated STs have adapted to a new environmental niche that facilitates long-distance spread, another possibility also exists.

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

Transmission of *L. pneumophila* from humans back into the environment is possible since *L. pneumophila* is regularly isolated from sputum samples of legionellosis patients, and has also been isolated from human feces (Rowbotham, 1998). Thus, contamination of man-made water systems via human respiratory or faecal secretions, or a combination of both, could be a possible mechanism of transmission back to the environment. Interestingly, the first probable case of human-to-human transmission has also recently been reported (Correia *et al.*, 2016). However, this reported case occurred

in particularly unusual circumstances whereby the first patient, who was severely ill, was nursed by his mother for several hours in a small and non-ventilated room. Thus, due to the unusual nature of this case, and the fact that person-to-person has never previously been reported, it can be assumed that direct transmission between humans is an extremely rare event and is unlikely to play a major role in the spread of *L. pneumophila*. Instead, it is more likely that transmission between humans would occur indirectly (via an intermediate environmental source).

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

Finally, a number of methods were used in this chapter to explore whether the five disease-associated STs share genomic features that could explain their increased propensity to cause human infection. Sixty-four genes contained within a large region of ~700kb were identified that contain higher than expected nucleotide similarity between representative isolates from STs 1, 23, 37, 47 and 62. Some of these genes have been previously reported to be involved in intracellular infection and virulence. By searching for recombination events on the branches of the species tree leading to each of the five disease-associated STs, it was shown that genes within this region have been horizontally exchanged between these STs prior to their emergence. It is hypothesised that this shared pool of allelic variants that has arisen *via* recombination may be related to the increased disease propensity of these five STs. Future confirmation could come from genomic analyses of other major disease-associated strains together with a

comparative collection of strains that are never or that are very rarely implicated in human disease.

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

4. Dynamics and impact of homologous recombination on the evolution of *L. pneumophila*

Declaration of work contributions

Julian Parkhill, Simon Harris and Timothy Harrison supervised this work. Massimo Mentasti performed culture and DNA extraction of all newly sequenced isolates. Leonor Sánchez-Busó contributed to the detection of MGEs and the inference of recombination donors. Jukka Corander performed the Bayesian analysis of population structure (BAPS) clustering. I conducted the remaining bioinformatics analyses and generated all the figures.

Publication

The following work has been prepared for publication:

David, S., Sánchez-Busó, L., Harris, S. R., Harrison, T. G. & Parkhill, J. Dynamics and impact of homologous recombination on the evolution of *Legionella pneumophila*.

4.1 Introduction

While all bacteria reproduce clonally, some also import DNA from other organisms into their chromosomes in a process known as recombination or horizontal gene transfer. There are three known mechanisms through which this can take place including transduction (*via* phage infection), conjugation (*via* direct contact), and transformation (*via* the uptake of naked DNA from the environment). The imported DNA can comprise either novel genes that are new to the recipient genome (non-homologous recombination), or can replace an equivalent segment of the genome (homologous recombination). The latter, which is the main focus of this thesis chapter, results in the replacement of genes with alternative allelic variants and requires the DNA to be highly similar, and possibly identical, at both ends of the fragment (Majewski & Cohan, 1998). For this reason, homologous recombination usually occurs between closely related bacteria.

The importance of recombination in bacterial evolution first became clear through the analysis of MLST data, which showed that phylogenetic trees constructed from individual MLST genes were often incongruent (Feil *et al.*, 2001). These analyses also predicted that the rate of homologous recombination varies considerably between different species (Perez-Losada *et al.*, 2006). There are a number of hypotheses regarding why bacteria engage in homologous recombination (Vos, 2009). One explanation is that recombination is used as a mechanism by which DNA damage, such as double-strand breaks, can be repaired using foreign DNA as a template (Michod *et al.*, 2008). Another is that it is a side effect of DNA uptake for use as an energy source or for DNA synthesis from nucleotide precursors (Redfield, 1993). Finally, the ability of recombination events to remove deleterious mutations and rapidly introduce combinations of advantageous mutations could mean it increases the efficiency of natural selection and is selectively maintained (Narra & Ochman, 2006).

In recent years, the availability of WGS data from multiple closely related isolates has enabled homologous recombination to be studied in great detail in species such as *Streptococcus pneumoniae* (Croucher *et al.*, 2011; Chewapreecha *et al.*, 2014), *Chlamydia trachomatis* (Harris *et al.*, 2012) and *Neisseria meningitidis* (Kong *et al.*, 2013). These

studies have confirmed that homologous recombination plays an important role in the evolution and adaptation of important bacterial pathogens, for example by facilitating vaccine escape (Croucher *et al.*, 2011) and antibiotic resistance (Chewapreecha *et al.*, 2014) in *S. pneumoniae*.

L. pneumophila was first reported to have a clonal population structure based on mult-locus enzyme electrophoresis (MLEE) analysis (Selander *et al.*, 1985). However, the three primary mechanisms of bacterial recombination (conjugation, transduction and transformation) have since all been described in *L. pneumophila* (Dreyfus & Iglewski, 1985; Mintz & Shuman, 1987; Stone & Abu Kwaik, 1999), and thus it was unsurprising when later studies reported its occurrence. Indeed, an early genomic study of the first sequenced genomes of *L. pneumophila* showed that recombination events are frequent and predicted that it can involve large chromosomal fragments of over 200kb (Gomez-Valero *et al.*, 2011). More recently, a study of closely related genomes belonging to ST578 demonstrated that recombination accounts for over 98% of the SNPs detected within the lineage and is therefore a dominant force in *L. pneumophila* evolution (Sanchez-Buso *et al.*, 2014). These findings are concordant with those made in the first results chapter of this thesis whereby over 96% of the SNPs found in STs 1, 23, 37 and 62 were found to be imported *via* recombination. Interestingly though, no recombination was detected in the ST47 lineage. In both the published study by Sanchez-Buso *et al.* (2014) and *Chapter 3*, the relative contribution of homologous and non-homologous recombination is not disentangled, nor is the impact of recombination on the adaptation and evolution of *L. pneumophila* studied in detail.

Therefore, the first aim of this results chapter is quantify the relative impact of homologous and non-homologous recombination on the evolution of major disease-associated lineages of *L. pneumophila* including STs 1, 23, 37, 62 (as studied in *Chapter 3*), ST578 (as studied by Sanchez-Buso *et al*., 2014) and an additional disease-associated lineage comprising ST42 isolates. The chapter subsequently focuses solely on homologous recombination, and seeks to characterise the regions that have been imported *via* this process. It explores whether there are "hotspots" in the genome where homologous recombination events are more likely to be selectively maintained, which could provide insight into important selection pressures of *L. pneumophila*. Finally, by

inferring the donor lineages of recombined regions, the last aim of the chapter is to examine the extent to which homologous recombination occurs between the two *L. pneumophila* subspecies and between and within major clades of the *L. pneumophila pneumophila* subspecies. This will provide further insight into the dynamics of genomic flux within the *L. pneumophila* species and reveal the extent to which major disease-associated STs from different clades are able to exchange DNA. As was suggested in *Chapter 3*, this process could represent an important mechanism by which diverse strains can adapt rapidly to new niches.

4.2 Materials & Methods

4.2.1 Bacterial isolates

L. pneumophila isolates belonging to six major disease-associated lineages are primarily used in this study (*n*=290). These include 81 ST1, or ST1-derived, isolates (including 71 used in *Chapter 3* and 10 from a study by Sanchez-Buso *et al.* (2014)), 42 ST23 isolates (including 37 used in *Chapter 3* and 5 from another study by Sanchez-Buso *et al.* (2016)), 72 ST37 and 35 ST62 isolates (all of which were used in *Chapter 3*), 46 ST578 (including one published by D'Auria *et al.* (2010) and 45 published by Sanchez-Buso *et al.* (2014)) and 15 ST42 isolates (including 2 previously published isolates (Schroeder *et al*., 2010; Underwood *et al.*, 2013) and 13 newly sequenced isolates). Those additional isolates belonging to these six STs that are not used in *Chapter 3* are listed in **Appendix Table 3**. A further 246 *L. pneumophila* isolates, listed in **Appendix Table 4** and which belong to a range of STs, were also used in the inference of recombination donors. Importantly, these include a set of previously published genomes, which were selected for sequencing using MLST data with the aim of encompassing as much of the species diversity as possible (Underwood *et al.*, 2013). Culture, DNA extraction and sequencing of all newly sequenced isolates were performed as described in *Chapter 2 (Materials & Methods).* Accession numbers or references for all sequence data are provided in **Appendix Tables 3 and 4**.

4.2.2 Reference genomes

Isolates belonging to each of the six disease-associated STs (1, 23, 37, 42, 62 and 578) were mapped to a reference genome of the same ST to enable each lineage to be studied at a high resolution. The complete genomes of Paris (Cazalet *et al.*, 2004) and Alcoy (D'Auria *et al.*, 2010) were already available for ST1 and ST578, respectively. Reference genomes were generated for the remaining four STs by sequencing a representative isolate from each ST on the PacBio RSII sequencer at the WTSI. The isolates chosen were EUL 28, EUL 120, EUL 165 and H044120014 belonging to STs 23, 42, 37 and 62, respectively. 1-2μg of DNA from each isolate was sheared using a 26G blunt-ended needle (ThermoFisher, UK) and used in library preparation according to the manufacturer's protocol. The P4 DNA polymerase was used with C2 chemistry to perform the sequencing. *De novo* assemblies were produced from the sequence reads using HGAP.3 (Pacific Biosciences). Assemblies that consisted of a single chromosomal contig were circularised using the overlapping sequence at the two ends and the start of the genome was set to the beginning of the *dnaA* gene. Each genome was subsequently confirmed by mapping Illumina sequence reads from each of the isolates to the PacBio assembly. Sequencing statistics for the four PacBio reference genomes are provided in **Appendix Table 5**. They were annotated using an in-house pipeline at the WTSI, which uses Prokka (Seemann, 2014), together with the complete genomes of Paris (ST1) and Alcoy (ST578).

Repetitive regions over 100bp were detected in the six reference genomes using repeat-match from MUMmer v3.0 (Kurtz *et al.*, 2004) (**Appendix Table 6**). This was performed by Leonor Sánchez-Busó.

4.2.3 Mapping, recombination detection, phylogenetic analysis and BAPS clustering

Sequence reads from all isolates belonging to the six major disease-associated STs were mapped to the appropriate reference genome of the same ST using SMALT v0.7.4 (available at: http://www.sanger.ac.uk/science/tools/smalt-0). All isolates used in the study (n=536) were also mapped to the Paris (ST1) reference genome (Cazalet *et al.*, 2004) in order to study the species-wide phylogenetic structure. An in-house pipeline at

the WTSI was used to call bases and identify SNPs as described in *Chapter 2 (Materials & Methods)*.

Recombined regions were detected in the alignments of the six disease-associated STs using Gubbins (Croucher *et al.*, 2015). Phylogenetic trees of these lineages were generated as described in *Chapter 2 (Materials & Methods)*, firstly using all SNPs to later allow ancestral sequence reconstruction (see *4.2.6 Inference of recombination donors*), and secondly using only the vertically inherited SNPs. A phylogenetic tree of the total 536 isolates was constructed using all the detected SNPs, as the high diversity renders recombination detection impossible. The alignment of all 536 genomes against the Paris reference genome was also used to group the isolates into clusters using hierBAPS (Cheng *et al.*, 2013), which was performed by Jukka Corander.

4.2.4 Detection of MGEs

The annotation files for each of the six reference genomes were parsed to detect genes annotated as "integrase", "transposase", "recombinase", "phage", "lvrA", "csrA", "HTX", "helix-turn-helix", "xre", "conjugal", "conjugation", "tra", "trb", "vir" and "mobile". Both the published annotation files of the Paris (ST1) and Alcoy (ST578) complete genomes and those generated using the in-house pipeline at the WTSI were used. However, the new annotations were only considered when the original one was a "hypothetical protein" in order to respect experimentally proven annotations. Plots showing the mapping coverage of all isolates in the six STs against the corresponding reference genome were also evaluated. Regions over 8kb with no coverage and that did not match repetitive regions were considered as potential mobile regions. These analyses were performed by Leonor Sánchez-Busó.

Other software to detect MGEs was also used including AlienHunter (Vernikos & Parkhill, 2006) and Island Viewer, the latter of which incorporates IslandPick, IslandPath-DIMOB and SIGI-HMM (Langille & Brinkman, 2009). However, these results were discarded due to major incongruences between them. Finally, manual curation of all predicted MGEs was performed using Artemis v15.0.0 (Carver *et al.*, 2012) (**Appendix Table 6**).

4.2.5 Identification of homologous recombination hotspots

In each of the six lineages, any predicted recombined regions that overlap with either repetitive regions or MGEs in the reference genome were identified and discarded for the majority of the analysis in this study, leaving only putative homologous recombination regions. An in-house script was used to calculate the number of times each gene had been involved in a homologous recombination event. Recombination "hotspots" were defined as genes with a recombination frequency above the 95th percentile observed in that particular ST.

4.2.6 Inference of recombination donors

A custom genome BLAST database (BLAST v2.2.30+) (Camacho *et al.*, 2009) was constructed using *de novo* assemblies from all 536 *L. pneumophila* isolates used in this study. The method by which *de novo* assemblies were generated is described in *Chapter 2 (Materials & Methods)*. Homologous recombination regions were extracted from the ancestral sequences inferred from the nodes of the phylogenetic trees (constructed prior to recombination removal) using PAML 4 (Yang, 2007). The reconstructed recombined regions were used as query sequences in BLAST searches against the custom genome database and the National Center for Biotechnology Information (NCBI) non-redundant nucleotide database. The resulting hits were filtered to remove those against isolates that are descended from the branch in which the recombination event was detected. Of the remaining hits, the one with the highest bit score was considered as the potential donor, provided it had a minimum of 99% nucleotide identity to the recombined fragment, and matched at least 50% of the fragment length.

4.3 Results

4.3.1 Contribution of homologous recombination to *L. pneumophila* diversity

To investigate the relative contribution of homologous recombination to diversity in each of the six major disease-associated STs (1, 23, 37, 42, 62 and 578), isolates were

first mapped to a reference genome of the same ST. Gubbins was used to detect recombined regions in each of the six alignments and construct a phylogenetic tree based on the vertically inherited SNPs located outside of these regions. As found in *Chapter 3* and a study by Sanchez-Buso *et al.* (2014), over 96% SNPs in STs 1, 23, 37, 62 and 578 are predicted to be derived from recombination events (**Table 4.1**). Furthermore, 99.0% SNPs in the ST42 lineage, which has not been studied previously, were also found in predicted recombined regions. The remaining number of vertically inherited SNPs in each of these lineages ranges from just 94 (ST42) to 1006 (ST1) (**Table 4.1**).

Table 4.1. Number of SNPs detected within each of the six disease-associated STs.

ST	Number of isolates	Total number of SNPs	Number (and %) of SNPs in recombined regions	Number (and %) of vertically inherited SNPs
1	81	73,044	72,038 (98.6%)	1006 (1.4%)
23	42	44,886	44,720 (99.6%)	166 (0.4%)
37	72	17,776	17,300 (97.3%)	476 (2.7%)
42	15	9,256	9,162 (99.0%)	94 (1.0%)
62	35	47,684	47,372 (99.3%)	312 (0.7%)
578	46	3,678	3,559 (96.8%)	119 (3.2%)

Any recombined regions that overlapped with either predicted MGE regions or repeat regions were subsequently identified, in order to determine the contribution of only homologous recombination to *L. pneumophila* diversity. It was found that between 33.0% (ST62) and 80.0% (ST578) of all SNPs are predicted to be in regions derived from homologous recombination events (**Table 4.2 and Figure 4.1**). However, the mean length of each individual genome affected by this process varies between just 1.2% (ST42/578) and 3.9% (ST1) (**Table 4.2**). It should be noted that the number of SNPs from homologous recombination might be slightly overestimated (and the number of *de novo* mutations slightly under estimated) since *de novo* mutations may have occurred on top of recombination events. However, the error should be no more than 1.2-3.9%, in

proportion with the average length of genome affected by homologous recombination events.

Table 4.2. Contribution of homologous recombination to the diversity of the six major disease-associated STs.

ST	Number of homologous recombination events	Number of SNPs in homologous recombination regions per vertically inherited SNP	Number of homologous recombination events per vertically inherited SNP (r/m ratio)	Mean length of sequence (and %) of each individual genome affected by homologous recombination (bp)	Total length (and %) of the reference genome affected by homologous recombination across all isolates (bp)
1	198	56.2	0.20	135,208 (3.9%)	1,430,288 (40.8%)
23	44	93.8	0.27	51,242 (1.5%)	520,584 (14.8%)
37	13	20.8	0.03	105,051 (3.0%)	251,988 (7.3%)
42	11	41.3	0.12	41,747 (1.2%)	120,545 (3.51%)
62	48	50.5	0.15	66,559 (1.9%)	456,451 (12.9%)
578	23	24.6	0.19	42,138 (1.2%)	204,114 (5.8%)

In each of the six lineages, the relative number of homologous recombination events to vertically inherited mutations (r/m ratio) was calculated per branch of each phylogenetic tree (**Figure 4.2**). Across all branches, the r/m ratio ranged from 0.03 (ST37) to 0.27 (ST23), indicating that recombination events have occurred less frequently than vertically inherited mutations in all six lineages, despite bringing in between 20.8 (ST37) and 93.8 (ST23) times as many SNPs (**Table 4.2**). The r/m ratios also differ significantly between lineages (Kruskal-Wallis test, $p<0.05$), highlighting different rates of recombination in the six major disease-associated STs.

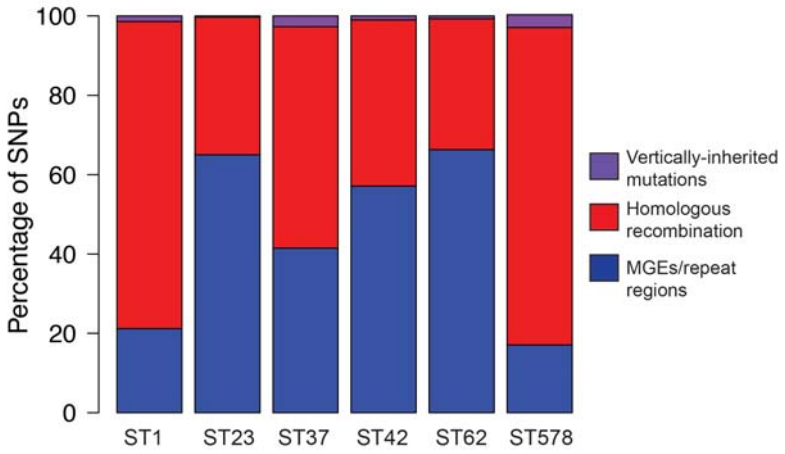


Figure 4.1 Generation of diversity in the six major disease-associated STs. The percentage of SNPs that are predicted to be derived from vertically inherited mutations, homologous recombination or from MGEs (i.e. non-homologous recombination) and repeat regions.

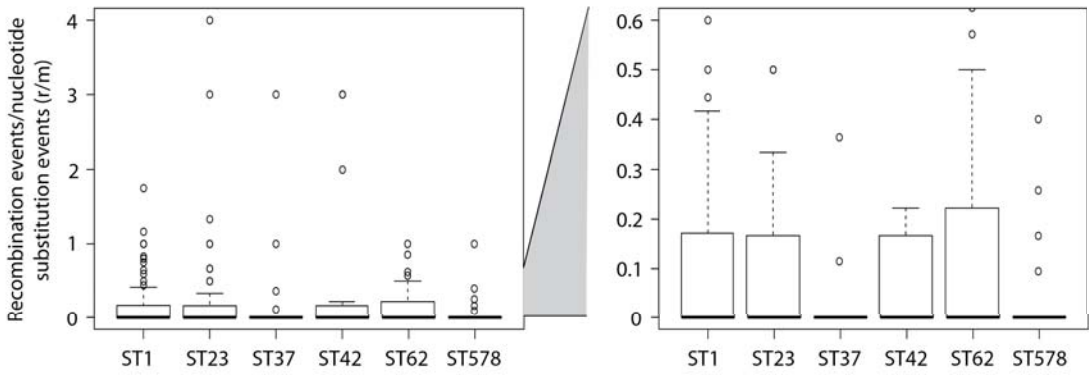


Figure 4.2. Relative frequency of homologous recombination events and vertically inherited mutations. Boxplots showing the number of homologous recombination events detected per vertically inherited SNP (r/m) on each of the branches of the phylogenetic trees belonging to the six STs.

To determine the relative impact of vertically inherited mutations and homologous recombination events on the coding sequence, the types of changes caused by the two processes were analysed. Vertically inherited mutations resulted in approximately

twice as many non-synonymous SNPs than synonymous SNPs, a result that is expected by chance when mutations occur at random in the genome and before selection has time to act on all but the most deleterious mutations (**Figure 4.3**). Interestingly though, the results are reversed for homologous recombination events, which result mostly in synonymous mutations (**Figure 4.3**). However, this observation is also not unexpected given that variants in sequences that are horizontally transferred between different lineages will have been subjected to a longer period of evolution and selection, which has purged harmful, non-synonymous mutations. The same phenomenon has also been observed in a previous study by Castillo-Ramirez *et al.* (2011). Furthermore, fewer SNPs that result in a stop codon are brought in by homologous recombination events than by vertically inherited mutations (**Figure 4.3**), which can also be explained by this process.

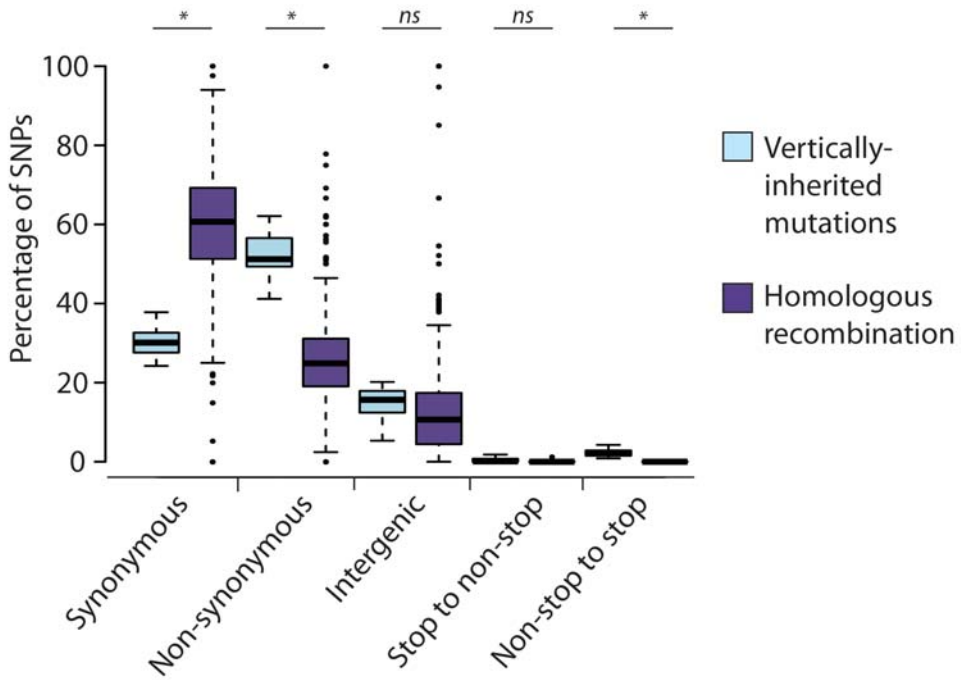


Figure 4.3. Types of change introduced by vertically inherited mutations and homologous recombination. Boxplots showing the percentage of SNPs per branch, derived from either vertically inherited mutations or homologous recombination that are synonymous, non-synonymous, intergenic, or result in a change from a stop to non-stop codon or a non-stop to stop codon. Statistically significant differences as determined by a Student's paired t-test are indicated by an asterisk. ns – not significant

The lengths of the recombined regions are exponentially distributed (rate of decay=7.52 x 10^{-5} bp^{-1}), with the majority of events being small (<10,000bp) and large events occurring relatively infrequently (**Figure 4.4**). The median recombination fragment length in each of the six lineages varies from 5,613bp (ST578) to 12,757bp (ST37), while the largest predicted region is 94,790bp (ST37).

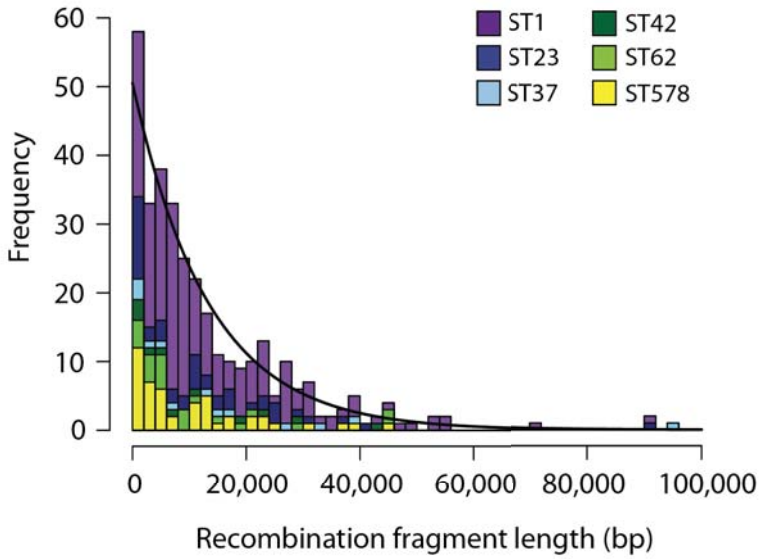


Figure 4.4. Size of detected homologous recombination regions in the six STs. An exponential decay curve (black line) is fitted to the distribution and the rate of decay is 7.52 x 10^{-5} bp^{-1}.

4.3.2 Hotspots of homologous recombination in *L. pneumophila*

To identify genomic regions associated with a high number of homologous recombination events, the number of events that overlap with each gene was calculated with respect to the reference genomes of the six STs. Hotspot regions were defined as genes with equal to or greater than the 95% recombination frequency detected in each lineage, and which were involved in at least two recombination events. This accounted for the recombination frequency, population size and diversity of each lineage. Based on these criteria, the minimum number of recombination events that a gene must have been involved in to be considered within a hotspot region was four events in the ST1

lineage and two events in the remaining five STs. A total of 32 hotspot regions were defined, including at least one in all six STs (**Table 4.3 and Appendix Table 7**). The most notable hotspot regions were observed in the ST1 lineage, which is also predicted to contain the highest number of homologous recombination events. A total of ten hotspot regions were defined in the ST1 lineage and, remarkably, one region contains genes that are predicted to have been involved in up to 27 recombination events and individual bases that have been involved in up to 25 recombination events (**Figure 4.5**). In the other five STs, the highest number of events affecting genes ranges from 2 (ST37/ST578) to 4 (ST42/ST62).

Table 4.3. Recombination hotspots in the six major disease-associated STs.

ST	Number of recombination events affecting genes	Genomic region (with respect to ST-specific reference genome)	Genes (defined in ST-specific reference genome and the Paris genome)
1	4-5	23,666-30,454	*lpp0019-lpp0024*
1	4	405,129-407,048	*lpp0356*
1	4-7	916,526-930,133	*lpp0819-lpp0830*
1	4	1,067,640-1,071,158	*lpp0961-lpp0963*
1	3-4	1,837,986-1,846,399	*lpp1640-lpp1645*
1	4-27	1,981,301-2,028,475	*lpp1761-lpp1794*
1	4	2,529,239-2,532,139	*lpp2198*
1	4	2,894,069-2,902,985	*lpp2543-lpp2550*
1	5-6	2,960,106-2,968,849	*lpp2595-lpp2604*
1	4	3,393,015-3,396,715	*lpp2977-lpp2979*
23	2	451,136-467,120	*ST23_00399-ST23_00417/lpp0453-lpp0471*
23	3	667,828-669,415	*ST23_00625-ST23_00626/lpp0668-lpp0669*
23	2	694,147-696,095	*ST23_00647-ST23_00648/lpp0690-lpp0691*
23	2	779,739-800,490	*ST23_00703-ST23_00713/lpp0748-lpp0758*

23	2-3	1,972,009-1,978,787	*ST23_01779-ST23_01781/lpp1768-lpp1770*
23	2	2,136,974-2,160,755	*ST23_01931-ST23_01947/lpp1925-lpp1942*
23	2	2,202,408-2,203,172	*ST23_01990/lpp1977*
23	2	2,865,124-2,877,303	*ST23_02606-ST23_02617/lpp2517-lpp2528*
23	2	3,365,161-3,367,970	*ST23_03044-ST23_03046/lpp2944-lpp2946*
37	2	1,326,840-1,329,791	*ST37_01205-ST37_01206/lpp1189-lpp1190*
42	2-4	2,830,437-2,845,009	*ST42_02559-ST42_02567/lpp2687-lpp2695*
62	2-4	284,602-299,923	*ST62_00255-ST62_00267/lpp0262-lpp0274*
62	2	310,597-311,994	*ST62_00277/lpp0285*
62	2	329,218-336,516	*ST62_00287-ST62_00292/lpp0305-lpp0310*
62	2	841,221-852,380	*ST62_00754-ST62_00764/lpp0756-lpp0766*
62	3	910,009-911,253	*ST62_00817/lpp0829*
62	2	918,029-919,546	*ST62_00823/lpp0835*
62	2	1,919,344-1,924,240	*ST62_01733-ST62_01736/lpp1667-lpp1670*
578	2	990,286-1,011,913	*lpa_01248-lpa_01273/lpp0880-lpp0902*
578	2	1,021,391-1,021,984	*lpa_01289/lpp0914*
578	2	1,693,186-1,696,080	*lpa_02154/lpp1435*
578	2	3,230,002-3,259,808	*lpa_04035-lpa_04063/lpp2815-lpp2839*

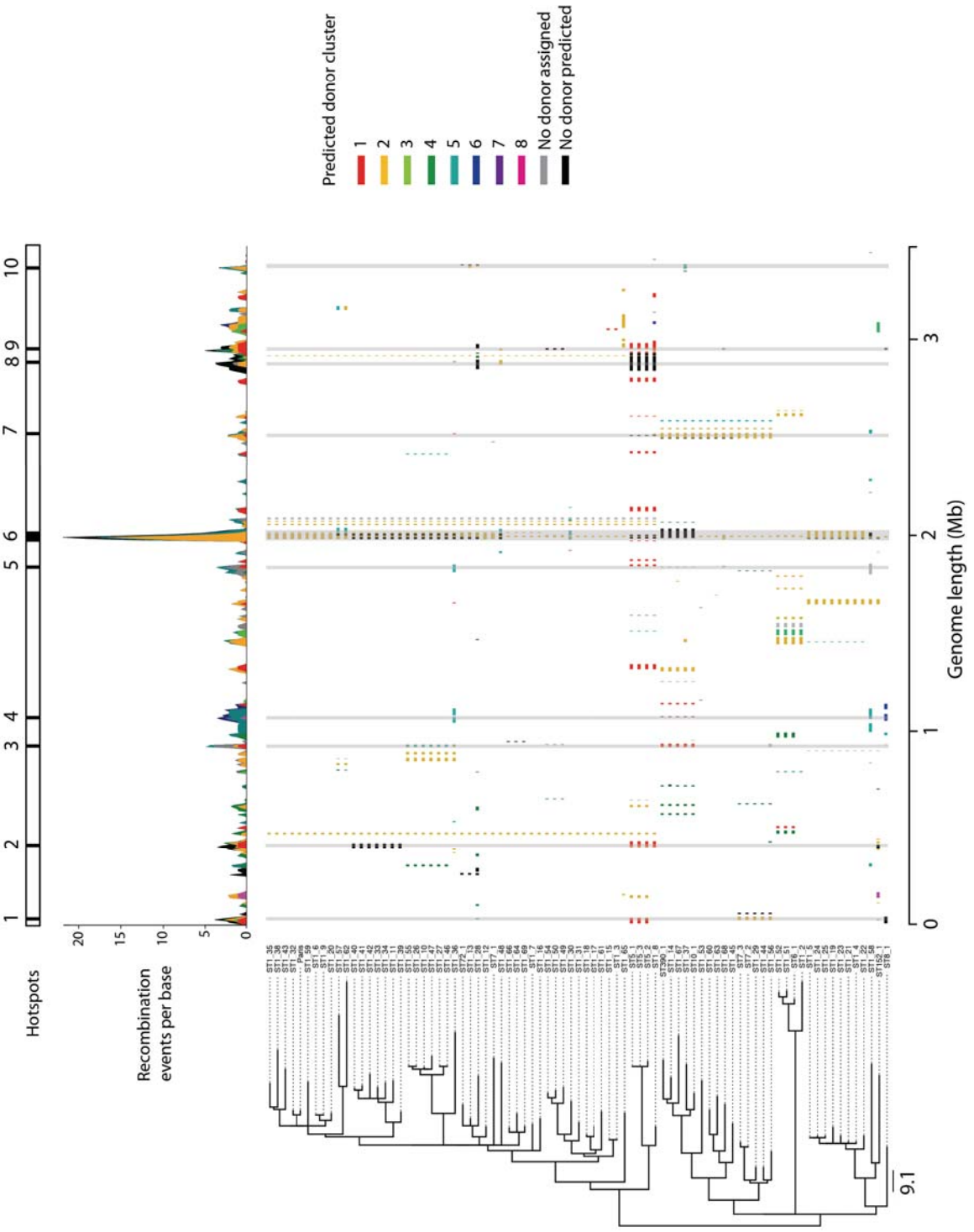

Predicted donor cluster
1
2
3
4
5
6
7
8
No donor assigned
No donor predicted
Hotspots
Recombination events per base
Genome length (Mb)
9.1

Figure 4.5. Homologous recombination events detected in the ST1 lineage (previous page). A phylogenetic tree, constructed using only vertically inherited mutations, is shown on the left and the scale indicates the number of SNPs. Homologous recombination events are shown by blocks, which are coloured according to the donor cluster from which they are predicted to have been derived (see *4.4.3 Inference of recombination donors*). The plot above shows the number of recombination events that have affected each base in the genome using a stacked visualisation to also indicate the number of events derived from different clusters. The ten genomic regions identified as recombination hotspots are marked at the top of the plot.

To further study the recombination hotspots, the gene content of the regions was analysed and compared between lineages. The most prominent hotspot (hotspot 6) identified in the ST1 lineage that contains genes involved in up to 27 recombination events is a 47,174bp region that ranges from *lpp1761* to *lpp1794* in the Paris (ST1) genome (**Figure 4.6 and Appendix Table 7**). The gene in this region that is predicted to have been involved in 27 events is *hemB*/*lpp1771*, a porphobilinogen synthase (delta-aminolevulinic acid dehydratase), which is an enzyme involved in the biosynthesis of tetrapyrroles. Since there is no obvious reason why this metabolic enzyme should be under a high selective pressure, the genes flanking this locus were also investigated. While the two immediate flanking genes (*lpp1770* and *lpp1772*) both encode "hypothetical proteins", *lpp1773*, which has been involved in 25 recombination events, has been shown to encode an outer membrane protein of *L. pneumophila* in a previous study (Khemiri *et al.*, 2008) and has high homology to the *fadL* gene conserved across many bacterial species. Interestingly, a *fadL*-like gene (*ST62_00760*; *lpp0762*) is also found within a recombination hotspot in the ST62 lineage, where it is involved in two recombination events, although it is found in a different part of the genome to the ST1 hotspot region. Furthermore, a smaller 6,778bp hotspot region in the ST23 region (*ST23_01779-ST23_01781; lpp1768-lpp1770*) overlaps with this hotspot region in the ST1 lineage. However, the region in the ST23 lineage centres on the gene, *ST23_01780*/*lpp1769*, which is involved in three recombination events and encodes the outer membrane protein assembly factor, BamA. Interestingly, *lpp1769* is involved in "just" 18 recombination events in the ST1 lineage, compared with *lpp1771* that is involved in 27.

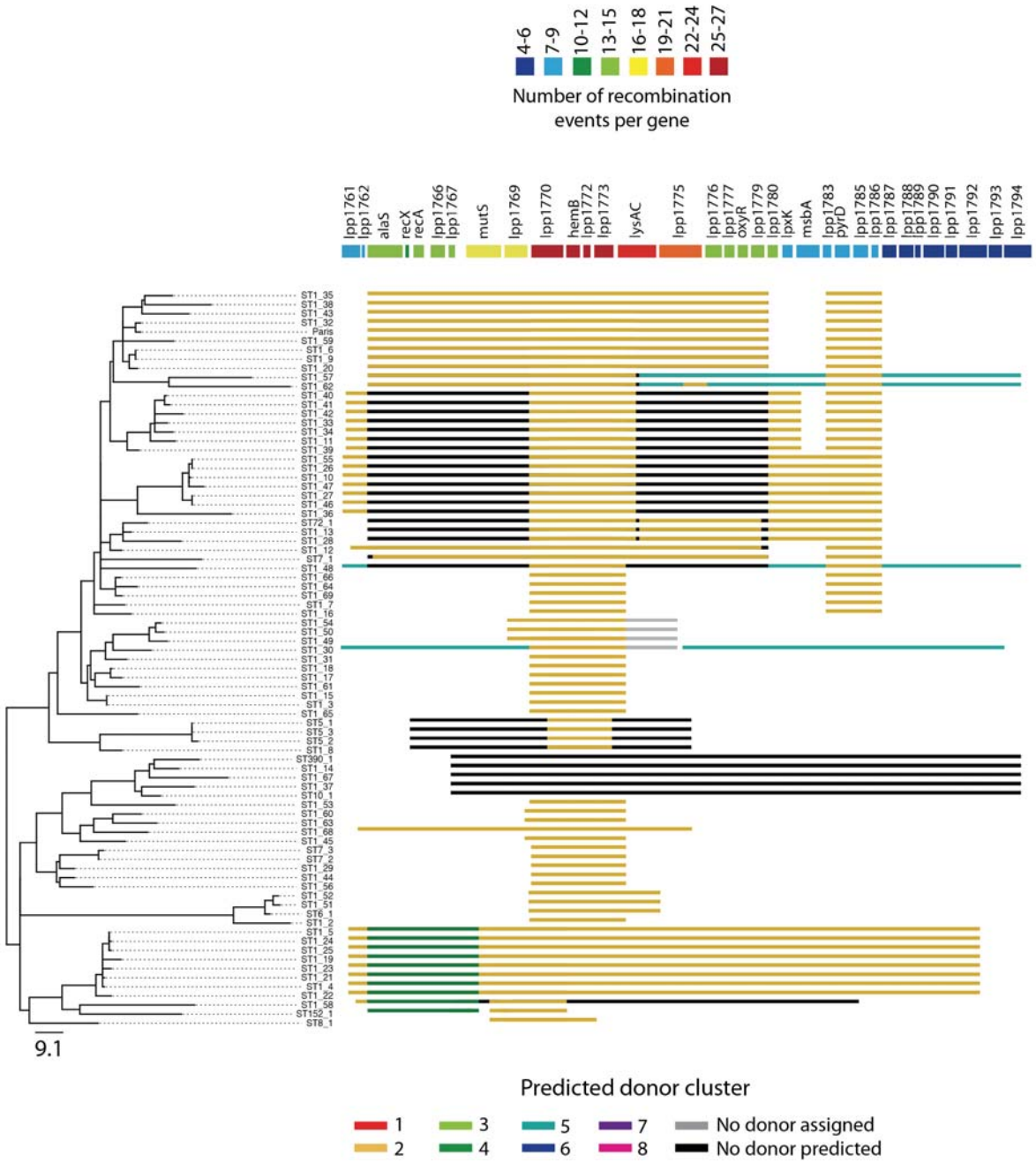


Figure 4.6. Hotspot 6 in the ST1 lineage. A maximum likelihood tree, constructed using only vertically inherited mutations, is shown on the left and the scale indicates the number of SNPs. The homologous recombination events are displayed as blocks, coloured according to the BAPS cluster from which they are predicted to be derived (see *4.4.3 Inference of recombination donors*). The genes shown at the top of the figure are coloured by the number of overlapping recombination regions.

The second most prominent hotspot (hotspot 3) in the ST1 lineage is a 13,607bp region that ranges from *lpp0819* to *lpp0830* in the Paris genome, and which contains genes affected by up to seven recombination events (**Figure 4.7 and Appendix Table 7**). This hotspot is fully contained within the lipopolysaccharide (LPS) locus that spans a region from *lpp0814* to *lpp0843*. Many of the genes in this hotspot region have been implicated in LPS core oligosaccharide biosynthesis including those belonging to the *rml* family, and O-antigen biosynthesis such as *neuA*, *neuB*, *neuC*, *wecA*, *wzt* and *wzm* (Lueneberg *et al.*, 2000). Interestingly, the genes affected by the highest number of recombination events are *wecA* but also *lpp0829a-c*, which are annotated as pseudogenes in the original annotation of the Paris genome (Cazalet *et al.*, 2004). All three genes encode "hypothetical proteins" although *lpp0829a* has a signal peptide and thus may be secreted, while *lpp0829b* has a pectin lyase fold, which has also been found in genes belonging to *L. longbeachae* and is thought to degrade the pectic components of plant cell walls. Furthermore, the ST62 lineage also has two genes from the LPS locus that are in hotspot regions. The first is *ST62_00817*, homologous to the three genes, *lpp0829a-c*, in the Paris genome and which has been involved in three recombination events. The second is *ST62_00823*, homologous to *lpp0835/rmlD* in the Paris genome, which has been involved in two events.

Another notable hotspot in the ST1 lineage is an 8,743bp region comprising genes from *lpp2595* to *lpp2604* (**Appendix Table 7**). The hotspot centres on the *lpp2599/tehB* gene, involved in six recombination events, and which encodes the tellurite resistance protein, TehB.

Across all six disease-associated STs, outer membrane proteins are commonly found within recombination hotspot regions (**Appendix Table 7**). Excluding those mentioned already (i.e. FadL and BamA), these include TolC (encoded by *ST23_00709*/*lpp0754*), involved in two recombination events in the ST23 lineage, and which has been implicated in the virulence of *L. pneumophila* (Ferhat *et al.*, 2009). Another is *lpa_01256*/*lpp0889*, also a TolC-like protein, which has been involved in two recombination events in the ST578 lineage. A small hotspot region in the ST37 lineage is immediately next to a known outer membrane protein (*ST37_01207*/*lpp1191*) described by Khemiri *et al.* (2008), and a hotspot region in the ST23 lineage is also very close to

the major outer membrane protein (*ST23_00628*/*lpp0671*). Furthermore, the *lpp0961* gene, involved in four recombination events in the ST1 lineage, encodes a protein homologous to AsmA, which is known to be involved in the assembly of outer membrane proteins in *E. coli*.

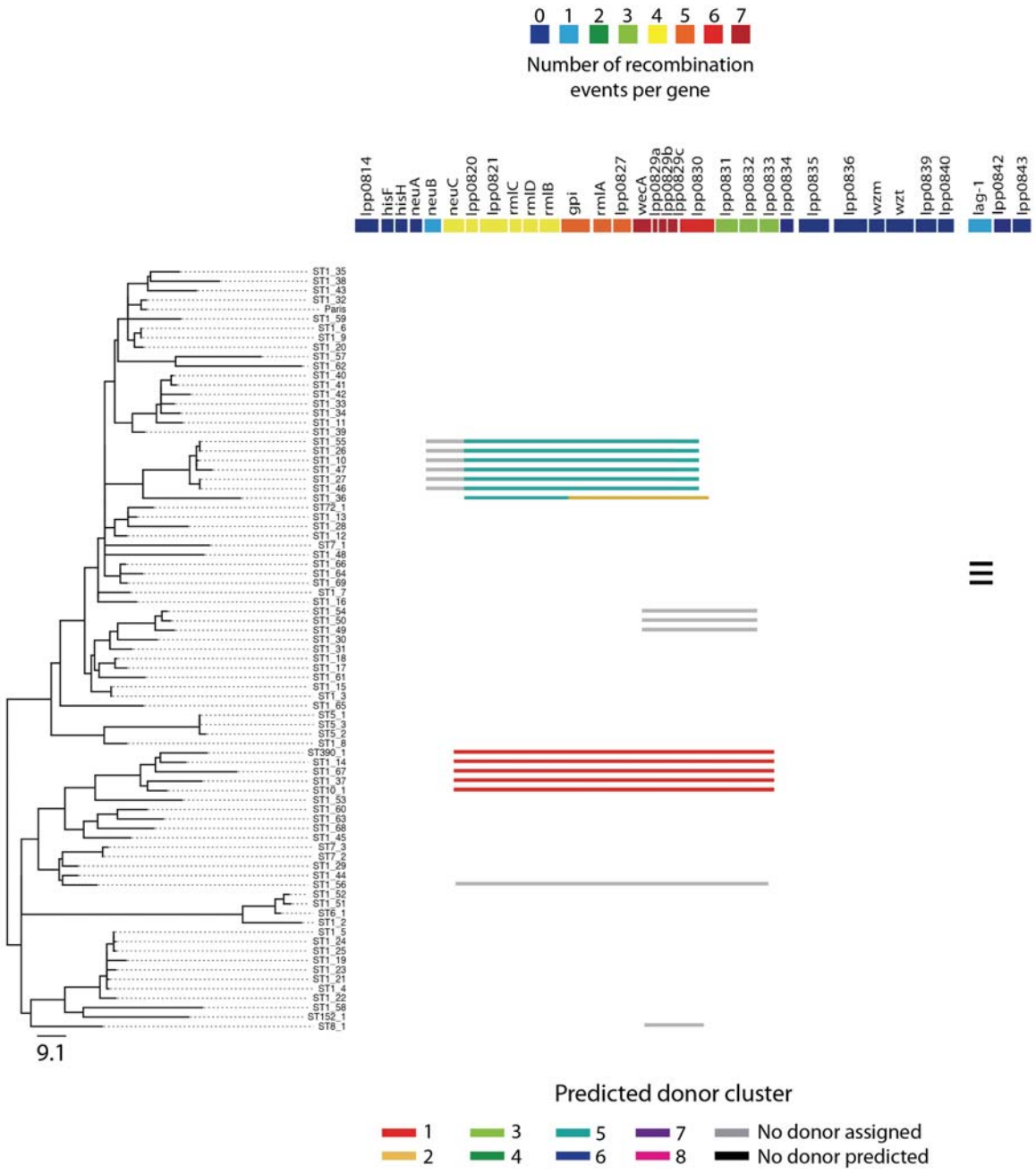


Figure 4.7. The LPS locus comprising hotspot 3 in the ST1 lineage. A maximum likelihood tree, constructed using only vertically inherited SNPs, is shown on the left and the scale indicates the number of SNPs. The recombination events are displayed as blocks, coloured according to

the BAPS cluster from which they are predicted to be derived. The genes within the LPS locus are shown at the top of the figure and coloured by the number of overlapping recombination regions.

A number of genes encoding putative or confirmed Dot/Icm effectors are also found within recombination hotspots across the different lineages (**Appendix Table 7**). These include *lpp0356*, involved in four recombination events in the ST1 lineage, which encodes an ankyrin repeat-containing protein that was originally found only in the Paris genome (Cazalet *et al.*, 2004). The *lpp2546* gene, which encodes the SdbB effector, has also been involved in four recombination events in the ST1 lineage. A further three ankyrin repeat-containing effector genes were identified within ST23 hotspots including *ST23_02606* (encoding LegA14), *ST23_00705* (encoding LegA8) and *ST23_00415* (encoding LegA7), all of which have been involved in two recombination events. Furthermore, the first described Dot/Icm effector, RalF, encoded by *ST23_01938*/*lpp1932*, was also found within a ST23 hotspot and predicted to have been involved in two recombination events.

Finally, while only 11 homologous recombination events were detected within the ST42 lineage, genes within one 14,572bp region have been affected by up to four recombination events. The hotspot region is centred on *ST42_02565*/*lpp2693*, which encodes the enhanced entry protein, EnhB, but also includes genes encoding the other enhanced entry proteins, EnhA and EnhC.

4.3.3 Inference of recombination donors

To predict the origin of the homologous recombination regions, the 536 *L. pneumophila* genomes used in this study were first divided into BAPS clusters, which were mapped onto a phylogenetic tree (**Figure 4.8**). Eight clusters were identified, seven of which comprised isolates from the *L. pneumophila pneumophila* subspecies (BAPS clusters 1-6, 8), and one with isolates from *L. pneumophila fraseri* (BAPS cluster 7).

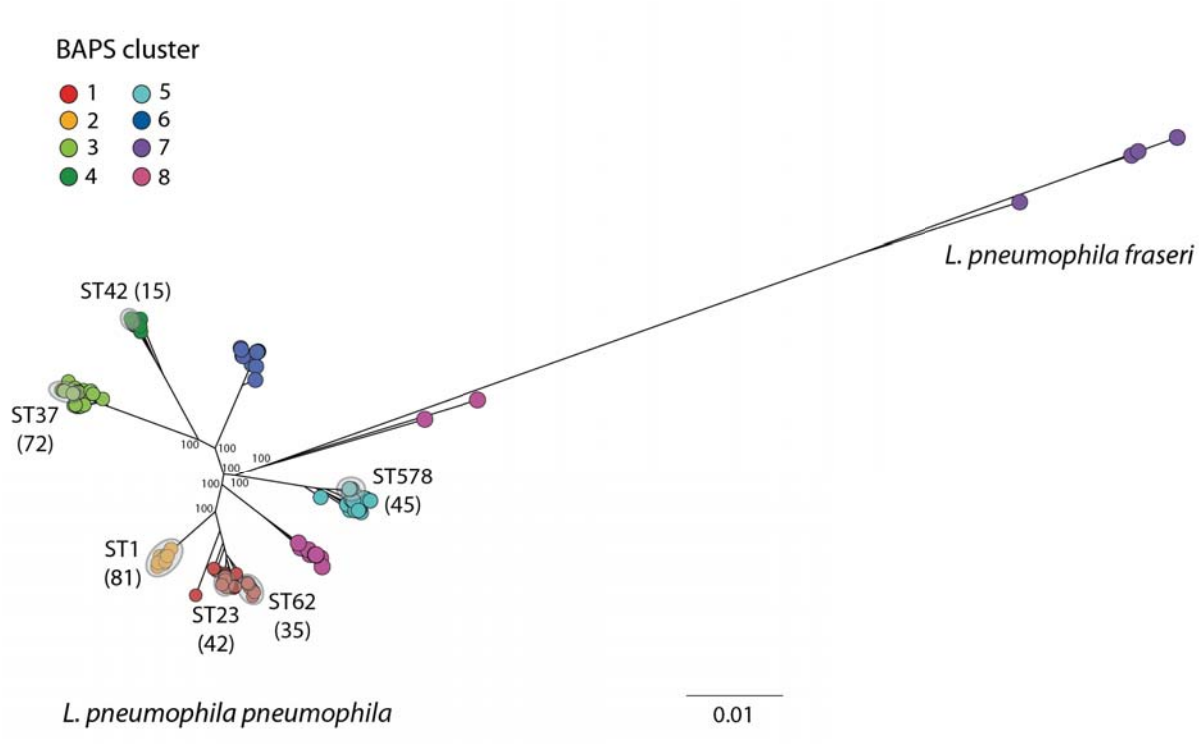


Figure 4.8. Maximum likelihood tree of 536 *L. pneumophila* isolates that are coloured by BAPS cluster. Grey circles also highlight the position of the six major disease-associated STs and the number of isolates belonging to each ST is indicated in brackets. The scale shows the number of SNPs per site. Bootstrap values, based on 1000 resamples, are shown for the major nodes of the tree.

All ancestral recombination sequences were extracted from the node downstream of the phylogenetic tree branch on which the recombination event was predicted to have occurred. Only regions greater than 500bp were used in this analysis, firstly because they were deemed more likely to be a "true" event, and secondly because small regions would likely have high similarity to many genomes. Each of the recombined fragments was used as a query in a BLASTn search against a database comprising all 536 *L. pneumophila* assembled genomes and the NCBI non-redundant database. The isolate with the highest bit score, together with the BAPS cluster from which it is derived, was considered the potential donor, provided that it covers at least 50% of the recombination fragment length and has a minimum of 99% nucleotide identity. Recombination fragments with no hits that met these thresholds were not assigned a donor ("No donor predicted"). Of the total 318 homologous recombination events

greater than 500bp predicted in the six STs, potential donors were predicted for 292 (91.8%) (**Table 4.4**). Many of the hits were almost perfect matches with 122 (41.8%) of the fragments having over 99.9% nucleotide identity, and 155 (53.1%) having hits that cover the full length of the recombination fragment (**Figure 4.9**).

Table 4.4. Number of homologous recombination events with predicted donors in each of the six STs.

ST	Total number of homologous recombination events	Number of recombination events >500bp	Number (and %) of filtered events with a predicted donor
1	198	193	176 (91.2%)
23	44	42	39 (92.9%)
37	13	12	9 (75.0%)
42	11	10	10 (100%)
62	48	39	36 (92.3%)
578	23	22	22 (100%)
Total	**337**	**318**	**292 (91.8%)**

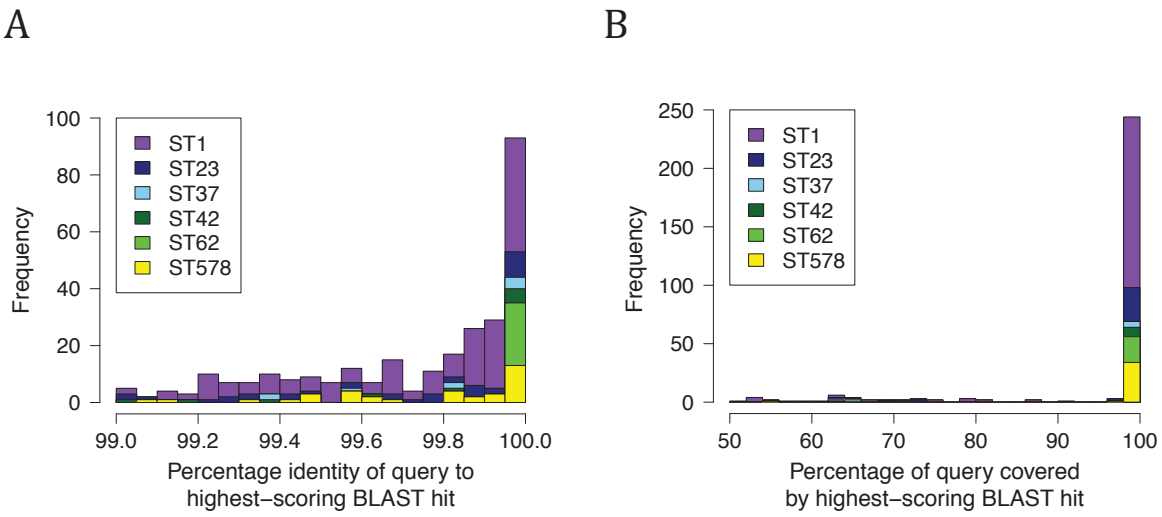


Figure 4.9. Similarity of the recombined regions to the predicted donors. The percentage nucleotide identity of the recombination fragments to the highest-scoring BLAST hit (A) and the percentage length covered by the highest-scoring BLAST hit (B).

The number of homologous recombination events in each of the six STs that are predicted to be derived from each of the eight BAPS clusters was calculated and visualised in a heat plot (**Figure 4.10**). Any events with equally good hits to isolates in more than one BAPS cluster were discarded for this analysis ("No donor assigned"). The heat plot illustrates that, in five of the six STs, recombination donors are most often from the same BAPS cluster as the recipient. The exception is ST37 in which the highest number of recombination fragments is derived from BAPS cluster 4, although its own cluster (BAPS cluster 3) accounts for the second highest number. However, all STs, with the exception of ST578, are also predicted to have acquired recombination fragments from clusters other than their own, demonstrating the occurrence of homologous recombination between major clusters of the *L. pneumophila pneumophila* subspecies. Interestingly, some BAPS clusters act frequently as donors (e.g. BAPS clusters 4 and 5) to other clusters, while others hardly donate except to isolates of their own cluster (e.g. BAPS clusters 2 and 3). Furthermore, just two events (one each in ST23 and ST62) are derived from the *L. pneumophila fraseri* subspecies (BAPS cluster 7).

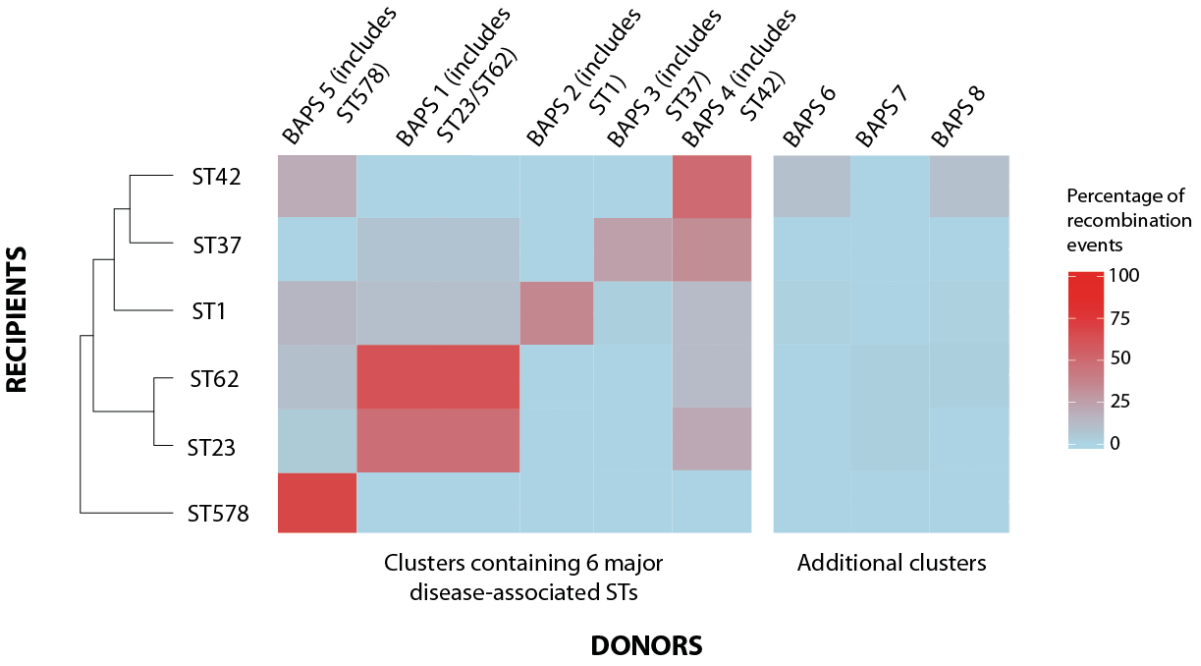


Figure 4.10. Predicted recombination donor clusters. A heat-map showing the percentage of recombination events detected in each of the six lineages that are predicted to be derived from each of the eight BAPS clusters. The six STs are shown in the left dendrogram constructed using

hierarchical clustering and based on the similarity of the predicted recombination donor lineages. The BAPS clusters are ordered from left to right based on the ordering of the six STs in the dendrogram. The column representing BAPS cluster 1, which contains both ST23 and ST62, is given twice the width as the other columns. The three BAPS clusters (6-8) that do not contain one of the six major disease-associated STs are shown on the right.

Recombination hotspot regions were next re-analysed to investigate whether the hotspots were driven by recombination events from the same or different BAPS clusters. The analysis focused on the ST1 lineage, which was previously found to contain the highest number of recombination events and the most prominent hotspots. The most notable hotspot region (hotspot 6), which was found to contain genes involved in up to 27 recombination events, was found to be driven mostly by recombination regions derived from the same BAPS cluster to which ST1 belongs (BAPS cluster 2) (**Figure 4.11**). However, a small number of recombination events that are predicted to be from BAPS cluster 5 were also observed in this region. Meanwhile, although some of the recombination events affecting the LPS locus (hotspot 3) could not be assigned a donor, others were derived from BAPS clusters 1, 2 and 5, suggesting that high diversity in this region may be important. Hotspot 4 appears to be driven by recombination events from BAPS clusters 5, 6 and 8 and contains no events derived from BAPS cluster 2 (to which ST1 belongs). However, the small number of events with predicted donors in most of these hotspots limits the conclusions that can be made.

For all homologous recombination events detected in the six STs, the percentage nucleotide identity between the imported fragment and the replaced fragment was calculated (**Figure 4.12A**). This analysis showed that 70% of homologous recombination events occurred between closely related isolates with >98% nucleotide similarity in the affected region, which agrees with our previous finding that most fragments are derived from the same BAPS cluster as the recipient. Interestingly, two peaks can be observed at ~98% identity and ~99.5-100% identity. These levels of divergence correspond to the nucleotide similarity observed between isolates belonging to different clusters or the same cluster, respectively (**Figure 4.12B**), and thus they represent recombination between and within clusters. Indeed, the recombination events that were predicted to be derived from the same BAPS cluster as the recipient have a

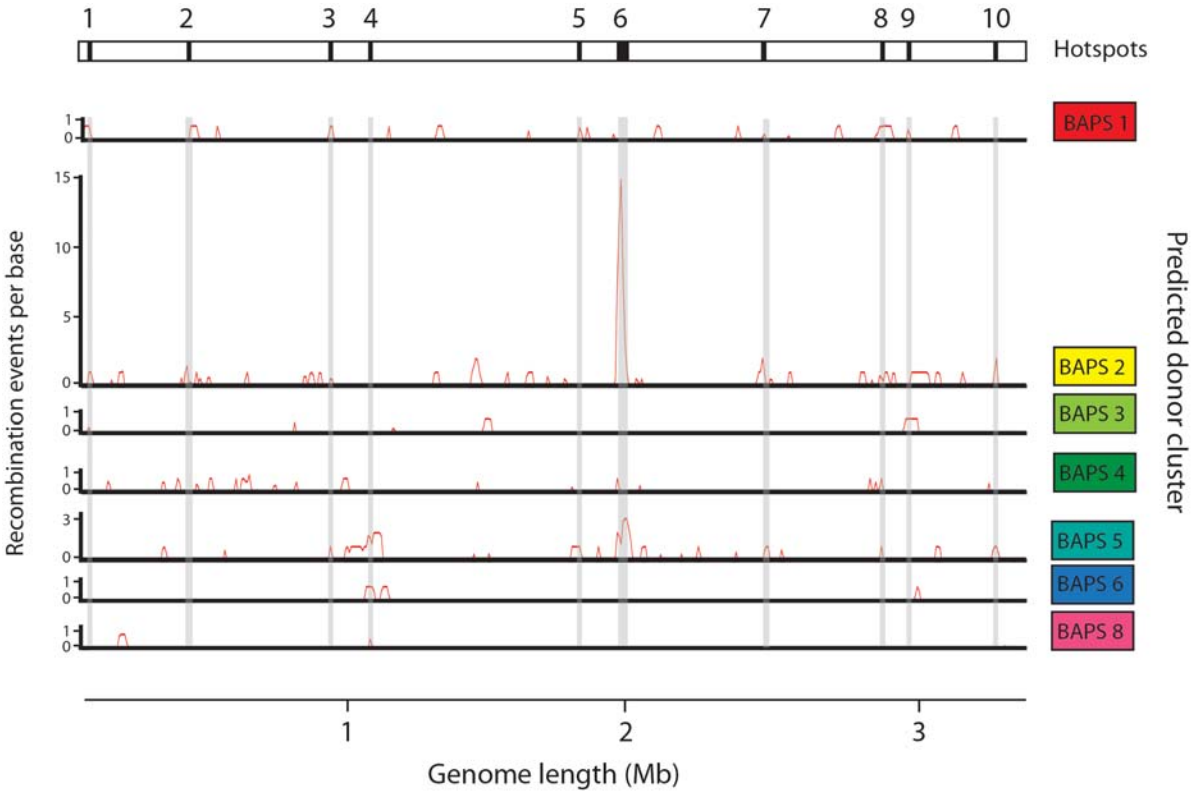


Figure 4.11. Diversity of recombination donors across the genome in the ST1 lineage. The number of recombination events per base that are derived from the different BAPS clusters are plotted. No events were predicted to be derived from BAPS cluster 7, which is thus excluded. The vertical grey bars correspond to the recombination hotspots.

mean nucleotide identity of 98.9% to the recipient genome while those predicted to be from a different cluster have a mean identity of 98.3%. Furthermore, very few recombination events were observed between isolates with <95% nucleotide identity, which is also concordant with our previous finding that very little recombination occurs between the two subspecies that share less than 95% nucleotide identity (**Figure 4.12A-B**).

Finally, the homologous recombination events that were predicted within the ST1 lineage were mapped onto the phylogenetic tree. This was to search for evidence of multi-fragment recombination, a process in which multiple non-contiguous segments that originate from the same molecule of DNA are imported into a recipient genome, and which is well documented in *S. pneumoniae* (Croucher *et al.*, 2012). Since the recombining fragments are non-contiguous, Gubbins will detect these as separate events

although the events should be predicted to have occurred on the same branch and have the same predicted donor. Indeed, **Figure 4.13** provides good evidence for the occurrence of this process in *L. pneumophila*, since many events with the same predicted donor, down to the BAPS cluster level and even the individual isolate level, are co-localised on branches. For example, over half of all recombination events in the ST1 lineage (100/193) occur on the same branch as another that is predicted to be derived from the same BAPS cluster.

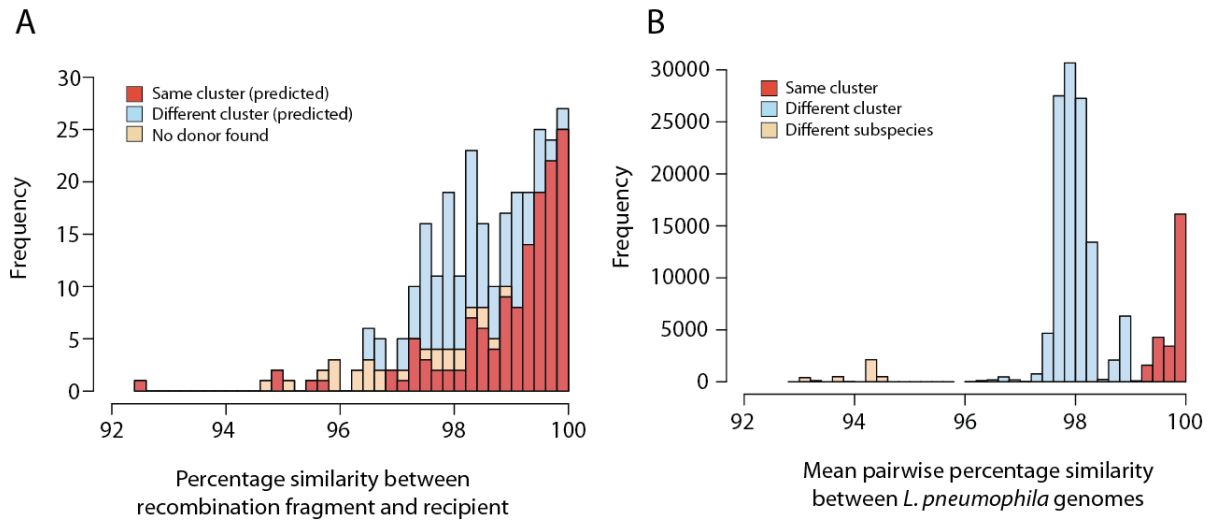


Figure 4.12. Sequence similarity between donors and recipients. A) Distribution of the percentage nucleotide similarities between the imported recombination fragments and the recipient sequence in all of the six STs. The events are categorised as being derived from the same or different BAPS clusters or with no donor lineage identified. B) Distribution of pairwise nucleotide similarities across the core genome amongst the 536 *L. pneumophila* isolates used in this study.

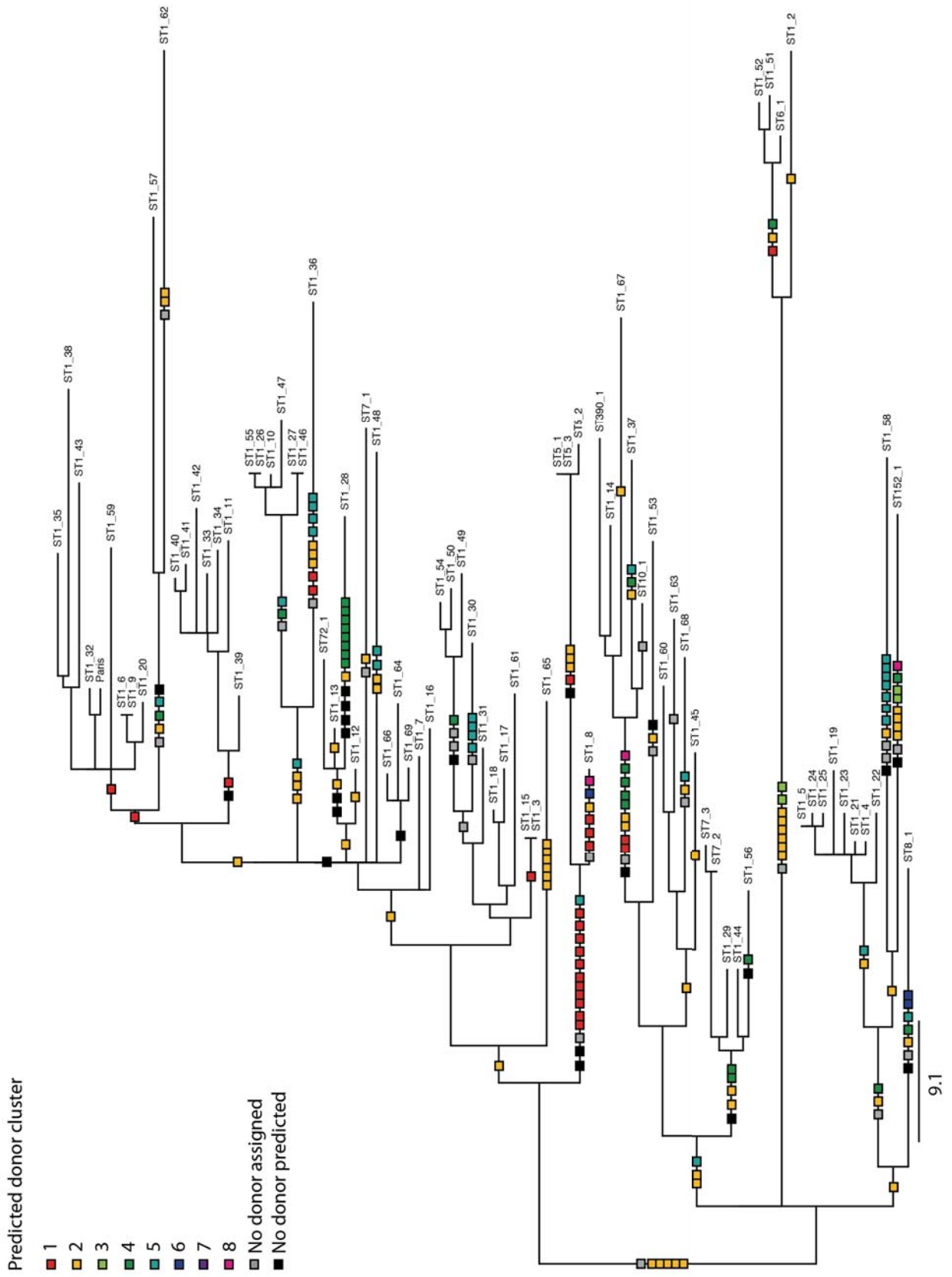

Predicted donor cluster
1
2
3
4
5
6
7
8
No donor assigned
No donor predicted
ST1_35
ST1_38
ST1_43
ST1_32
Paris
ST1_59
ST1_6
ST1_9
ST1_20
ST1_62
ST1_57
ST1_40
ST1_41
ST1_42
ST1_33
ST1_34
ST1_11
ST1_39
ST1_55
ST1_26
ST1_10
ST1_47
ST1_27
ST1_46
ST1_36
ST1_28
ST72_1
ST1_13
ST1_12
ST7_1
ST1_48
ST1_64
ST1_66
ST1_69
ST1_7
ST1_16
ST1_54
ST1_50
ST1_49
ST1_30
ST1_31
ST1_61
ST1_17
ST1_18
ST1_15
ST1_3
ST1_65
ST5_1
ST5_3
ST5_2
ST1_8
ST390_1
ST1_67
ST1_14
ST1_37
ST10_1
ST1_53
ST1_63
ST1_60
ST1_68
ST1_45
ST7_3
ST7_2
ST1_56
ST1_29
ST1_44
ST1_52
ST1_51
ST6_1
ST1_2
ST1_5
ST1_24
ST1_25
ST1_19
ST1_23
ST1_21
ST1_4
ST1_22
ST1_58
ST152_1
ST8_1
9.1

Figure 4.13. Maximum likelihood tree of 81 ST1 isolates with predicted recombination events mapped onto the branches (previous page). The tree was constructed using only vertically inherited SNPs and the scale bar indicates the number of SNPs. Predicted recombination events are represented by squares and coloured according to the BAPS cluster from which they are predicted to have been derived. Squares representing events with the same predicted donor at the isolate level and that have occurred on the same branches are joined together, and possibly represent multi-fragment recombination.

4.4 Discussion

Previous studies of several major disease-associated STs of *L. pneumophila* including ST578 (Sanchez-Buso *et al.*, 2014) and ST1, ST23, ST37 and ST62 (*Chapter 3* in this thesis) have shown that recombination is a dominant force in *L. pneumophila* evolution. However, the relative impacts of homologous and non-homologous recombination on *L. pneumophila* diversity have not been disentangled in any previous studies, and nor have the dynamics of homologous recombination been studied in detail in this species. Therefore, by studying six major disease-associated STs of *L. pneumophila*, including the five mentioned above and ST42, the aims of this thesis chapter were i) to determine the relative impact of homologous and non-homologous recombination on *L. pneumophila* evolution; ii) to identify homologous recombination hotspots; and iii) to explore the dynamics of recombination flux within the *L. pneumophila* species.

Analysis of all six lineages confirmed findings from *Chapter 3* that over 96% of SNPs in some lineages are found in recombined regions. However, when homologous recombination regions were distinguished from those associated with mobile genetic elements (non-homologous recombination) and repeat regions, the former were found to account for between 33.0% (ST62) and 80.0% (ST578) of the total SNPs. Remarkably, while homologous recombination events were shown to have occurred less frequently than *de novo* mutations in all lineages, they have contributed to between 20.8 and 93.8 times as many SNPs. These results support a very important role for homologous recombination in shaping the population structure and evolution of *L. pneumophila*, and highlight its potential to facilitate rapid adaptation to new niches such as modern, man-made water systems.

The fragments derived from homologous recombination were mostly found to be small (<10kb) although a small number of large events up to ~95kb were identified. A similar distribution of fragment sizes has also been reported in a previous study of recombination in *S. pneumoniae* in which it was suggested that transformation is optimised for exchanging short sequences rather than large features such as complete operons (Croucher *et al.*, 2012). This scenario could be favoured as it allows for larger numbers of potentially advantageous allele combinations to be tested.

Analysis of the genomic distribution of recombination events identified a total of 32 hotspot regions across the six STs. The most prominent hotspots were found in the ST1 lineage, in which the highest number of homologous recombination events was also detected. This is in concordance with the finding from *Chapter 3* that ST1 also has the highest number of vertically inherited mutations. Of particular note is a region containing genes that have been involved in up to 27 recombination events. The region appears to centre on the *hemB*/*lpp1771* gene, a porphobilinogen synthase (delta-aminolevulinic acid dehydratase), which is an enzyme involved in the biosynthesis of tetrapyrroles. However, the surrounding genes were also analysed and it seems more likely that the nearby gene, *lpp1773*, which encodes the outer membrane protein, FadL, may be responsible for driving the selection of recombination events in this region. Outer membrane proteins such as FadL could be under a high selection pressure to vary in order to interact with a highly variable aspect of the environment (e.g. diverse protozoan hosts), to escape protozoan predation, or to cope with an immune response during infection of host cells. However, since protozoa do not have an adaptive immune response, the latter possibility is unlikely unless higher organisms (e.g. humans) are also part of the infection cycle. As suggested in *Chapter 3*, it could be possible that this is indeed the case and that FadL is eliciting an immune response from humans.

Despite the prominence of this hotspot region in the ST1 lineage, it was not identified in any of the other STs apart from ST23, in which the hotspot region appeared to be centred on the outer membrane protein, BamA (encoded by *lpp1769*). BamA is also conserved across Gram-negative bacteria and is required for the assembly and insertion of beta-barrel proteins into the outer membrane (Tommassen, 2010). A *fadL*-like gene was also found within a recombination hotspot in the ST62 lineage, although this hotspot was found in a different part of the genome to the hotspot identified in the ST1

lineage. Further studies, perhaps involving larger number of isolates, would be useful to confirm the gene(s) that are driving these hotspots and to determine whether the prominent hotspot region in the ST1 lineage is also an important hotspot region in other lineages, or whether it represents a unique selection pressure in ST1 isolates.

Across the six STs, a number of other outer membrane proteins such as TolC were also identified within recombination hotspot regions. Of the many outer membrane proteins likely expressed on the surface of *L. pneumophila*, these results provide clues as to which ones are being selected for variation and part of dynamic environmental interactions. Furthermore, the LPS locus was also found in recombination hotspots in both the ST1 and ST62 lineages. Given that the LPS has been shown to be the major immunodominant antigen of *L. pneumophila* in the laboratory (Ciesielski *et al.*, 1986; Petzold *et al.*, 2013), it could be that it is also generating an immune response from humans and is thus under strong selection to vary. Horizontal exchange of the LPS locus also explains a previous observation that sg 1 isolates can have diverse genomic backgrounds, and that serogroups often do not correlate with overall genomic relatedness (Cazalet *et al.*, 2008).

A number of recombination hotspots also contain putative or confirmed effectors of the type IVB Dot/Icm secretion system of *L. pneumophila*. Dot/Icm effectors, of which there are over 300 described, manipulate a wide range of host cell processes and are essential to *L. pneumophila* pathogenesis (Ensminger, 2016). The genes found within recombination hotspots include that which encodes the first described effector, RalF. They likely represent those at the forefront of the arms race between *L. pneumophila* and its protozoan (or even human) hosts. It will be intriguing to decipher whether variation is being selected for within these effectors in order to take advantage of a wide variety of host species, or to counter changes in individual hosts. Larger sets of genomic data would also be useful to confirm the existence of these hotspots and further explore differences between lineages, which could suggest differences in hosts and infection strategies.

A number of other identified hotspots are also worthy of further investigation. These include a region within the ST1 lineage that appears to be centred on the *lpp2599/tehB* gene, which encodes the tellurite resistance protein, TehB. This has been identified in

both Gram-positive and Gram-negative bacterial pathogens and is involved in the detoxification of tellurite (Taylor, 1999). However, due to the apparent rarity of tellurium compounds in the environment, it has been suggested that this may not be the main function of this gene. For example, one study found that when the *tehB* gene from *S. pneumoniae* is expressed in *E. coli*, it results in a filamentous morphology (Liu & Taylor, 1999). The authors hypothesise that it might act as a methyltransferase that can alter the methylation of proteins related to cell division, thereby resulting in the generation of elongated cells (Liu & Taylor, 1999). Several studies have shown that *L. pneumophila* can also form long filamentous cells, particularly in response to stress conditions such as antibiotic exposure (Smalley *et al.*, 1980; Elliott & Rodgers, 1985), but it is unknown whether TehB is involved in this process. Further understanding of the function of TehB is therefore required to understand why this gene is associated with a recombination hotspot in the ST1 lineage.

Just one hotspot region was identified in the ST42 lineage, which appears to be centred on the enhanced entry gene, *enhB*. While little is known about *enhB*, the neighbouring gene, *ST42_02564*/*lpp2692*, which encodes the enhanced entry protein, EnhC, has been shown to be important for entry into host cells (Cirillo *et al.*, 2000) and to facilitate intracellular growth of *L. pneumophila* by evading immune recognition by the pattern recognition receptor (PRR), Nod1, in macrophages (Liu *et al.*, 2012). Further studies are required to understand why variability within the enhanced entry proteins might be advantageous, and also why these genes were found in a hotspot in the ST42 lineage and not others.

Recombination donors were predicted for over 90% of homologous recombination events (over 500bp) identified in the six STs. In all but one lineage, the highest number of recombination events was predicted to be from the same BAPS cluster as the recipient. This is an expected finding since homologous recombination is thought to require high, or even perfect, sequence homology between the donor and recipient at both ends of the recombination fragment (Majewski & Cohan, 1998), a scenario which is more likely between closely-related bacteria. However, all disease-associated STs, with the exception of ST578, have imported regions from BAPS clusters other than their own, thus also demonstrating evidence for homologous recombination between major clades

of the *L. p. pneumophila* subspecies. This suggests that different clades at least partially share the same ecological niche and that new adaptations can be shared freely between these disease-associated STs. This is in concordance with the findings of *Chapter 3*, whereby large regions were found to be transferred from the ST62 lineage to the ST47 lineage, and from the ST1 lineage to the ST37 lineage. Thus, the findings from both *Chapter 3* and the current chapter suggest that one of the disease-associated lineages could have initially acquired mutations or genes facilitating adaptation to human infection and these were subsequently shared with other lineages via homologous recombination (rather than independent acquisition by different lineages). Man-made water systems could provide a mixing vessel in which this process occurs. The findings also highlight the potential risk of more disease-associated strains from wide-ranging genomic backgrounds emerging rapidly in the future after having acquired adaptive features for human infection via homologous recombination.

Interestingly though, some BAPS clusters were predicted to act frequently as donors (e.g. BAPS 4 and 5), while others hardly donate, apart from to isolates of their own cluster (e.g. BAPS 2 and 3). A possible explanation for this could be related to the presence of restriction-modification systems in some lineages that prevent horizontal acquisition of DNA from lineages other than their own. Similar patterns whereby different lineages donate and receive DNA at different rates have also been observed in other species such as *S. pneumoniae* (Chewapreecha *et al.*, 2014), *C. trachomatis* (Harris *et al.*, 2012) and *E. coli* (Didelot *et al.*, 2012).

Only two recombination events detected within the six lineages were predicted to be from the *L. p. fraseri* subspecies. Given that this subspecies shares less than 95% nucleotide identity with the *L. p. pneumophila* subspecies, this was not an unexpected finding, given the high identity required for homologous recombination. It could be that these two subspecies have gradually diverged due to differing ecologies, and that eventually they may become different species that are fully incapable of exchange *via* homologous recombination.

Finally, the detection of multiple recombination events that are derived from the same donor and predicted on the same tree branch suggests the occurrence of multi-fragment

recombination. This is a process in which multiple non-contiguous segments of DNA originating from the same donor molecule are imported into the recipient genome during transformation and which has been documented in several studies of *S. pneumoniae* (Hiller *et al.*, 2010; Golubchik *et al.*, 2012; Croucher *et al.*, 2012). However, it could also be that the recombining isolates have shared a common niche for a prolonged period of time, and that multiple independent recombination events have occurred during this time. Thus further experimental studies will be required to confirm the occurrence of this process in *L. pneumophila*.

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

Declaration of work contributions

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

Publication

The following work has been published:

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

5.1 Introduction

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

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

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

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

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

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

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

5.2 Materials & Methods

5.2.1 Bacterial isolates

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

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

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

5.2.2 Study design

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

5.2.3 *De novo* assembly

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

5.2.4 Mapping/SNP-based analysis

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

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

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

5.2.5 Extended MLST

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

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

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

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

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

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

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

5.2.6 Gene presence/absence profiling

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

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

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

5.2.7 Kmer-based analysis

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

5.3 Results

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

5.3.1 Typability

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

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

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

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

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

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

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

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

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

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

5.3.2 Reproducibility

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

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

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

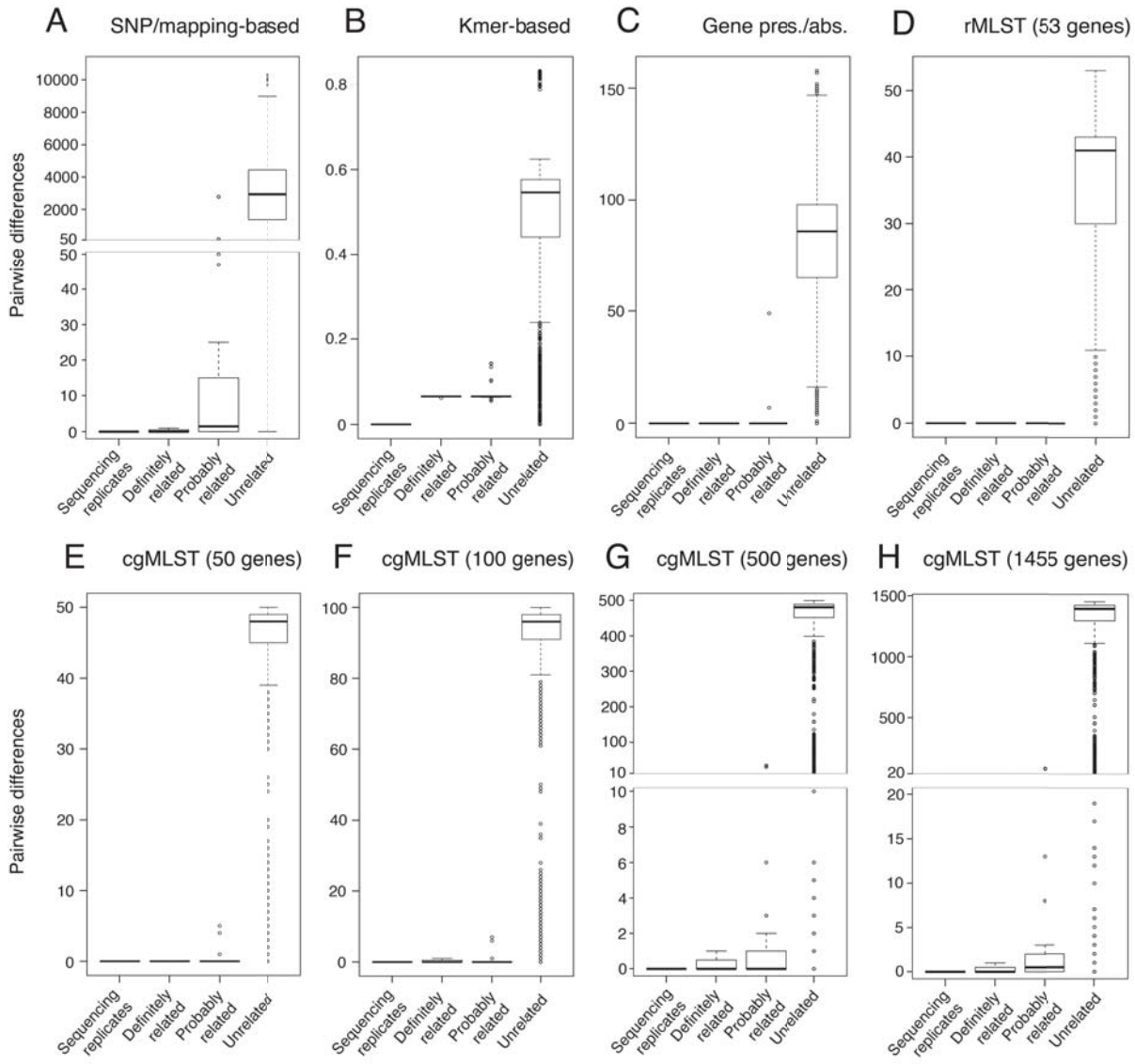


Figure 5.1. Pairwise differences between typing panel isolates using different WGS-based methods (A-H). Included are sequencing replicates (6 pairs), "definitely related" isolates (10 isolates comprising 4 sets), "probably related" isolates (34 isolates comprising 13 sets) and 79 epidemiologically "unrelated"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

5.3.3 Epidemiological concordance

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

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

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

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

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

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

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

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

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

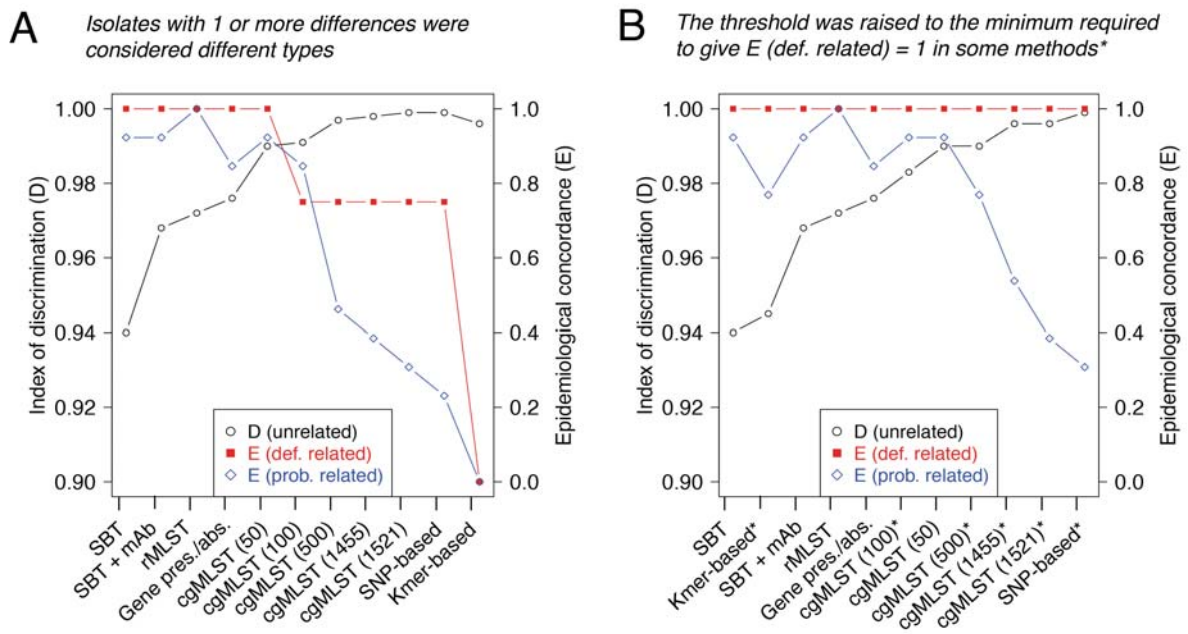


Figure 5.2. Index of discrimination (*D*) and epidemiological concordance (*E*) of the current and WGS-based methods. A) Isolates from the typing panel (*n*=106) were classified as the same type if they shared no differences and a different type if they shared 1 or more differences, except using the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

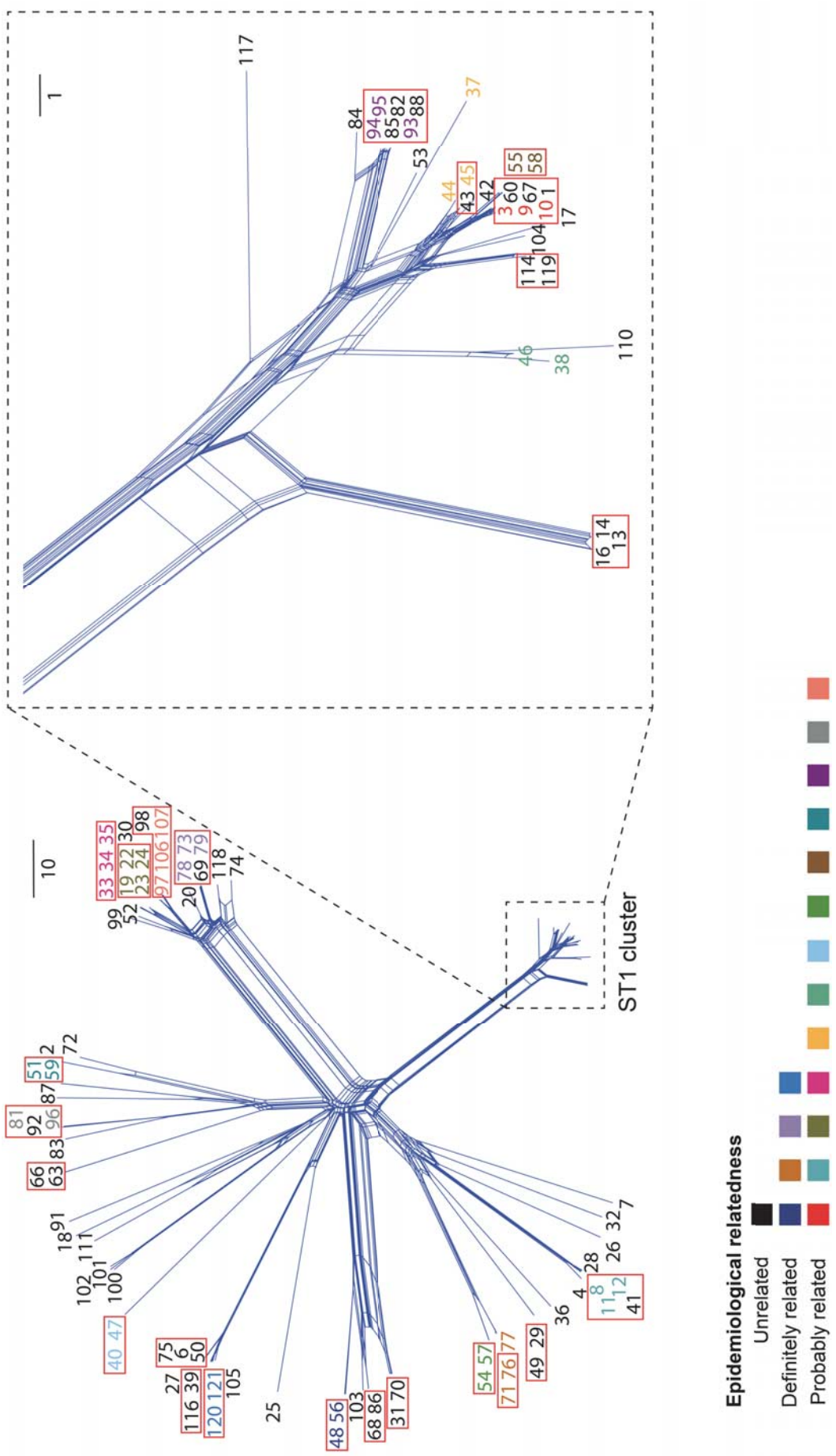

1
10
ST1 cluster
Epidemiological relatedness
Unrelated
Definitely related
Probably related

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

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

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

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

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

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

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

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

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

5.3.4 Discriminatory power

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

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

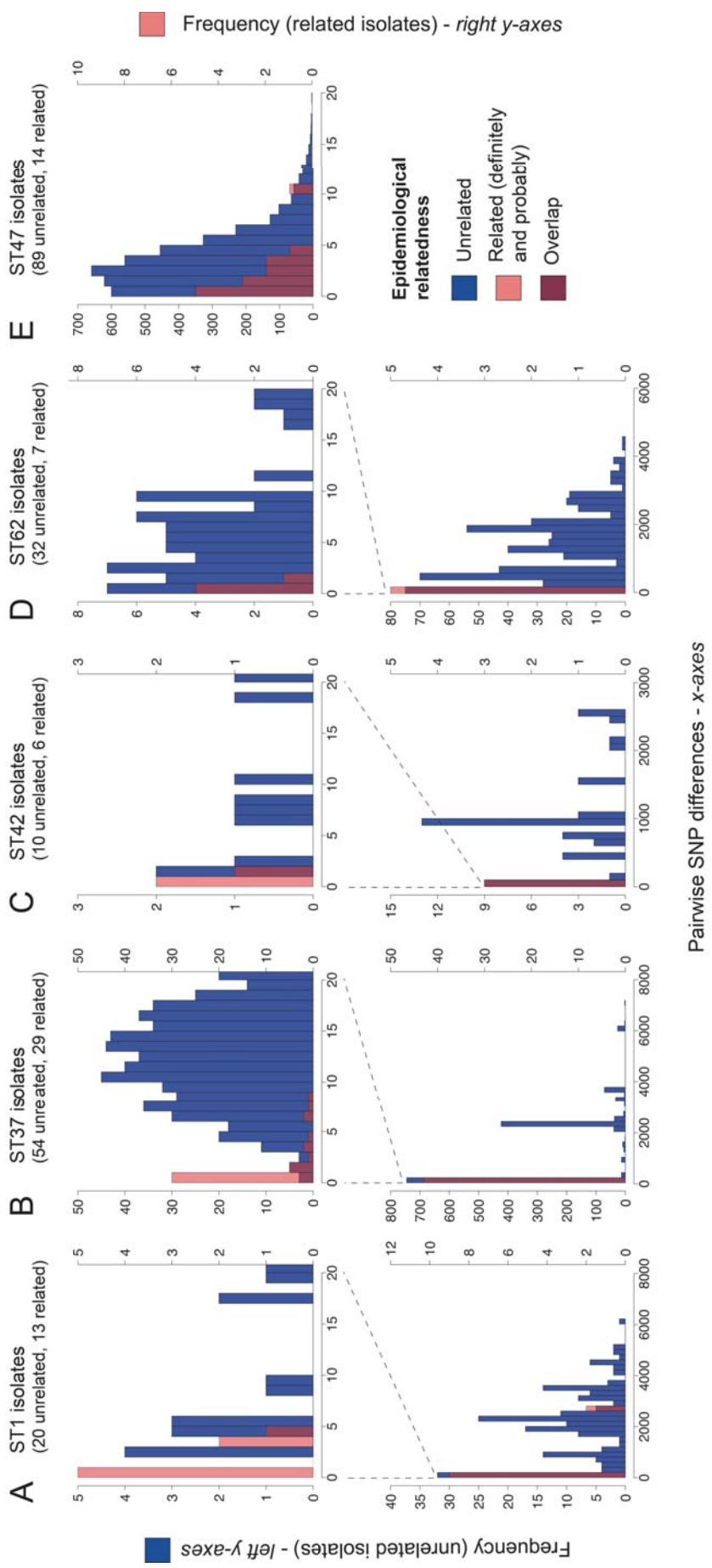

Frequency (related isolates) - right y-axes
A
ST1 isolates
(20 unrelated, 13 related)
B
ST37 isolates
(54 unrelated, 29 related)
C
ST42 isolates
(10 unrelated, 6 related)
D
ST62 isolates
(32 unrelated, 7 related)
E
ST47 isolates
(89 unrelated, 14 related)
Epidemiological relatedness
Unrelated
Related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

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

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


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

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

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

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

5.3.5 Stability

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

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

5.4 Discussion

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

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

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

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

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

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

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

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

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

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

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

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

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

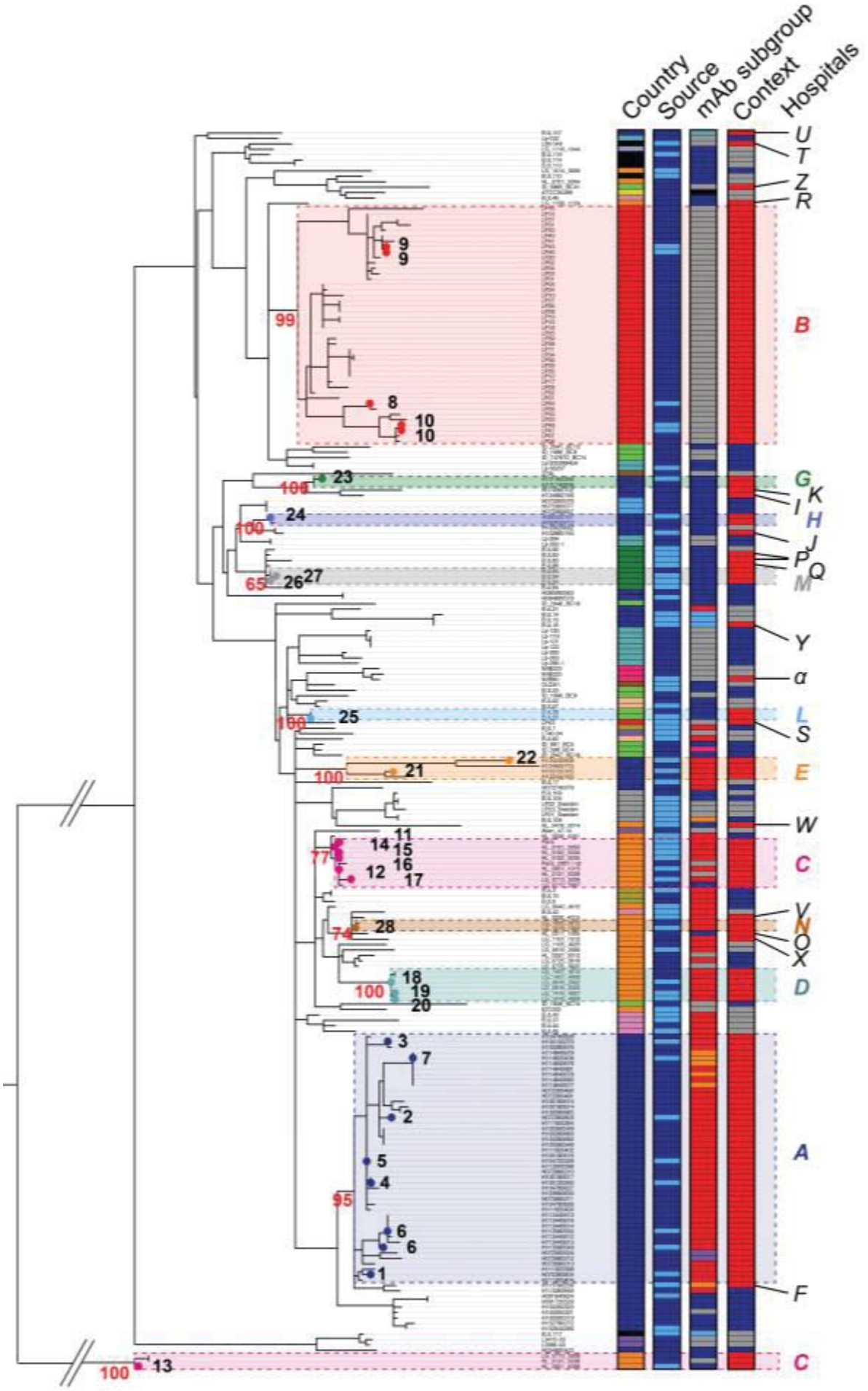
Sequence reads were mapped to the Paris (ST1) reference genome (Cazalet *et al.*, 2004) using SMALT v0.7.4 (available from: http://www.sanger.ac.uk/science/tools/smalt-0) and bases were called as described in *Chapter 2 (Materials & Methods)*. Recombined regions were identified and removed from the alignment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

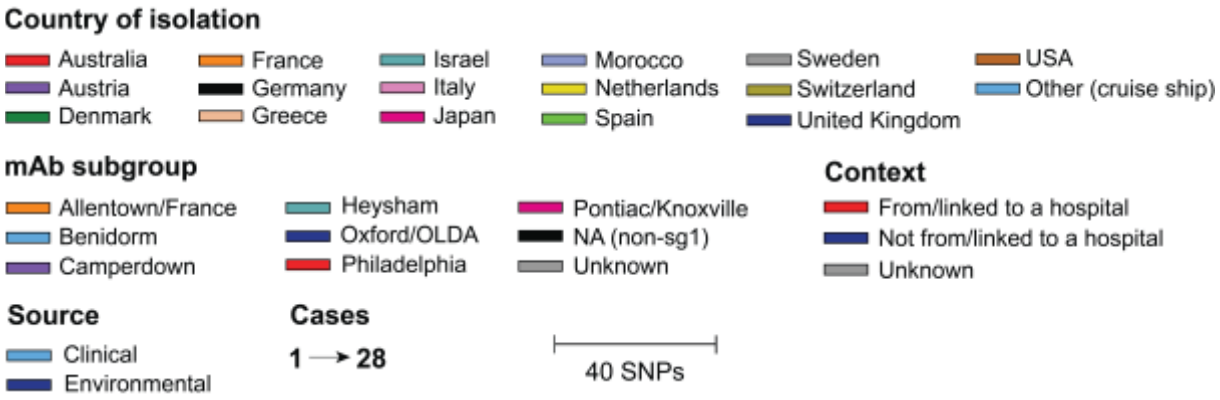


Figure 6.1. Maximum likelihood tree of 229 ST1 and ST1-derived isolates including those from or associated with hospitals (previous and current page). The tree was constructed using 1,741 SNPs identified after the removal of recombined regions. Environmental isolates from and clinical isolates linked 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.

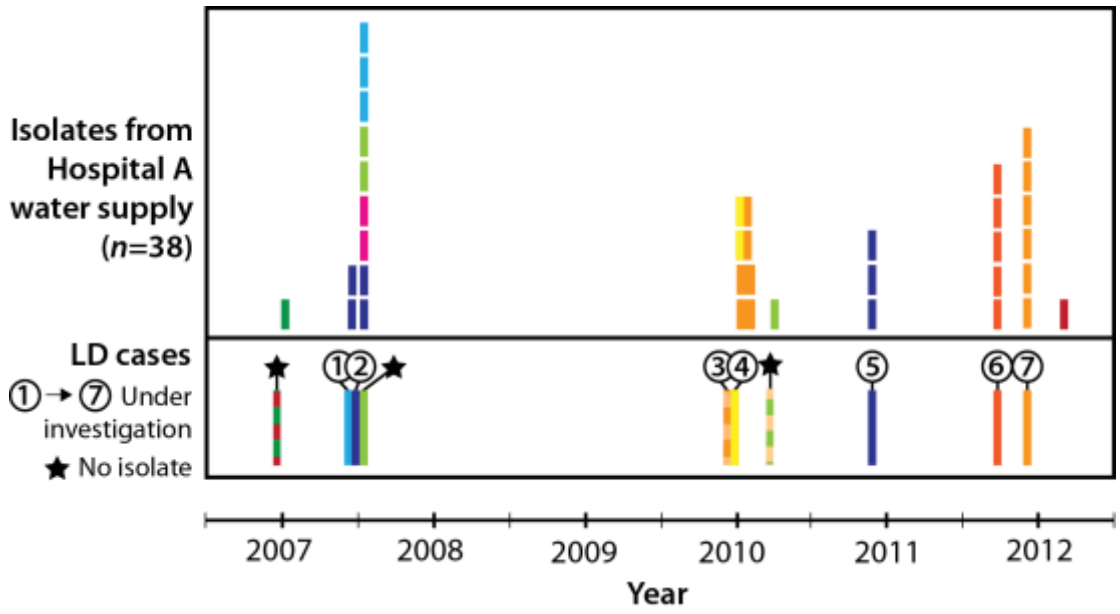


Figure 6.2. Time frame of legionellosis incidents and collection of environmental isolates at Hospital A. The time frame in which ten cases of Legionnaires' disease that were considered to have been acquired from Hospital A between the end of 2006 and 2011 is shown (bottom panel). Clinical isolates were 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

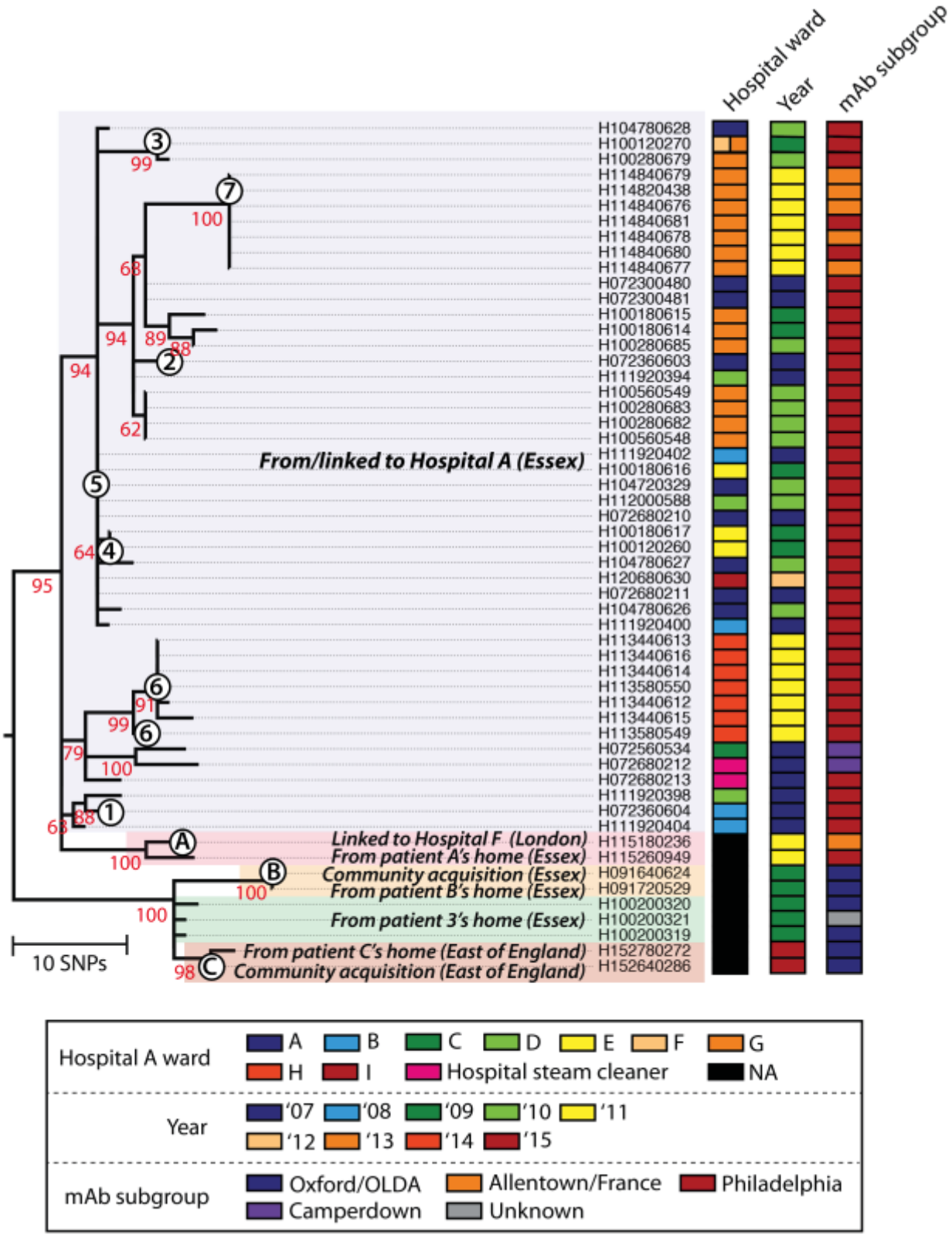


Figure 6.3. Phylogeny of isolates from Hospital A and the surrounding area. A zoomed-in section of the maximum likelihood tree presented in Figure 6.1 is shown, comprising environmental isolates from and clinical isolates linked to Hospital A. Clinical isolates from

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.

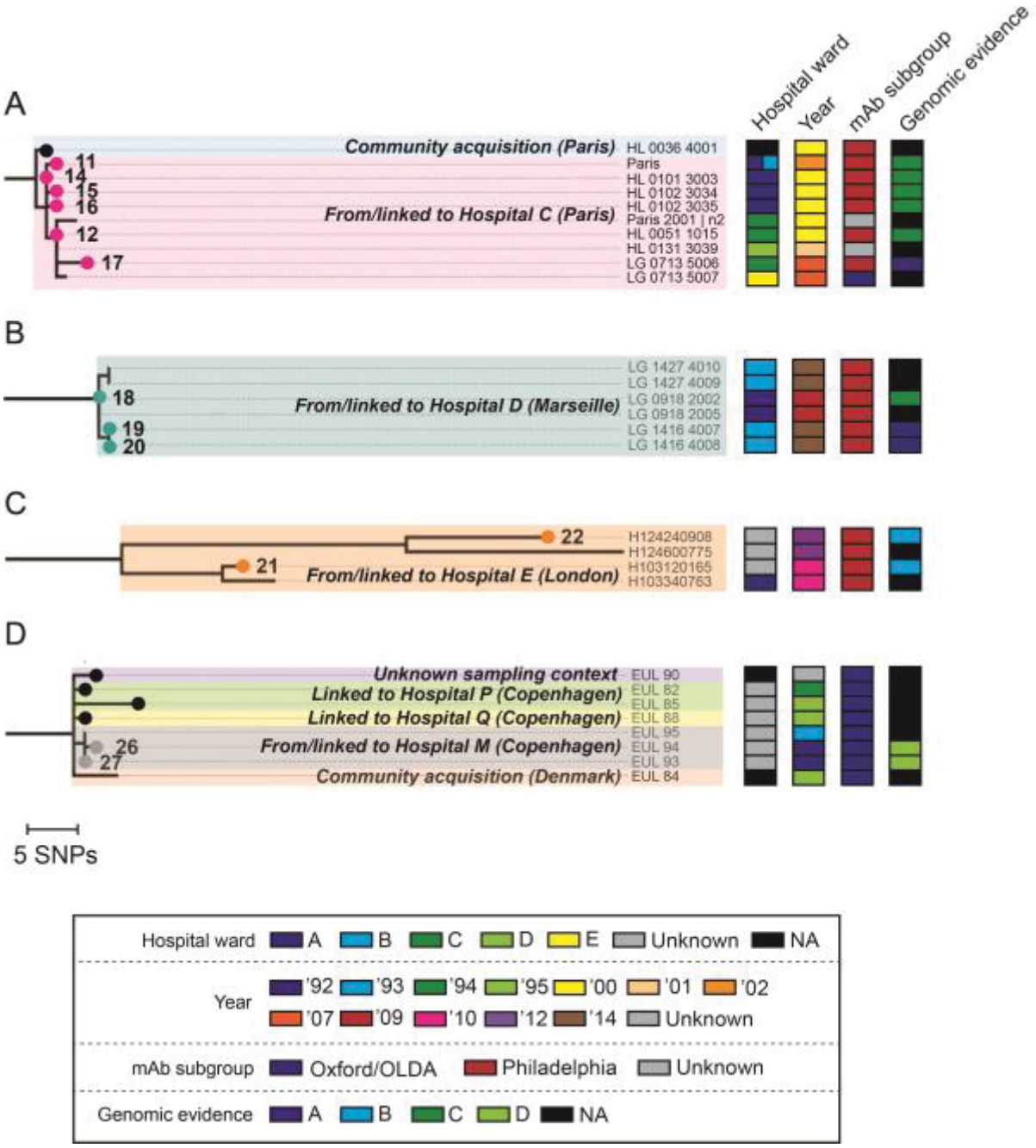


Figure 6.4. A-D) Zoomed-in sections of the maximum likelihood tree presented in Figure 1. All clinical isolates are indicated by small circles, with those from the 28 cases under investigation coloured and numbered as in Figure 1. Where applicable, isolates are additionally coloured in the right hand 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).

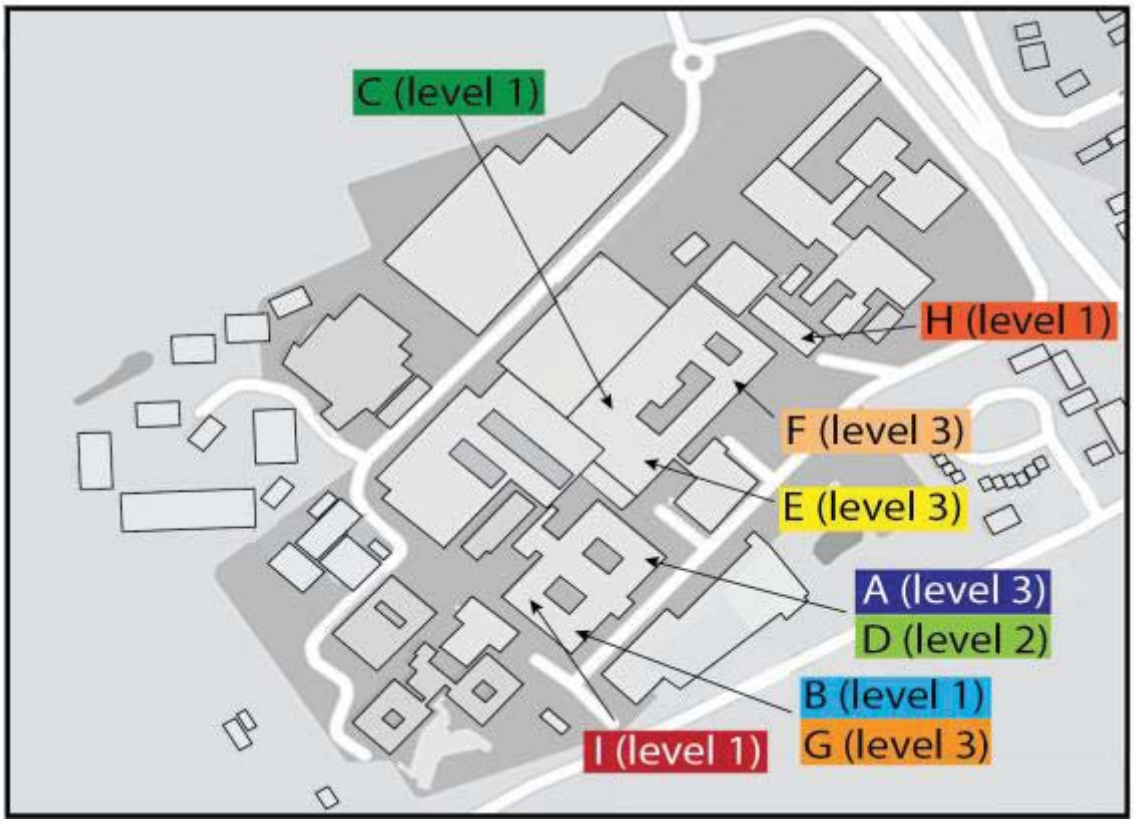


Figure 6.5. A plan of Hospital A. The wards in which the patients stayed are shown, as well as those in which the environmental isolates were obtained.

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.

7. Conclusions and future directions

7.1 A restatement of the research questions and aims

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

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

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

7.2 Key findings and future directions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

7.3 Closing remarks

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

8. References

Abu Kwaik, Y., Gag, L. Y., Stone, B. J., Venkataraman, C. & Harb, O. S. Invasion of protozoa by *Legionella pneumophila* and its role in bacterial ecology and pathogenesis. *Applied and Environmental Microbiology* **64**, 3127-3133 (1998).

Addiss, D. G. *et al.* Community-acquired Legionnaires' disease associated with a cooling tower - evidence for longer-distance transport of *Legionella pneumophila*. *American Journal of Epidemiology* **130**, 557-568 (1989).

Aguero-Rosenfeld, M. E. & Edelstein, P. H. Retrospective evaluation of the Du Pont radioimmunoassay kit for detection of *Legionella pneumophila* serogroup 1 antigenuria in humans. *Journal of Clinical Microbiology* **26**, 1775-1778 (1988).

Akermi, M. *et al.* Characterization of the *Legionella anisa* population structure by pulsed-field gel electrophoresis. *FEMS Microbiology Letters* **258**, 204-207 (2006).

Albert-Weissenberger, C., Cazalet, C. & Buchrieser, C. *Legionella pneumophila* - a human pathogen that co-evolved with fresh water protozoa. *Cellular and Molecular Life Sciences* **64**, 432-448 (2007).

Alexiou, S. D., Antoniadis, A., Papapaganagiotou, J. & Stefanou, T. Isolation of *Legionella pneumophila* from hotels of Greece. *European Journal of Epidemiology* **5**, 47-50 (1989).

Alleron, L., Merlet, N., Lacombe, C. & Frere, J. Long-term survival of *Legionella pneumophila* in the viable but nonculturable state after monochloramine treatment. *Current Microbiology* **57**, 497-502 (2008).

Amaro, F., Gilbert, J. A., Owens, S., Trimble, W., Shuman, H. A. Whole-genome sequence of the human pathogen *Legionella pneumophila* serogroup 12 strain 570-CO-H. *Journal of Bacteriology* **194**, 1613-1614 (2012).

Amemura-Maekawa, J. *et al.* Distribution of monoclonal antibody subgroups and sequence-based types among *Legionella pneumophila* serogroup 1 isolates derived from cooling tower water, bathwater, and soil in Japan. *Applied and Environmental Microbiology* **78**, 4263-4270 (2012).

Amemura-Maekawa, J. *et al.* Characterization of *Legionella pneumophila* isolates from patients in Japan according to serogroups, monoclonal antibody subgroups and sequence types. *Journal of Medical Microbiology* **59**, 653-659 (2010).

Arnow, P. M., Chou, T., Weil, D., Shapiro, E. N. & Kretzschmar, C. Nosocomial Legionnaires' disease caused by aerosolized tap water from respiratory devices. *Journal of Infectious Diseases* **146**, 460-467 (1982).

Atlas, R. M. *Legionella*: from environmental habitats to disease pathology, detection and control. *Environmental Microbiology* **1**, 283-293 (1999).

Baele, G. *et al.* Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution* **29**, 2157-2167 (2012).

Bangsborg, J. M., Jensen, B. N., Friismoller, A. & Bruun, B. Legionellosis in patients with HIV infection. *Infection* **18**, 342-346 (1990).

Bartley, P. B. *et al.* Hospital-wide eradication of a nosocomial *Legionella pneumophila* serogroup 1 outbreak. *Clinical Infectious Diseases* **62**, 273-279 (2016).

Beaute, J., Zucs, P., de Jong, B. & European Legionnaires' Disease Surveillance Network. Legionnaires' disease in Europe, 2009-2010. *Eurosurveillance* **18**, 6-12 (2013).

Bencini, M. A. *et al.* A case of Legionnaires' disease caused by aspiration of ice water. *Archives of Environmental & Occupational Health* **60**, 302-306 (2005).

Bennett, E. *et al.* Barrow-in-Furness: a large community legionellosis outbreak in the UK. *Epidemiology and Infection* **142**, 1763-1777 (2014).

Beres, S. B. *et al.* Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 4371-4376 (2010).

Bertelli, C. & Greub, G. Rapid bacterial genome sequencing: methods and applications in clinical microbiology. *Clinical Microbiology and Infection* **19**, 803-813 (2013).

Beyrer, K. *et al.* Legionnaires' disease outbreak associated with a cruise liner, August 2003: epidemiological and microbiological findings. *Epidemiology and Infection* **135**, 802-810 (2007).

Birtles, R. J., Harrison, T. G., Samuel, D. & Taylor, A. G. Evaluation of urinary antigen ELISA for diagnosing *Legionella pneumophila* serogroup 1 infection. *Journal of Clinical Pathology* **43**, 685-690 (1990).

Blatny, J. M. *et al.* Dispersion of *Legionella*-containing aerosols from a biological treatment plant, Norway. *Frontiers in Bioscience* **3**, 1300-1309 (2011).

Blatt, S. P. *et al.* Nosocomial Legionnaires' disease - aspiration as a primary mode of disease acquisition. *American Journal of Medicine* **95**, 16-22 (1993).

Blazquez Garrido, R. M. *et al.* Antimicrobial chemotherapy for Legionnaires disease: Levofloxacin versus macrolides. *Clinical Infectious Diseases* **40**, 800-806 (2005).

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579 (2011).

Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biology* **13** (2012).

Borchardt, J., Helbig, J. H. & Lueck, P. C. Occurrence and distribution of sequence types among *Legionella pneumophila* strains isolated from patients in Germany: common features and differences to other regions of the world. *European Journal of Clinical Microbiology & Infectious Diseases* **27**, 29-36 (2008).

Borella, P. *et al. Legionella* contamination in hot water of Italian hotels. *Applied and Environmental Microbiology* **71**, 5805-5813 (2005).

Bosch, T. *et al.* High resolution typing by whole genome mapping enables discrimination of LA-MRSA (CC398) strains and identification of transmission events. *PLOS ONE* **8** (2013).

Boshuizen, H. C. *et al.* Subclinical *Legionella* infection in workers near the source of a large outbreak of Legionnaires' disease. *Journal of Infectious Diseases* **184**, 515-518 (2001).

Bou, R. & Ramos, P. Outbreak of nosocomial Legionnaires' disease caused by a contaminated oxygen humidifier. *Journal of Hospital Infection* **71**, 381-383 (2009).

Brenner, D. J. *et al. Legionella pneumophila* serogroup Lansing 3 isolated from a patient with fatal pneumonia, and descriptions of *Legionella pneumophila* subsp. *pneumophila* subsp. nov., *Legionella pneumophila* subsp. *fraseri* subsp. nov., and *Legionella pneumophila* subsp. *pascullei* subsp. nov. *Journal of Clinical Microbiology* **26**, 1695-1703 (1988).

Bruggemann, H. *et al.* Virulence strategies for infecting phagocytes deduced from the *in vivo* transcriptional program of *Legionella pneumophila*. *Cellular Microbiology* **8**, 1228-1240 (2006).

Bryant, J. M. *et al.* Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. *Lancet* **381**, 1551-1560 (2013).

Burstein, D. *et al.* Genomic analysis of 38 *Legionella* species identifies large and diverse effector repertoires. *Nature Genetics* **48**, 167-175 (2016).

Camacho, C. *et al.* BLAST plus: architecture and applications. *BMC Bioinformatics* **10** (2009).

Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28** (2012).

Cassier, P. *et al.* Epidemiologic characteristics associated with ST23 clones compared to ST1 and ST47 clones of Legionnaires' disease cases in France. *New Microbes and New Infections* **3**, 29-33 (2015).

Castillo-Ramirez, S. *et al.* The impact of recombination of dN/dS within recently emerged bacterial clones. *PLOS Pathogens* **7** (2011).

Cazalet, C. *et al.* Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. *Nature Genetics* **36**, 1165-1173 (2004).

C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**, 2012-2018 (1998).

Chao, Y. *et al.* Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. *Scientific Reports* **3** (2013).

Chen, J. *et al. Legionella* effectors that promote nonlytic release from protozoa. *Science* **303**, 1358-1361 (2004).

Cheng, L., Connor, T. R., Siren, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular Biology and Evolution* **30**, 1224-1228 (2013).

Cherry, W. B. *et al.* Detection of Legionnaires' disease bacteria by direct immunofluorescent staining. *Journal of Clinical Microbiology* **8**, 329-338 (1978).

Chewapreecha, C. *et al.* Dense genomic sampling identifies highways of pneumococcal recombination. *Nature Genetics* **46**, 305-309 (2014).

Chien, M. C. *et al.* The genomic sequence of the accidental pathogen *Legionella pneumophila*. *Science* **305**, 1966-1968 (2004).

Christie, P. J., Atmakuri, K., Krishnamoorthy, V., Jakubowski, S. & Cascales, E. Biogenesis, architecture, and function of bacterial type IV secretion systems. *Annual Review*

of Microbiology **59**, 451-485 (2005).

Ciesielski, C. A., Blaser, M. J. & Wang, W. L. L. Serogroup specificity of *Legionella pneumophila* is related to lipopolysaccharide characteristics. *Infection and Immunity* **51**, 397-404 (1986).

Cirillo, S. L. G., Lum, J. & Cirillo, J. D. Identification of novel loci involved in entry by *Legionella pneumophila*. *Microbiology* **146**, 1345-1359 (2000).

Coetzee, N. *et al.* An outbreak of Legionnaires' disease associated with a display spa pool in retail premises, Stoke-on-Trent, United Kingdom, July 2012. *Eurosurveillance* **17**, 6-8 (2012).

Cordes, L. G. *et al.* Isolation of *Legionella pneumophila* from hospital shower heads. *Annals of Internal Medicine* **94**, 195-197 (1981).

Correia, A. M. *et al.* Probable person-to-person transmission of Legionnaires' disease. *New England Journal of Medicine* **374**, 497-498 (2016).

Coscolla, M., Comas, I. & Gonzalez-Candelas, F. Quantifying nonvertical inheritance in the evolution of *Legionella pneumophila*. *Molecular Biology and Evolution* **28**, 985-1001 (2011).

Coscolla, M., Fernandez, C., Colomina, J., Sanchez-Buso, L. & Gonzalez-Candelas, F. Mixed infection by *Legionella pneumophila* in outbreak patients. *International Journal of Medical Microbiology* **304**, 307-313 (2014).

Cristino, S., Legnani, P. P. & Leoni, E. Plan for the control of *Legionella* infections in long-term care facilities: role of environmental monitoring. *International Journal of Hygiene and Environmental Health* **215**, 279-285 (2012).

Croucher, N. J., Harris, S. R., Barquist, L., Parkhill, J. & Bentley, S. D. A high-resolution view of genome-wide pneumococcal transformation. *PLOS Pathogens* **8** (2012).

Croucher, N. J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430-434 (2011).

Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* **43** (2015).

Currie, S. L., Beattie, T. K., Knapp, C. W. & Lindsay, D. S. J. *Legionella spp*. in UK composts - a potential public health issue? *Clinical Microbiology and Infection* **20**, O224-O229 (2014).

D'Auria, G., Jimenez-Hernandez, N., Peris-Bondia, F., Moya, A., Latorre, A. *Legionella pneumophila* pangenome reveals strain-specific virulence factors. *BMC Genomics* **11** (2010).

Dallman, T. J. *et al.* An investigation of the diversity of strains of enteroaggregative *Escherichia coli* isolated from cases associated with a large multi-pathogen foodborne outbreak in the UK. *PLOS One* **9** (2014).

de Been, M. *et al.* Core genome multilocus sequence typing scheme for high-resolution typing of *Enterococcus faecium*. *Journal of Clinical Microbiology* **53**, 3788-3797 (2015).

de Felipe, K. S. *et al.* Evidence for acquisition of *Legionella* type IV secretion substrates via interdomain horizontal gene transfer. *Journal of Bacteriology* **187**, 7716-7726 (2005).

Dennis, P. J. L. Isolation of *Legionella* from environmental specimens in *A laboratory manual for Legionella* (ed. Harrison, T. G. & Taylor, A. G.) 31-44 (Wiley & Sons Ltd, 1988).

De Zoysa, A. S. & Harrison, T. G. Molecular typing of *Legionella pneumophila* serogroup 1 by pulsed-field gel electrophoresis with SfiI and comparison of this method with restriction fragment-length polymorphism analysis. *Journal of Medical Microbiology* **48**, 269-278 (1999).

Den Boer, J. W., Nijhof, J. & Friesema, I. Risk factors for sporadic community-acquired Legionnaires' disease. A 3-year national case-control study. *Public Health* **120**, 566-571 (2006).

Den Boer, J. W. *et al.* A large outbreak of Legionnaires' disease at a flower show, the Netherlands, 1999. *Emerging Infectious Diseases* **8**, 37-43 (2002).

Desai, R., Welsh, C., Summy, M., Farone, M. & Newsome, A. L. The potential of *in situ* hybridization and an immunogold assay to identify *Legionella* associations with other microorganisms. *Journal of Microbiological Methods* **37**, 155-164 (1999).

Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A. & Crook, D. W. Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics* **13**, 601-612 (2012).

Didelot, X., Meric, G., Falush, D. & Darling, A. E. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* **13** (2012).

Diederen, B. M. W. *Legionella spp.* and Legionnaires' disease. *Journal of Infection* **56**, 1-12 (2008).

Dominguez, A. *et al.* Factors influencing the case-fatality rate of Legionnaires' disease. *International Journal of Tuberculosis and Lung Disease* **13**, 407-412 (2009).

Dominguez, J. *et al.* Evaluation of a rapid immunochromatographic assay for the detection of *Legionella* antigen in urine samples. *European Journal of Clinical Microbiology & Infectious Diseases* **18**, 896-898 (1999).

Dominguez, J. A. *et al.* Comparison of the Binax *Legionella* urinary antigen enzyme immunoassay (EIA) with the biotest *Legionella* Urin Antigen EIA for detection of *Legionella* antigen in both concentrated and nonconcentrated urine samples. *Journal of Clinical Microbiology* **36**, 2718-2722 (1998).

Dreyfus, L. A. & Iglewski, B. H. Conjugation-mediated genetic exchange in *Legionella pneumophila*. *Journal of Bacteriology* **161**, 80-84 (1985).

Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**, 1969-1973 (2012).

Ducret, A., Chabalier, M. & Dukan, S. Characterization and resuscitation of 'non-culturable' cells of *Legionella pneumophila*. *BMC Microbiology* **14** (2014).

Edelstein, P. H. Comparative study of selective media for isolation of *Legionella pneumophila* from potable water. *Journal of Clinical Microbiology* **16**, 697-699 (1982).

Edelstein, P. H. Laboratory diagnosis of infections caused by Legionellae. *European Journal of Clinical Microbiology & Infectious Diseases* **6**, 4-10 (1987).

Edelstein, P. H. Laboratory diagnosis of Legionnaires' disease - An update from 1984, in *Legionella: Current Status and Emerging Perspectives* (ed. Barbaree, J. M., Breiman, R. F. & Dufour, A. P.) 7-11 (American Society for Microbiology, 1993).

Edelstein, P. H., Snitzer, J. B. & Bridge, J. A. Enhancement of recovery of *Legionella pneumophila* from contaminated respiratory tract specimens by heat. *Journal of Clinical Microbiology* **16**, 1061-1065 (1982).

Eickhoff, T. C. Epidemiology of Legionnaires' disease. *Annals of Internal Medicine* **90**, 499-502 (1979).

Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138 (2009).

Eitrem, R., Forsgren, A. & Nilsson, C. Pneumonia and acute pancreatitis most probably caused by a *Legionella longbeachae* infection. *Scandinavian Journal of Infectious Diseases* **19**, 381-382 (1987).

Elliott, T. S. J. & Rodgers, F. G. Morphological response and growth characteristics of *Legionella pneumophila* exposed to ampicillin and erythromycin. *Journal of Medical Microbiology* **19**, 383-390 (1985).

Ensminger, A. W. *Legionella pneumophila*, armed to the hilt: justifying the largest arsenal of effectors in the bacterial world. *Current Opinion in Microbiology* **29**, 74-80 (2016).

Ensminger, A. W., Yassin, Y., Miron, A. & Isberg, R. R. Experimental evolution of *Legionella pneumophila* in mouse macrophages leads to strains with altered determinants of environmental survival. *PLOS Pathogens* **8** (2012).

Epalle, T. *et al.* Viable but not culturable forms of *Legionella pneumophila* generated after heat shock treatment are infectious for macrophage-like and alveolar epithelial cells after resuscitation on *Acanthamoeba polyphaga*. *Microbial Ecology* **69**, 215-224 (2015).

European Centre for Disease Prevention and Control. Legionnaires' disease in Europe, 2013 (2015).

European Working Group for *Legionella* Infections (EWGLI). Sequence-based typing (SBT) database. Available at: http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php.

Euser, S. M. *et al. Legionella* prevention in the Netherlands: an evaluation using genotype distribution. *European Journal of Clinical Microbiology & Infectious Diseases* **32**, 1017-1022 (2013).

Farris, J. S. Methods for computing Wagner trees. *Systematic Zoology* **19**, 83-92 (1970).

Feil, E. J. *et al.* Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 4276-4276 (2001).

Feldman, M., Zusman, T., Hagag, S. & Segal, G. Coevolution between nonhomologous but functionally similar proteins and their conserved partners in the *Legionella* pathogenesis system. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 12206-12211 (2005).

Ferhat, M. *et al.* The TolC protein of *Legionella pneumophila* plays a major role in multi-

drug resistance and the early steps of host invasion. *PLOS One* **4** (2009).

Fields, B. S. The molecular ecology of legionellae. *Trends in Microbiology* **4**, 286-290 (1996).

Fields, B. S., Benson, R. F. & Besser, R. E. *Legionella* and Legionnaires' disease: 25 years of investigation. *Clinical Microbiology Reviews* **15**, 506-526 (2002).

Fields, B. S. *et al.* Pontiac fever due to *Legionella micdadei* from a whirlpool spa: Possible role of bacterial endotoxin. *Journal of Infectious Diseases* **184**, 1289-1292 (2001).

Fields, B. S., Shotts, E. B., Feeley, J. C., Gorman, G. W. & Martin, W. T. Proliferation of *Legionella pneumophila* as an intracellular parasite of the ciliated protozoan *Tetrahymena pyriformis*. *Applied and Environmental Microbiology* **47**, 467-471 (1984).

Fisman, D. N. *et al.* It's not the heat, it's the humidity: Wet weather increases legionellosis risk in the greater Philadelphia metropolitan area. *Journal of Infectious Diseases* **192**, 2066-2073 (2005).

Fleischmann, R. D. *et al.* Whole genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512 (1995).

Fliermans, C. B. *et al.* Ecological distribution of *Legionella pneumophila*. *Applied and Environmental Microbiology* **41**, 9-16 (1981).

Fliermans, C. B., Cherry, W. B., Orrison, L. H. & Thacker, L. Isolation of *Legionella pneumophila* from nonepidemic-related aquatic habitats. *Applied and Environmental Microbiology* **37**, 1239-1242 (1979).

Flournoy, D. J., Belobraydic, K. A., Silberg, S. L., Lawrence, C. H. & Guthrie, P. J. False positive *Legionella pneumophila* direct immunofluorescent monoclonal antibody test caused by *Bacillus cereus* spores. *Diagnostic Microbiology and Infectious Disease* **9**, 123-125 (1988).

Ford, C. B. *et al. Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature Genetics* **45**, 784-790 (2013).

Fraser, D. W., Deubner, D. C., Hill, D. L. & Gilliam, D. K. Nonpneumonic, short incubation period legionellosis (Pontiac fever) in men who cleaned a steam turbine condenser. *Science* **205**, 690-691 (1979).

Fraser, D. W. *et al.* Legionnaires' disease - Description of an epidemic of pneumonia. *New England Journal of Medicine* **297**, 1189-1197 (1977).

Friedman, S., Spitalny, K., Barbaree, J., Faur, Y. & McKinney, R. Pontiac fever outbreak associated with a cooling tower. *American Journal of Public Health* **77**, 568-572 (1987).

Fry, N. K. *et al.* A multicenter evaluation of genotypic methods for the epidemiologic typing of *Legionella pneumophila* serogroup 1: results of a pan-European study. *Clinical Microbiology & Infection* **5**, 462-477 (1999).

Fry, N. K. *et al.* Designation of the European Working Group on *Legionella* Infection (EWGLI) amplified fragment length polymorphism types of *Legionella pneumophila* serogroup 1 and results of intercentre proficiency testing using a standard protocol. *European Journal of Clinical Microbiology & Infectious Diseases*

21, 722-728 (2002).

Fry, N. K. *et al.* Assessment of intercentre reproducibility and epidemiological concordance of *Legionella pneumophila* serogroup 1 genotyping by amplified fragment length polymorphism analysis. *European Journal of Clinical Microbiology & Infectious Diseases* **19**, 773-780 (2000).

Gaia, V. *et al.* Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila*. *Journal of Clinical Microbiology* **43**, 2047-2052 (2005).

Gaia, V., Fry, N. K., Harrison, T. G. & Peduzzi, R. Sequence-based typing of *Legionella pneumophila* serogroup 1 offers the potential for true portability in legionellosis outbreak investigation. *Journal of Clinical Microbiology* **41**, 2932-2939 (2003).

Gao, L. Y., Harb, O. S. & Abu Kwaik, Y. Utilization of similar mechanisms by *Legionella pneumophila* to parasitize two evolutionarily distant host cells, mammalian macrophages and protozoa. *Infection and Immunity* **65**, 4738-4746 (1997).

Garau, J., Fritsch, A., Arvis, P. & Read, R. C. Clinical efficacy of moxifloxacin versus comparator therapies for community-acquired pneumonia caused by *Legionella spp. Journal of Chemotherapy* **22**, 264-266 (2010).

Garcia, M. T., Jones, S., Pelaz, C., Millar, R. D. & Abu Kwaik, Y. *Acanthamoeba polyphaga* resuscitates viable non-culturable *Legionella pneumophila* after disinfection. *Environmental Microbiology* **9**, 1267-1277 (2007).

Gardy, J. L. *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *New England Journal of Medicine* **364**, 730-739 (2011).

Ginevra, C. *et al.* Lorraine strain of *Legionella pneumophila* serogroup 1, France. *Emerging Infectious Diseases* **14**, 673-675 (2008).

Ginevra, C. *et al. Legionella pneumophila* sequence type 1/Paris pulsotype subtyping by spoligotyping. *Journal of Clinical Microbiology* **50**, 696-701 (2012).

Ginevra, C. *et al.* Evaluation of a nested-PCR-derived sequence-based typing method applied directly to respiratory samples from patients with Legionnaires' disease. *Journal of Clinical Microbiology* **47**, 981-987 (2009).

Girod, J. C. *et al.* Pneumonic and non-pneumonic forms of legionellosis - the result of a common-source exposure to *Legionella pneumophila*. *Archives of Internal Medicine* **142**, 545-547 (1982).

Gladman, S. & Seemann, T. Velvet Optimiser. Available at: http://bioinformatics.net.au/software.velvetoptimiser.shtml

Glick, T. H. *et al.* Pontiac fever – an epidemic of unknown etiology in a health department: I. Clinical and epidemiologic aspects. *American Journal of Epidemiology* **107**, 149-160 (1978).

Gloeckner, G. *et al.* Identification and characterization of a new conjugation/type IVA secretion system (*trb/tra*) of *Legionella pneumophila* Corby localized on two mobile genomic islands. *International Journal of Medical Microbiology* **298**, 411-428 (2008).

Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 563-567 (1996).

Golubchik, T. *et al.* Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. *Nature Genetics* **44**, 352-

355 (2012).

Gomez-Alvarez, V., Revetta, R. P. & Santo Domingo, J. W. Metagenomic analyses of drinking water receiving different disinfection treatments. *Applied and Environmental Microbiology* **78**, 6095-6102 (2012).

Gomez-Valero, L. *et al.* Extensive recombination events and horizontal gene transfer shaped the *Legionella pneumophila* genomes. *BMC Genomics* **12** (2011).

Gomgnimbou, M. K. *et al.* Validation of a microbead-based format for spoligotyping of *Legionella pneumophila*. *Journal of Clinical Microbiology* **52**, 2410-2415 (2014).

Graham, R. M. A., Doyle, C. J. & Jennison, A. V. Real-time investigation of a *Legionella pneumophila* outbreak using whole genome sequencing. *Epidemiology and Infection* **142**, 2347-2351 (2014).

Graman, P. S., Quinlan, G. A. & Rank, J. A. Nosocomial legionellosis traced to a contaminated ice machine. *Infection Control and Hospital Epidemiology* **18**, 637-640 (1997).

Green, E. D. Strategies for the systematic sequencing of complex genomes. *Nature Reviews Genetics* **2**, 573-583 (2001).

Green, P. N. Efficacy of biocides on laboratory-generated *Legionella* biofilms. *Letters in Applied Microbiology* **17**, 158-161 (1993).

Griffin, A. T., Peyrani, P., Wiemken, T. & Arnold, F. Macrolides versus quinolones in *Legionella* pneumonia: results from the Community-Acquired Pneumonia Organization international study. *International Journal of Tuberculosis and Lung Disease* **14**, 495-499 (2010).

Guiguet, M. *et al.* Epidemiologic survey of a major outbreak of nosocomial legionellosis. *International Journal of Epidemiology* **16**, 466-471 (1987).

Habyarimana, F. *et al.* Role for the ankyrin eukaryotic-like genes of *Legionella pneumophila* in parasitism of protozoan hosts and human macrophages. *Environmental Microbiology* **10**, 1460-1474 (2008).

Haldane, D. J., Peppard, R. & Sumarah, R. K. Direct immunofluorescence for the diagnosis of legionellosis. *The Canadian Journal of Infectious Diseases* **4**, 101-104 (1993).

Haranaga, S. *et al.* Intravenous ciprofloxacin versus erythromycin in the treatment of *Legionella* pneumonia. *Internal Medicine* **46**, 352-356 (2007).

Harding, C. R. *et al.* The Dot/Icm effector SdhA is necessary for virulence of *Legionella pneumophila* in *Galleria mellonella* and A/J mice. *Infection and Immunity* **81**, 2598-2605 (2013).

Harris, S. R. *et al.* Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. *Nature Genetics* **44**, 413-419 (2012).

Harris, S. R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469-474 (2010).

Harrison, T. G., Afshar, B., Doshi, N., Fry, N. K. & Lee, J. V. Distribution of *Legionella pneumophila* serogroups, monoclonal antibody subgroups and DNA sequence types in recent clinical and environmental isolates from England and Wales (2000-2008). *European Journal of Clinical Microbiology & Infectious Diseases* **28**,

781-791 (2009).

Harrison, T. G. & Doshi, N. Evaluation of the Bartels *Legionella* urinary antigen enzyme immunoassay. *European Journal of Clinical Microbiology & Infectious Diseases* **20**, 738-740 (2001).

Harrison, T. G. & Taylor, A. G. *A Laboratory Manual for Legionella*. (Wiley & Sons Ltd, 1988).

Hatem, A., Bozdag, D., Toland, A. E. & Catalyuerek, U. V. Benchmarking short sequence mapping tools. *BMC Bioinformatics* **14** (2013).

Hay, J., Seal, D. V., Billcliffe, B. & Freer, J. H. Non-culturable *Legionella pneumophila* associated with *Acanthamoeba castellanii* - detection of the bacterium using DNA amplification and hybridization. *Journal of Applied Bacteriology* **78**, 61-65 (1995).

Health and Safety Executive (HSE*). Legionnaires' disease. The control of legionella bacteria in water systems* (2013).

Helbig, J. H., Luck, P. C., Knirel, Y. A., Witzleb, W. & Zahringer, U. Molecular characterization of a virulence-associated epitope on the lipopolysaccharide of *Legionella pneumophila* serogroup 1. *Epidemiology and Infection* **115**, 71-78 (1995).

Helbig, J. H. *et al.* Pan-European study on culture-proven Legionnaires' disease: Distribution of *Legionella pneumophila* serogroups and monoclonal subgroups. *European Journal of Clinical Microbiology & Infectious Diseases* **21**, 710-716 (2002).

Helbig, J. H., Kurtz, J. B., Pastoris, M. C., Pelaz, C. & Luck, P. C. Antigenic lipopolysaccharide components of *Legionella pneumophila* recognized by monoclonal antibodies: Possibilities and limitations for division of the species into serogroups. *Journal of Clinical Microbiology* **35**, 2841-2845 (1997).

Heller, R., Holler, C., Sussmuth, R. & Gundermann, K. O. Effect of salt concentration and temperature on survival of *Legionella pneumophila*. *Letters in Applied Microbiology* **26**, 64-68 (1998).

Hiller, N. L. *et al.* Generation of genic diversity among *Streptococcus pneumoniae* strains via horizontal gene transfer during a chronic polyclonal pediatric infection. *PLOS Pathogens* **6** (2010).

Hlady, W. G. *et al.* Outbreak of Legionnaires' disease linked to a decorative fountain by molecular epidemiology. *American Journal of Epidemiology* **138**, 555-562 (1993).

Holden, E. P., Winkler, H. H., Wood, D. O. & Leinbach, E. D. Intracellular growth of *Legionella pneumophila* within *Acanthamoeba castellani* Neff. *Infection and Immunity* **45**, 18-24 (1984).

Horwitz, M. A. Formation of a novel phagosome by the Legionnaires' disease bacterium (*Legionella pneumophila*) in human monocytes. *Journal of Experimental Medicine* **158**, 1319-1331 (1983).

Horwitz, M. A. & Maxfield, F. R. *Legionella pneumophila* inhibits acidification of its phagosome in human monocytes. *Journal of Cell Biology* **99**, 1936-1943 (1984).

Hubber, A. & Roy, C. R. Modulation of host cell function by *Legionella pneumophila* type

IV effectors. *Annual Review of Cell and Developmental Biology* **26**, 261-283 (2010).

Hunter, P. R. & Gaston, M. A. Numerical index of the discriminatory ability of typing systems - an application of Simpson's index of diversity *Journal of Clinical Microbiology* **26**, 2465-2466 (1988).

Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**, 254-267 (2006).

Hussong, D. *et al.* Viable *Legionella pneumophila* not detectable by culture on agar media. *Nature Biotechnology* **5**, 947-950 (1987).

Isberg, R. R., O'Connor, T. J. & Heidtman, M. The *Legionella pneumophila* replication vacuole: making a cosy niche inside host cells. *Nature Reviews Microbiology* **7**, 12-24 (2009).

Ito, I. *et al.* Hot spring bath and *Legionella pneumonia*: an association confirmed by genomic identification. *Internal Medicine* **41**, 859-863 (2002).

Jernigan, D. B. *et al.* Outbreak of Legionnaires' disease among cruise ship passengers exposed to a contaminated whirlpool spa. *Lancet* **347**, 494-499 (1996).

Johnson, J. D., Raff, M. J. & Vanarsdall, J. A. Neurologic manifestations of Legionnaires' disease. *Medicine* **63**, 303-310 (1984).

Jolley, K. A. *et al.* Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* **158**, 1005-1015 (2012).

Jolley, K. A. & Maiden, M. C. J. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11** (2010).

Joly, J. R. *et al.* Ecological distribution of Legionellaceae in the Quebec city area. *Canadian Journal of Microbiology* **30**, 63-67 (1984).

Joly, J. R. *et al.* Development of a standardised subgrouping scheme for *Legionella pneumophila* serogroup 1 using monoclonal antibodies. *Journal of Clinical Microbiology* **23**, 768-771 (1986).

Judge, K., Harris, S. R., Reuter, S., Parkhill, J. & Peacock, S. J. Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy* **70**, 2775-2778 (2015).

Karagiannis, I., Brandsema, P. & Van der Sande, M. Warm, wet weather associated with increased Legionnaires' disease incidence in The Netherlands. *Epidemiology and Infection* **137**, 181-187 (2009).

Kashuba, A. D. M. & Ballow, C. H. *Legionella* urinary antigen testing: Potential impact on diagnosis and antibiotic therapy. *Diagnostic Microbiology and Infectious Disease* **24**, 129-139 (1996).

Kaufmann, A. F. *et al.* Pontiac fever - Isolation of the etiologic agent (*Legionella pneumophila*) and demonstration of its mode of transmission. *American Journal of Epidemiology* **114**, 337-347 (1981).

Keller, D. W. *et al.* Community outbreak of Legionnaires' disease: An investigation confirming the potential for cooling towers to transmit *Legionella* species. *Clinical Infectious Diseases* **22**, 257-261 (1996).

Khan, M. A. *et al.* Comparative genomics reveal that host-innate immune responses influence the clinical prevalence of *Legionella pneumophila* serogroups. *PLOS*

ONE **8** (2013).

Khemiri, A. *et al.* Outer-membrane proteomic maps and surface-exposed proteins of *Legionella pneumophila* using cellular fractionation and fluorescent labelling. *Analytical and Bioanalytical Chemistry* **390**, 1861-1871 (2008).

Kim, B. R., Anderson, J. E., Mueller, S. A., Gaines, W. A. & Kendall, A. M. Literature review - efficacy of various disinfectants against *Legionella* in water systems. *Water Research* **36**, 4433-4444 (2002).

Koeser, C. U. *et al.* Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *New England Journal of Medicine* **366**, 2267-2275 (2012).

Kohl, T. A. *et al.* Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach. *Journal of Clinical Microbiology* **52**, 2479-2486 (2014).

Kohler, R. B., Winn, W. C. & Wheat, L. J. Onset and duration of urinary antigen excretion in Legionnaires' disease. *Journal of Clinical Microbiology* **20**, 605-607 (1984).

Kong, Y. *et al.* Homologous recombination drives both sequence diversity and gene content variation in *Neisseria meningitidis*. *Genome Biology and Evolution* **5**, 1611-1627 (2013).

Kool, J. L., Carpenter, J. C. & Fields, B. S. Effect of monochloramine disinfection of municipal drinking water on risk of nosocomial Legionnaires' disease. *Lancet* **353**, 272-277 (1999).

Kool, J. L. *et al.* More than 10 years of unrecognized nosocomial transmission of Legionnaires' disease among transplant patients. *Infection Control and Hospital Epidemiology* **19**, 898-904 (1998).

Koren, S. *et al.* Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology* **14** (2013).

Koren, S. & Phillippy, A. M. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology* **23**, 110-120 (2015).

Kozak, N. A. *et al.* Distribution of *lag-1* alleles and sequence-based types among *Legionella pneumophila* serogroup 1 clinical and environmental isolates in the United States. *Journal of Clinical Microbiology* **47**, 2525-2535 (2009).

Kozak-Muiznieks, N. *et al.* Prevalence of sequence types among clinical and environmental isolates of *Legionella pneumophila* serogroup 1 in the United States from 1982 to 2012. *Journal of Clinical Microbiology* **52**, 201-211 (2014).

Kubori, T., Hyakutake, A. & Nagai, H. *Legionella* translocates an E3 ubiquitin ligase that has multiple U-boxes with distinct functions. *Molecular Microbiology* **67**, 1307-1319 (2008).

Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biology* **5** (2004).

Kwong, J. C., McCallum, N., Sintchenko, V. & Howden, B. P. Whole genome sequencing in clinical and public health microbiology. *Pathology* **47**, 199-210 (2015).

Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187-197 (2011).

Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**,

860-921 (2001).

Langille, M. G. I. & Brinkman, F. S. L. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* **25**, 664-665 (2009).

Laver, T. *et al.* Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification* **3**, 1-8 (2015).

Lawrence, C. *et al.* Single clonal origin of a high proportion of *Legionella pneumophila* serogroup 1 isolates from patients and the environment in the area of Paris, France, over a 10-year period. *Journal of Clinical Microbiology* **37**, 2652-2655 (1999).

Lechat, P., Souche, E. & Moszer, I. SynTView - an interactive multi-view genome browser for next-generation comparative microorganism genomics. *BMC Bioinformatics* **14** (2013).

Leekitcharoenphon, P., Nielsen, E. M., Kaas, R. S., Lund, O. & Aarestrup, F. M. Evaluation of whole genome sequencing for outbreak detection of *Salmonella enterica*. *PLOS ONE* **9** (2014).

Leoni, E. *et al. Legionella* waterline colonization: detection of *Legionella* species in domestic, hotel and hospital hot water systems. *Journal of Applied Microbiology* **98**, 373-379 (2005).

Leopold, S. R., Goering, R. V., Witten, A., Harmsen, D. & Mellmann, A. Bacterial whole-genome sequencing revisited: Portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. *Journal of Clinical Microbiology* **52**, 2365-2370 (2014).

Lepine, L. A. *et al.* A recurrent outbreak of nosocomial Legionnaires' disease detected by urinary antigen testing: Evidence for long-term colonization of a hospital plumbing system. *Infection Control and Hospital Epidemiology* **19**, 905-910 (1998).

Lettinga, K. D. *et al.* Legionnaires' disease at a Dutch flower show: Prognostic factors and impact of therapy. *Emerging Infectious Diseases* **8**, 1448-1454 (2002).

Levesque, S. *et al.* Genomic characterization of a large outbreak of *Legionella pneumophila* serogroup 1 strains in Quebec City, 2012. *PLOS ONE* **9** (2014).

Lewis, T. *et al.* High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak. *Journal of Hospital Infection* **75**, 37-41 (2010).

Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

Li, L., Mendis, N., Trigui, H. & Faucher, S. P. Transcriptomic changes of *Legionella pneumophila* in water. *BMC Genomics* **16** (2015).

Lim, W. S. *et al.* Study of community acquired pneumonia aetiology (SCAPA) in adults admitted to hospital: implications for management guidelines. *Thorax* **56**, 296-301 (2001).

Lin, Y. E., Lu, W. M., Huang, H. I. & Huang, W. K. Environmental survey of *Legionella*

pneumophila in hot springs in Taiwan. *Journal of Toxicology and Environmental Health* **70**, 84-87 (2007).

Lin, Y. E., Stout, J. E. & Yu, V. L. Controlling *Legionella* in hospital drinking water: An evidence-based review of disinfection methods. *Infection Control and Hospital Epidemiology* **32**, 166-173 (2011).

Lin, Y. E., Stout, J. E. & Yu, V. L. Prevention of hospital-acquired legionellosis. *Current Opinion in Infectious Diseases* **24**, 350-356 (2011).

Liu, L. *et al.* Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology* (2012).

Liu, M. *et al.* The *Legionella pneumophila* EnhC protein interferes with immunostimulatory muramyl peptide production to evade innate immunity. *Cell Host & Microbe* **12**, 166-176 (2012).

Liu, M. F. & Taylor, D. E. Characterization of Gram-positive tellurite resistance encoded by the *Streptococcus pneumonia tehB* gene. *FEMS Microbiology Letters* **174**, 385-392 (1999).

Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nature Methods* **12**, 733-735 (2015).

Lowry, P. W., Blankenship, R. J., Gridley, W., Troup, N. J. & Tompkins, L. S. A cluster of *Legionella* sternal wound infections due to postoperative topical exposure to contaminated tap water. *New England Journal of Medicine* **324**, 109-113 (1991).

Lowry, P. W. & Tompkins, L. S. Nosocomial legionellosis - A review of pulmonary and extrapulmonary syndromes. *American Journal of Infection Control* **21**, 21-27 (1993).

LPSN. List of prokaryotic names with standing in nomenclature: Genus *Legionella*. Available at: http://www.bacterio.net/legionella.html

Lück, C., Fry, N. K., Helbig, J. H., Jarraud, S. & Harrison, T. G. Typing methods for *Legionella* in *Legionella: Methods and Protocols* (ed. Buchrieser, C. & Hilbi, H.) 119-148 (Humana Press, 2013).

Lück, P. C., Bender, L., Ott, M., Helbig, J. H. & Hacker, J. Analysis of *Legionella pneumophila* serogroup 6 strains isolated from a hospital warm water supply over a 3-year period by using genomic long-range mapping techniques and monoclonal antibodies. *Applied and Environmental Microbiology* **57**, 3226-3231 (1991).

Lück, P. C. *et al.* Epidemiologic investigation by macrorestriction analysis and by using monoclonal antibodies of nosocomial pneumonia caused by *Legionella pneumophila* serogroup 10. *Journal of Clinical Microbiology* **32**, 2692-2697 (1994).

Lück, P. C., Kohler, J., Maiwald, M. & Helbig, J. H. DNA polymorphisms in strains of *Legionella pneumophila* serogroup 3 and serogroup 4 detected by macrorestriction analysis and their use for epidemiologic investigation of nosocomial legionellosis. *Applied and Environmental Microbiology* **61**, 2000-2003 (1995).

Lueneberg, E. *et al.* Cloning and functional characterization of a 30kb gene locus required for lipopolysaccharide biosynthesis in *Legionella pneumophila*.

International Journal of Medical Microbiology **290**, 37-49 (2000).

Luo, T. *et al.* Whole-genome sequencing to detect recent transmission of *Mycobacterium tuberculosis* in settings with a high burden of tuberculosis. *Tuberculosis* **94**, 434-440 (2014).

Lurie-Weinberger, M. N. *et al.* The origins of eukaryotic-like proteins in *Legionella pneumophila*. *International Journal of Medical Microbiology* **300**, 470-481 (2010).

Ma, J., He, Y., Hu, B., Luo, Z-Q. Genome sequence of an environmental isolate of the bacterial pathogen *Legionella pneumophila*. *Genome Announcements* **1** (2013).

Maiwald, M., Helbig, J. H. & Luck, P. C. Laboratory methods for the diagnosis of *Legionella* infections. *Journal of Microbiological Methods* **33**, 59-79 (1998).

Majewski, J. & Cohan, F. M. The effect of mismatch repair and heteroduplex formation on sexual isolation in *Bacillus*. *Genetics* **148**, 13-18 (1998).

Mandell, L. A. *et al.* Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. *Clinical Infectious Diseases* **44**, S27-S72 (2007).

Marchesi, I. *et al.* Effectiveness of different methods to control *Legionella* in the water supply: ten-year experience in an Italian university hospital. *Journal of Hospital Infection* **77**, 47-51 (2010).

Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380 (2005).

Marrao, G., Verissimo, A., Bowker, R. G. & Dacosta, M. S. Biofilms as major sources of *Legionella spp* in hydrothermal areas and their dispersion into stream water. *FEMS Microbiology Ecology* **12**, 25-33 (1993).

Marston, B. J., Lipman, H. B. & Breiman, R. F. Surveillance for Legionnaires' disease - risk factors for morbidity and mortality. *Archives of Internal Medicine* **154**, 2417-2422 (1994).

Marttinen, P. *et al.* Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Research* **40** (2012).

Matsui, M. *et al.* Isolation of *Legionella rubrilucens* from a pneumonia patient co-infected with *Legionella pneumophila*. *Journal of Medical Microbiology* **59**, 1242-1246 (2010).

Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 560-564 (1977).

McAdam, P. R. *et al.* Gene flow in environmental *Legionella pneumophila* leads to genetic and pathogenic heterogeneity within a Legionnaires' disease outbreak. *Genome Biology* **15** (2014).

McDade, J. E. *et al.* Legionnaires' disease - Isolation of a bacterium and demonstration of its role in other respiratory disease. *New England Journal of Medicine* **297**, 1197-1203 (1977).

Mellmann, A. *et al.* Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLOS ONE* **6** (2011).

Mentasti, M. *et al.* Application of *Legionella pneumophila*-specific quantitative real-time

PCR combined with direct amplification and sequence-based typing in the diagnosis and epidemiological investigation of Legionnaires' disease. *European Journal of Clinical Microbiology & Infectious Diseases* **31**, 2017-2028 (2012).

Mentasti, M. *et al.* Extension of the *Legionella pneumophila* sequence-based typing scheme to include strains carrying a variant of the *N-acylneuraminate cytidylyltransferase* gene. *Clinical Microbiology and Infection* **20**, O435-O441 (2014).

Mentasti, M. & Fry, N. K. Nested sequence-based typing (SBT) protocol for epidemiological typing of *Legionella pneumophila* directly from clinical samples. Available at: http://bioinformatics.phe.org.uk/legionella/legionella_sbt/php/protocols/ESGLI%20NESTED%20SBT%20GUIDELINE%20v2.0.pdf

Mercante, J. W. & Winchell, J. M. Current and emerging *Legionella* diagnostics for laboratory and outbreak investigations. *Clinical Microbiology Reviews* **28**, 95-133 (2015).

Michod, R. E., Bernstein, H. & Nedelcu, A. M. Adaptive value of sex in microbial pathogens. *Infection Genetics and Evolution* **8**, 267-285 (2008).

Mintz, C. S. & Shuman, H. A. Transposition of bacteriophage mu in the Legionnaires' disease bacterium. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 4645-4649 (1987).

Molmeret, M. & Abu Kwaik, Y. How does *Legionella pneumophila* exit the host cell? *Trends in Microbiology* **10**, 258-260 (2002).

Molmeret, M., Horn, M., Wagner, M., Santic, M. & Abu Kwaik, Y. Amoebae as training grounds for intracellular bacterial pathogens. *Applied and Environmental Microbiology* **71**, 20-28 (2005).

Molofsky, A. B. & Swanson, M. S. Differentiate to thrive: lessons from the *Legionella pneumophila* life cycle. *Molecular Microbiology* **53**, 29-40 (2004).

Moran-Gilad, J. *et al.* Molecular epidemiology of Legionnaires' disease in Israel. *Clinical Microbiology and Infection* **20**, 690-696 (2014).

Moran-Gilad, J. *et al.* Design and application of a core genome multilocus sequence typing scheme for investigation of Legionnaires' disease incidents. *Eurosurveillance* **20** (2015).

Mouchtouri, V. A., Goutziana, G., Kremastinou, J. & Hadjichristodoulou, C. *Legionella* species colonization in cooling towers: Risk factors and assessment of control measures. *American Journal of Infection Control* **38**, 50-55 (2010).

Muder, R. R. & Yu, V. L. Infection due to *Legionella* species other than *L. pneumophila*. *Clinical Infectious Diseases* **35**, 990-998 (2002).

Muder, R. R., Yu, V. L. & Woo, A. H. Mode of transmission of *Legionella pneumophila* - A critical review. *Archives of Internal Medicine* **146**, 1607-1612 (1986).

Munro, R., Neville, S., Daley, D. & Mercer, J. Microbiological aspects of an outbreak of Legionnaires' disease in South Western Sydney. *Pathology* **26**, 48-51 (1994).

Muraca, P. W., Yu, V. L. & Goetz, A. Disinfection of water distribution systems for *Legionella* - a review of application procedures and methodologies. *Infection Control and Hospital Epidemiology* **11**, 79-88 (1990).

Murdoch, D. R. Diagnosis of *Legionella* infection. *Clinical Infectious Diseases* **36**, 64-69 (2003).

Murga, R. *et al.* Role of biofilms in the survival of *Legionella pneumophila* in a model potable-water system. *Microbiology* **147**, 3121-3126 (2001).

Musso, D. & Raoult, D. Serological cross-reactions between *Coxiella burnetii* and *Legionella micdadei*. *Clinical and Diagnostic Laboratory Immunology* **4**, 208-212 (1997).

Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462-465 (2011).

Mykietiuk, A. *et al.* Clinical outcomes for hospitalized patients with *Legionella pneumonia* in the antigenuria era: The influence of Levofloxacin therapy. *Clinical Infectious Diseases* **40**, 794-799 (2005).

Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research* **39** (2011).

Narra, H. P. & Ochman, H. Of what use is sex to bacteria? *Current Biology* **16**, R705-R710 (2006).

Naylor, J. & Cianciotto, N. P. Cytochrome c maturation proteins are critical for *in vivo* growth of *Legionella pneumophila*. *FEMS Microbiology Letters* **241**, 249-256 (2004).

Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 5269-5273 (1979).

Nelson, D. P., Rensimer, E. R. & Raffin, T. A. *Legionella pneumophila* pericarditis without pneumonia. *Archives of Internal Medicine* **145**, 926-926 (1985).

Newton, H. J., Ang, D. K. Y., van Driel, I. R. & Hartland, E. L. Molecular pathogenesis of infections caused by *Legionella pneumophila*. *Clinical Microbiology Reviews* **23**, 274-298 (2010).

Ng, V. *et al.* Going with the flow: Legionellosis risk in Toronto, Canada is strongly associated with local watershed hydrology. *Ecohealth* **5**, 482-490 (2008).

Nguyen, T. M. N. *et al.* A community-wide outbreak of Legionnaires' disease linked to industrial cooling towers - How far can contaminated aerosols spread? *Journal of Infectious Diseases* **193**, 102-111 (2006).

Nygard, K. *et al.* An outbreak of Legionnaires' disease caused by long-distance spread from an industrial air scrubber in Sarpsborg, Norway. *Clinical Infectious Diseases* **46**, 61-69 (2008).

O'Loughlin, R. E. *et al.* Restaurant outbreak of Legionnaires' disease associated with a decorative fountain: an environmental and case-control study. *BMC Infectious Diseases* **7** (2007).

Orsi, G. B. *et al. Legionella* control in the water system of antiquated hospital buildings by shock and continuous hyperchlorination: 5 years experience. *BMC Infectious Diseases* **14** (2014).

Ortiz-Roque, C. M. & Hazen, T. C. Abundance and distribution of Legionellaceae in Puerto-Rican waters. *Applied and Environmental Microbiology* **53**, 2231-2236 (1987).

Osawa, K. *et al.* A case of nosocomial *Legionella pneumonia* associated with a contaminated hospital cooling tower. *Journal of Infection and Chemotherapy* **20**, 68-70 (2014).

Osterholm, M. T. *et al.* A 1957 outbreak of Legionnaires' disease associated with a meat packing plant. *American Journal of Epidemiology* **117**, 60-67 (1983).

Ott, M., Bender, L., Marre, R. & Hacker, J. Pulsed field electrophoresis of genomic restriction fragments for the detection of nosocomial *Legionella pneumophila* in hospital water supplies. *Journal of Clinical Microbiology* **29**, 813-815 (1991).

Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693 (2015).

Palmer, C. J., Tsai, Y. L., Paszkokolva, C., Mayer, C. & Sangermano, L. R. Detection of *Legionella* species in sewage and ocean water by polymerase chain reaction, direct fluorescent antibody, and plate culture methods. *Applied and Environmental Microbiology* **59**, 3618-3624 (1993).

Palmore, T. N. *et al.* A cluster of cases of nosocomial Legionnaires' disease linked to a contaminated hospital decorative water fountain. *Infection Control and Hospital Epidemiology* **30**, 764-768 (2009).

Pan, X., Luhrmann, A., Satoh, A., Laskowski-Arce, M. A. & Roy, C. R. Ankyrin repeat proteins comprise a diverse family of bacterial type IV effectors. *Science* **320**, 1651-1654 (2008).

Pancer, K., Matuszewska, R., Bartosik, M., Kacperski, K. & Krogulska, B. Persistent colonization of 2 hospital water supplies by *L. pneumophila* strains through 7 years - Sequence-based typing and serotyping as useful tools for complex risk analysis. *Annals of Agricultural and Environmental Medicine* **20**, 687-694 (2013).

Pastoris, M. C. *et al.* Legionnaires' disease on a cruise ship linked to the water supply system: Clinical and public health implications. *Clinical Infectious Diseases* **28**, 33-38 (1999).

Pastoris, M. C., Passi, C. & Maroli, M. Evidence of *Legionella pneumophila* in some arthropods and related natural aquatic habitats. *FEMS Microbiology Ecology* **62**, 259-263 (1989).

Perez-Losada, M. *et al.* Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infection Genetics and Evolution* **6**, 97-112 (2006).

Perola, O. *et al.* Persistent *Legionella pneumophila* colonization of a hospital water supply: efficacy of control methods and a molecular epidemiological analysis. *APMIS* **113**, 45-53 (2005).

Petzold, M. *et al.* A structural comparison of lipopolysaccharide biosynthesis loci of *Legionella pneumophila* serogroup 1 strains. *BMC Microbiology* **13** (2013).

PHE. Health Technical Memorandum 04-01 Safe water in healthcare premises. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/524880/DH_HTM_0401_PART_A_acc.pdf

Phin, N. *et al.* Epidemiology and clinical management of Legionnaires' disease. *Lancet Infectious Diseases* **14**, 1011-1021 (2014).

Pop, M. & Salzberg, S. L. Bioinformatics challenges of new sequencing technology. *Trends in Genetics* **24**, 142-149 (2008).

Ponstingl, H. SMALT. Available at: http://www.sanger.ac.uk/science/tools/smalt-0

Quail, M. A. *et al.* Optimal enzymes for amplifying sequencing libraries. *Nature Methods* **9**, 10-11 (2012).

Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13** (2012).

Quaranta, G. *et al. Legionella* on board trains: effectiveness of environmental surveillance and decontamination. *BMC Public Health* **12** (2012).

Rambaut, A. FigTree. Available at: http://tree.bio.ed.ac.uk/software/figtree/

Rambaut, A. Tracer. Available at: http://tree.bio.ed.ac.uk/software/tracer

Rambaut, A., Lam, T. T., Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, doi:10.1093/ve/vew007 (2016).

Ramirez, J. A. & Summersgill, J. T. Rapid tests for the diagnosis of *Legionella* infections. *Journal of the Kentucky Medical Association* **92**, 62-65 (1994).

Rangel-Frausto, M. S. *et al.* Persistence of *Legionella pneumophila* in a hospital's water system: A 13-year survey. *Infection Control and Hospital Epidemiology* **20**, 793-797 (1999).

Ratzow, S., Gaia, V., Helbig, J. H., Fry, N. K. & Lueck, P. C. Addition of *neuA*, the gene encoding N-acylneuraminate cytidylyl transferase, increases the discriminatory ability of the consensus sequence-based scheme for typing *Legionella pneumophila* serogroup 1 strains. *Journal of Clinical Microbiology* **45**, 1965-1968 (2007).

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing (2013).

Redfield, R. J. Genes for breakfast - the have-your-cake-and-eat-it-too of bacterial transformation. *Journal of Heredity* **84**, 400-404 (1993).

Reischl, U. *et al.* Direct detection and differentiation of *Legionella spp*. and *Legionella pneumophila* in clinical specimens by dual-color real-time PCR and melting curve analysis. *Journal of Clinical Microbiology* **40**, 3814-3817 (2002).

Reuter, S. *et al.* A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak. *BMJ Open* **3** (2013).

Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* **5**, 278-289 (2015).

Ricketts, K. D. *et al.* Weather patterns and Legionnaires' disease: a meteorological study. *Epidemiology and Infection* **137**, 1003-1012 (2009).

Rodgers, F. G. Ultrastructure of *Legionella pneumophila*. *Journal of Clinical Pathology* **32**, 1195-1202 (1979).

Rodgers, F. G., Greaves, P. W., Macrae, A. D. & Lewis, M. J. Electron microscopic evidence of flagella and pili on *Legionella pneumophila*. *Journal of Clinical Pathology* **33**,

1184-1188 (1980).

Rosmini, F. *et al.* Febrile illness in successive cohorts of tourists at a hotel on the Italian Adriatic coast - evidence for a persistent focus of *Legionella* infection. *American Journal of Epidemiology* **119**, 124-134 (1984).

Rota, M. C., Caporali, M. G. & Massari, M. European guidelines for control and prevention of travel-associated Legionnaires' disease: the Italian experience. *Eurosurveillance* **9** (2004).

Rowbotham, T. J. Preliminary report on the pathogenicity of *Legionella pneumophila* for freshwater and soil amoebas. *Journal of Clinical Pathology* **33**, 1179-1183 (1980).

Rowbotham, T. J. Current views on the relationships between amoebas, legionellae and man. *Israel Journal of Medical Sciences* **22**, 678-689 (1986).

Rowbotham, T. J. Isolation of *Legionella pneumophila* serogroup 1 from human feces with use of amebic cocultures. *Clinical Infectious Diseases* **26**, 502-503 (1998).

Roy, T. M., Fleming, D. & Anderson, W. H. Tularemic pneumonia mimicking Legionnaires' disease with false positive direct fluorescent antibody stains for *Legionella*. *Southern Medical Journal* **82**, 1429-1431 (1989).

Ruffalo, M., LaFramboise, T. & Koyutuerk, M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* **27**, 2790-2796 (2011).

Rusin, P. A., Rose, J. B., Haas, C. N. & Gerba, C. P. Risk assessment of opportunistic bacterial pathogens in drinking water. *Reviews of Environmental Contamination and Toxicology* **152**, 57-83 (1997).

Sabria, M. *et al.* A community outbreak of Legionnaires' disease: evidence of a cooling tower as the source. *Clinical Microbiology and Infection* **12**, 642-647 (2006).

Sahr, T. *et al.* Deep sequencing defines the transcriptional map of *L. pneumophila* and identifies growth phase-dependent regulated ncRNAs implicated in virulence. *RNA Biology* **9**, 503-519 (2012).

Salipante, S. J. *et al.* Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology. *Journal of Clinical Microbiology* **53**, 1072-1079 (2015).

Sánchez-Busó, L., Comas, I., Jorques, G. & Gonzalez-Candelas, F. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. *Nature Genetics* **46**, 1205-1211 (2014).

Sánchez-Busó, L. *et al.* Genomic investigation of a legionellosis outbreak in a persistently colonized hotel. *Frontiers in Microbiology* **6** (2016).

Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463-5467 (1977).

Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94**, 441-448 (1975).

Sauer, J. D., Bachman, M. A. & Swanson, M. S. The phagosomal transporter A couples threonine acquisition to differentiation and replication of *Legionella*

pneumophila in macrophages. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 9924-9929 (2005).

Schadt, E. E., Turner, S. & Kasarskis, A. A window into third generation sequencing. *Human Molecular Genetics* **20**, 853-853 (2010).

Schaefer, U. KmerID. Available at: https://github.com/phe-bioinformatics/kmerid

Schroeder, G. N. *et al. Legionella pneumophila* strain 130b possesses a unique combination of type IV secretion systems and novel Dot/Icm secretion system effector proteins. *Journal of Bacteriology* **192**, 6001-6016 (2010).

Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).

Selander, R. K. *et al.* Genetic structure of populations of *Legionella pneumophila*. *Journal of Bacteriology* **163**, 1021-1037 (1985).

Shands, K. N. *et al.* Potable water as a source of Legionnaires' disease. *Jama-Journal of the American Medical Association* **253**, 1412-1416 (1985).

Shang, J. *et al.* Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *BioMed Research International* (2014).

Shelburne, S. A., Kielhofner, M. A. & Tiwari, P. S. Cerebellar involvement in legionellosis. *Southern Medical Journal* **97**, 61-64 (2004).

Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature Biotechnology* **26**, 1135-1145 (2008).

Silk, B. J. *et al.* Eight years of Legionnaires' disease transmission in travellers to a condominium complex in Las Vegas, Nevada. *Epidemiology and Infection* **140**, 1993-2002 (2012).

Smalley, D. L., Jaquess, P. A., Ourth, D. D. & Layne, J. S. Antibiotic-induced filament formation of *Legionella pneumophila*. *American Journal of Clinical Pathology* **74**, 852-852 (1980).

Snipen, L. & Liland, K. H. Micropan: an R-package for microbial pan-genomics. *BMC Bioinformatics* **16** (2015).

Stamatakis, A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).

Stanek, G. *et al.* Indirect immunofluorescence assay (IFA), microagglutination test (MA) and enzyme-linked immunosorbent assay (ELISA) in diagnosis of legionellosis. *Medical Microbiology Infectious Diseases Virology Parasitology* **255**, 108-114 (1983).

Steinert, M., Emody, L., Amann, R. & Hacker, J. Resuscitation of viable but nonculturable *Legionella pneumophila* Philadelphia JR32 by *Acanthamoeba castellanii*. *Applied and Environmental Microbiology* **63**, 2047-2053 (1997).

Stone, B. J. & Abu Kwaik, Y. Natural competence for DNA transformation by *Legionella pneumophila* and its association with expression of type IV pili. *Journal of Bacteriology* **181**, 1395-1402 (1999).

Stout, J. E. & Yu, V. L. Legionellosis. *New England Journal of Medicine* **337**, 682-687 (1997).

Stout, J. E., Yu, V. L. & Best, M. G. Ecology of *Legionella pneumophila* within water distribution systems. *Applied and Environmental Microbiology* **49**, 221-228

(1985).

Straus, W. L. *et al.* Risk factors for domestic acquisition of Legionnaires' disease. *Archives of Internal Medicine* **156**, 1685-1692 (1996).

Sturgill-Koszycki, S. & Swanson, M. S. *Legionella pneumophila* replication vacuoles mature into acidic, endocytic organelles. *Journal of Experimental Medicine* **192**, 1261-1272 (2000).

Taylor, D. E. Bacterial tellurite resistance. *Trends in Microbiology* **7**, 111-115 (1999).

Tijet, N. *et al.* New endemic *Legionella pneumophila* serogroup I clones, Ontario, Canada. *Emerging Infectious Diseases* **16**, 447-454 (2010).

Tilney, L. G., Harb, O. S., Connelly, P. S., Robinson, C. G. & Roy, C. R. How the parasitic bacterium *Legionella pneumophila* modifies its phagosome and transforms it into rough ER: implications for conversion of plasma membrane to the ER membrane. *Journal of Cell Science* **114**, 4637-4650 (2001).

Tommassen, J. Assembly of outer-membrane proteins in bacteria and mitochondria. *Microbiology* **156**, 2587-2596 (2010).

Tompkins, L. S., Roessler, B. J., Redd, S. C., Markowitz, L. E. & Cohen, M. L. *Legionella* prosthetic valve endocarditis. *New England Journal of Medicine* **318**, 530-535 (1988).

Tossa, P., Deloge-Abarkan, M., Zmirou-Navier, D., Hartemann, P. & Mathieu, L. Pontiac fever: an operational definition for epidemiological studies. *BMC Public Health* **6** (2006).

Tripp, S. & Grueber, M. Economic Impact of the Human Genome Project. Battelle Memorial Institute. (2011).

Tsai, T. F. *et al.* Legionnaires' disease - clinical features of the epidemic in Philadelphia. *Annals of Internal Medicine* **90**, 509-517 (1979).

Underwood, A. P., Jones, G., Mentasti, M., Fry, N. K. & Harrison, T. G. Comparison of the *Legionella pneumophila* population structure as determined by sequence-based typing and whole genome sequencing. *BMC Microbiology* **13** (2013).

Valouev, A. *et al.* A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Research* **18**, 1051-1063 (2008).

Valsangiacomo, C. *et al.* Use of amplified fragment length polymorphism in molecular typing of *Legionella pneumophila* and application to epidemiologic studies. *Journal of Clinical Microbiology* **33**, 1716-1719 (1995).

van Belkum, A. *et al.* Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clinical Microbiology and Infection* **13**, 1-46 (2007).

van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends in Genetics* **30**, 418-426 (2014).

van Heijnsbergen, E. *et al.* Viable *Legionella pneumophila* bacteria in natural soil and rainwater puddles. *Journal of Applied Microbiology* **117**, 882-890 (2014).

Vekens, E. *et al.* Sequence-based typing of *Legionella pneumophila* serogroup 1 clinical isolates from Belgium between 2000 and 2010. *Eurosurveillance* **17**, 9-14 (2012).

Venezia, R. A., Agresta, M. D., Hanley, E. M., Urquhart, K. & Schoonmaker, D. Nosocomial legionellosis associated with aspiration of nasogastric feedings diluted in tap water. *Infection Control and Hospital Epidemiology* **15**, 529-533 (1994).

Verissimo, A., Marrao, G., Dasilva, F. G. & Dacosta, M. S. Distribution of *Legionella spp* in hydrothermal areas in continental Portugal and the island of Sao-Miguel, Azores. *Applied and Environmental Microbiology* **57**, 2921-2927 (1991).

Vernikos, G. S. & Parkhill, J. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* **22**, 2196-2203 (2006).

Viswanathan, V. K. *et al.* The cytochrome c maturation locus of *Legionella pneumophila* promotes iron assimilation and intracellular infection and contains a strain-specific insertion sequence element. *Infection and Immunity* **70**, 1842-1852 (2002).

Vogel, J. P. & Isberg, R. R. Cell biology of *Legionella pneumophila*. *Current Opinion in Microbiology* **2**, 30-34 (1999).

von Baum, H. *et al.* Community-acquired *Legionella pneumonia*: new insights from the German Competence Network for Community Acquired Pneumonia. *Clinical Infectious Diseases* **46**, 1356-1364 (2008).

Vos, M. Why do bacteria engage in homologous recombination? *Trends in Microbiology* **17**, 226-232 (2009).

Wadowsky, R. M., Wolford, R., McNamara, A. M. & Yee, R. B. Effect of temperature, pH, and oxygen level on the multiplication of naturally-occurring *Legionella pneumophila* in potable water. *Applied and Environmental Microbiology* **49**, 1197-1205 (1985).

Wadowsky, R. M. & Yee, R. B. Glycine-containing selective medium for isolation of Legionellaceae from environmental specimens. *Applied and Environmental Microbiology* **42**, 768-772 (1981).

Wallis, L. & Robinson, P. Soil as a source of *Legionella pneumophila* serogroup 1 (Lp1). *Australian and New Zealand Journal of Public Health* **29**, 518-520 (2005).

Weissenmayer, B. A., Prendergast, J. G. D., Lohan, A. J. & Loftus, B. J. Sequencing illustrates the transcriptional response of *Legionella pneumophila* during infection and identifies seventy novel small non-coding RNAs. *PLOS ONE* **6** (2011).

Whiley, H. & Bentham, R. *Legionella longbeachae* and legionellosis. *Emerging Infectious Diseases* **17**, 579-583 (2011).

Wilkinson, H. W., Farshy, C. E., Fikes, B. J., Cruce, D. D. & Yealy, L. P. Measure of immunoglobulin G-specific, M-specific and A-specific titers against *Legionella pneumophila* and inhibition of titers against nonspecific, Gram-negative bacterial antigens in the indirect immunofluorescence test for legionellosis. *Journal of Clinical Microbiology* **10**, 685-689 (1979).

Xu, L. & Luo, Z. Q. Cell biology of infection by *Legionella pneumophila*. *Microbes and Infection* **15**, 157-167 (2013).

Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).

Yiallouros, P. K. *et al.* First outbreak of nosocomial *Legionella* infection in term neonates caused by a cold mist ultrasonic humidifier. *Clinical Infectious Diseases* **57**, 48-56 (2013).

Yu, V. L. Could aspiration be the major mode of transmission for *Legionella*? *American Journal of Medicine* **95**, 13-15 (1993).

Yu, V. L., Liu, Z. M., Stout, J. E. & Goetz, A. *Legionella* disinfection of water distribution systems - principles, problems and practice. *Infection Control and Hospital Epidemiology* **14**, 567-570 (1993).

Yzerman, E. P. F. *et al.* Sensitivity of three urinary antigen tests associated with clinical severity in a large outbreak of Legionnaires' disease in the Netherlands. *Journal of Clinical Microbiology* **40**, 3232-3236 (2002).

Zacheus, O. M. & Martikainen, P. J. Occurrence of legionellae in hot water distribution systems of Finnish apartment buildings. *Canadian Journal of Microbiology* **40**, 993-999 (1994).

Zacheus, O. M. & Martikainen, P. J. Effect of heat flushing on the concentrations of *Legionella pneumophila* and other heterotrophic microbes in hot water systems of apartment buildings. *Canadian Journal of Microbiology* **42**, 811-818 (1996).

Zerbino, D. R. & Birney, E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research* **18**, 821-829 (2008).

Zumla, A., Weyell, R. & Tettmar, R. E. Legionnaires' disease – early lessons from 1988 London outbreak. *Lancet* **1**, 1275 (1988).

9. Appendix

9.1 Chapter 3

Appendix Table 1. 32 previously published genomes of *L. pneumophila* that represent the known species diversity. ST - sequence type; Sg - serogroup; clin - clinical; env - environmental; U/k - unknown

Isolate name	ST	Sg	Source	Country	Year	Reference	Known epidemiologic relatedness
Alcoy	578	1	clin	Spain	1999	D'Auria *et al.* (2010)	None
Corby	51	1	clin	UK	1982	Gloeckner *et al.* (2007)	None
Lorraine/ ST47_1	47	1	clin	France	2004	Gomez-Valero *et al.* (2011)	None
Philadelphia-1 (ATCC 33152)	36	1	clin	USA	1981	Chien *et al.* (2004)	None
Wadsworth 130b	42	1	clin	USA	U/k	Schroeder *et al.* (2010)	None
LC6774	154	1	env	UK	2003	Underwood *et al.* (2013)	None
H093380153	179	1	clin	UK	2009	Underwood *et al.* (2013)	None
H044500045	186	1	clin	UK	2004	Underwood *et al.* (2013)	None
H075160080	188	1	env	UK	2007	Underwood *et al.* (2013)	None
H063280001 /ST23_1	23	1	clin	UK	2006	Underwood *et al.* (2013)	None
Lansing-3	336	15	clin	USA	1981	Underwood *et al.* (2013)	None
RR08000517	337	4	env	UK	2007	Underwood *et al.* (2013)	None
RR08000134	34	1	env	UK	2005	Underwood *et al.* (2013)	None
RR08000760	376	4	env	UK	2006	Underwood *et al.* (2013)	None
H100260089	44	1	clin	UK	2010	Underwood *et al.* (2013)	None
H091960011	454	1	env	UK	2009	Underwood *et al.* (2013)	None
H093620212	46	1	clin	UK	2009	Underwood *et al.* (2013)	None
H065000139	54	1	clin	UK	2006	Underwood *et al.* (2013)	None
H070840415	59	1	clin	UK	2007	Underwood *et al.* (2013)	None
H090500162	611	1	env	UK	2009	Underwood *et al.*	None

						(2013)	
H064180002 /ST62_1	62	1	clin	UK	2006	Underwood *et al.* (2013)	Related to ST62_19
H074360710	68	6	env	UK	2007	Underwood *et al.* (2013)	None
H091960009	707	4	env	UK	2009	Underwood *et al.* (2013)	None
LC6451	78	1	clin	UK	2002	Underwood *et al.* (2013)	None
H071260094	87	3	clin	Spain	2007	Underwood *et al.* (2013)	None
H053260229	74	1	clin	UK	2005	Underwood *et al.* (2013)	None
H043940028	84	1	clin	UK	2004	Underwood *et al.* (2013)	None
Paris/ST1_1	1	1	clin	France	2002	Cazalet *et al.* (2004)	None
H074360702 /ST152_1	152	1	env	UK	2007	Underwood *et al.* (2013)	None
EUL 13 /ST5_1	5	1	clin	UK	1994	Underwood *et al.* (2013)	None
EUL00165 /ST37_1	37	1	clin	UK	2003	Underwood *et al.* (2013)	Related to ST37_64
Lens	15	1	clin	France	2003	Cazalet *et al.* (2004)	None

Appendix Table 2. Additional *L. pneumophila* isolates belonging to five major disease-associated STs (1, 23, 37, 47 and 62). These include 58 ST1 and 10 ST1-derived, 36 ST23, 71 ST37, 121 ST47 and 34 ST62 isolates. Two isolates belonging to ST18 and ST146, used for rooting some of the disease-associated lineages, are also included. Isolates with (1) in the ST column refer to ST1-derived isolates. ST - sequence type; Sg - serogroup; clin - clinical; env - environmental; TA - travel-associated; U/k – unknown

Isolate name	**Other name**	**ST**	**Sg**	**Source**	**Country**	**Year**	**Known epidemiolo-gic relatedness**	**Accession number/ Reference**
H034800423	ST1_2	1	1	env	UK	2003	None	Reuter *et al.* (2013)
EUL 55	ST1_3	1	1	clin	Spain	1994	Related to ST1_15	ERR332141
EUL 88	ST1_4	1	1	clin	Denmark	1995	None	ERR332174
EUL 93	ST1_5	1	1	clin	Denmark	1992	Related to ST1_24, ST1_25	ERR332179
EUL 10	ST1_6	1	1	env	Switzerland	1989	Related to ST1_9, ST1_20	ERR376635
EUL 1	ST1_7	1	1	clin	Switzerland	1998	None	ERR376626
EUL 21	ST1_8	1	1	env	UK	1999	None	ERR376638

EUL 3	ST1_9	1	1	clin	Switzerland	1989	Related to ST1_6, ST1_20	ERR376628
EUL 109	ST1_10	1	1	env	Sweden	1992	None	ERR376662
EUL 42	ST1_11	1	1	clin	Italy	1999	None	ERR376667
EUL 43	ST1_12	1	1	clin	Italy	1999	None	ERR376668
EUL 44	ST1_13	1	1	env	Italy	1999	Related to ST1_28	ERR376669
EUL 46	ST1_14	1	1	env	Italy	1999	None	ERR376671
EUL 58	ST1_15	1	1	env	Spain	1994	Related to ST1_3	ERR376683
EUL 60	ST1_16	1	1	clin	Greece	1992	None	ERR376685
EUL 62	ST1_17	1	1	env	Greece	1989	None	ERR376687
EUL 67	ST1_18	1	1	clin	Greece	1995	None	ERR376692
EUL 85	ST1_19	1	1	clin	Denmark	1995	None	ERR376710
EUL 9	ST1_20	1	1	env	Switzerland	1989	Related to ST1_6, ST1_9	ERR376634
EUL 82	ST1_21	1	1	clin	Denmark	1994	None	ERR376733
EUL 84	ST1_22	1	1	clin	Denmark	1995	None	ERR376735
EUL 90	ST1_23	1	1	clin	Denmark	U/k	None	ERR376736
EUL 94	ST1_24	1	1	clin	Denmark	1992	Related to ST1_5, ST1_25	ERR376738
EUL 95	ST1_25	1	1	env	Denmark	1993	Related to ST1_5, ST1_24	ERR376739
EUL 104	ST1_26	1	1	clin	Sweden	1992	None	ERR376745
EUL 108	ST1_27	1	1	clin	Sweden	1992	None	ERR376748
EUL 37	ST1_28	1	1	clin	Italy	1999	Related to ST1_13	ERR376723
EUL 119	ST1_29	1	1	clin	Germany	2005	None	ERR376757
EUL 53	ST1_30	1	1	clin	Spain	1995	None	ERR376725
OLDA1 (NCTC1208)	ST1_31	1	1	clin	USA	1947	None	ERR434061
HL00364001	ST1_32	1	1	clin	France	2000	None	ERR922483
HL02304015	ST1_33	1	1	clin	France	2002	None	ERR922484
HL03111005	ST1_34	1	1	env	France	2003	None	ERR922485
HL03373012	ST1_35	1	1	env	France	2003	None	ERR922486
HL04163014	ST1_36	1	1	clin	France	2004	None	ERR922487
HL07013004	ST1_37	1	1	env	France	2007	None	ERR922488
LG09192006	ST1_38	1	1	clin	France	2009	None	ERR922489
LG09404015	ST1_39	1	1	clin	France	2009	None	ERR922490
LG10191002	ST1_40	1	1	clin	France	2010	Related to ST1_41	ERR922491
LG10203012	ST1_41	1	1	env	France	2010	Related to ST1_40	ERR922492
LG11011012	ST1_42	1	1	env	France	2010	None	ERR922493
LG11054025	ST1_43	1	1	env	France	2011	None	ERR922494
LG11181044	ST1_44	1	1	env	Morocco	2009	None	ERR922495
LG11391124	ST1_45	1	1	env	France	2011	None	ERR922496

LP21	ST1_46	1	1	clin	Sweden	1996 - 1999	U/k	ERR922497
LP23	ST1_47	1	1	clin	Sweden	1996 - 2000	None	ERR922498
LT 40/04	ST1_48	1	1	clin	Austria	2004	None	ERR922499
NIIB223	ST1_49	1	1	env	Japan	1986	U/k	ERR922500
NIIB225	ST1_50	1	1	env	Japan	1986	U/k	ERR922501
L 3386/03	ST1_51	1	1	env	Austria	2003	None	ERR922502
L 3415/03	ST1_52	1	1	env	Austria	2003	None	ERR922503
LG10143009	ST1_53	1	1	clin	France	2010	None	ERR922504
NIIB80	ST1_54	1	1	clin	Japan	1981	None	ERR923392
LP22	ST1_55	1	1	clin	Sweden	1996 - 1999	U/k	ERR923393
L00-549	ST1_56	1	1	clin	Germany	2000	None	ERR923394
E21203	ST1_57	1	1	clin	France	2004	None	ERR923395
2735	ST1_58	1	1	env	USA	2002	None	ERR923396
Wien 47-14	ST1_59	1	1	env	Austria	1996	None	ERR923397
EUL 14	ST5_2	5 (1)	1	clin	UK	1984	None	ERR376639
EUL 16	ST5_3	5 (1)	1	clin	UK	1984	None	ERR376641
EUL 17	ST7_1	7 (1)	1	clin	UK	1993	None	ERR376642
EUL 113	ST7_2	7 (1)	1	env	Germany	1995	None	ERR376751
EUL 114	ST7_3	7 (1)	1	env	Germany	1995	None	ERR376752
EUL 45	ST72_1	72 (1)	1	clin	Italy	1999	None	ERR376670
EUL 110	ST10_1	10 (1)	1	clin	Germany	1993	None	ERR376674
EUL 117	ST6_1	6 (1)	1	clin	Germany	2005	None	ERR376755
EUL 157	ST8_1	8 (1)	1	env	UK	2004	None	ERR376779
IN-23-G1-C2 (ATCC 35289)	ST390_1	390 (1)	9	env	Netherlands	1988	None	ERR923391
EUL 8	ST23_2	23	1	clin	Switzerland	1993	Related to ST23_3, ST23_4	ERR376633
EUL 11	ST23_3	23	1	env	Switzerland	1993	Related to ST23_2, ST23_4	ERR376636
EUL 12	ST23_4	23	1	env	Switzerland	1993	Related to ST23_2, ST23_3	ERR376637
EUL 41	ST23_5	23	1	clin	Italy	1999	None	ERR376666
EUL 130	ST23_6	23	1	clin	Croatia	1987	Related to ST23_7	ERR376703
EUL 129	ST23_7	23	1	clin	Croatia	1987	Related to ST23_6	ERR376762

EUL 4	ST23_8	23	1	clin	Switzerland	1991	None	ERR376721
EUL 28	ST23_9	23	1	clin	France	1994	None	ERR376722
HL01273027	ST23_10	23	1	clin	France	2001	None	ERR922505
HL02365014	ST23_11	23	1	clin	France	2002	None	ERR922506
HL02365015	ST23_12	23	1	clin	France	2002	None	ERR922507
HL03071012	ST23_13	23	1	clin	France	2003	None	ERR922508
HL03393028	ST23_14	23	1	clin	France	2003	None	ERR922509
HL04371017	ST23_15	23	1	clin	France	2004	None	ERR922510
HL04433031	ST23_16	23	1	clin	France	2004	None	ERR922511
HL05063005	ST23_17	23	1	clin	France	2005	None	ERR922512
HL05322037	ST23_18	23	1	env	France	2005	None	ERR922513
HL05415018	ST23_19	23	1	clin	France	2005	None	ERR922514
HL06043045	ST23_20	23	1	clin	France	2006	None	ERR922515
HL06373021	ST23_21	23	1	clin	France	2006	None	ERR922516
HL07093017	ST23_22	23	1	clin	France	2007	None	ERR922517
LG07512008	ST23_23	23	1	clin	France	2007	None	ERR922518
LG08345006	ST23_24	23	1	clin	France	2008	None	ERR922519
LG08392025	ST23_25	23	1	clin	France	2008	None	ERR922520
LG09153012	ST23_26	23	1	env	France	2009	None	ERR922521
LG09353013	ST23_27	23	1	env	France	2009	None	ERR922522
LG09403015	ST23_28	23	1	clin	France	2009	None	ERR922523
LG09454021	ST23_29	23	1	clin	France	2009	None	ERR922524
LG10255002	ST23_30	23	1	clin	France	2010	None	ERR922525
LG10363013	ST23_31	23	1	env	France	2010	None	ERR922526
LG10481020	ST23_32	23	1	clin	France	2010	None	ERR922527
LG11272006	ST23_33	23	1	clin	France	2011	None	ERR922528
LG11363009	ST23_34	23	1	clin	France	2011	None	ERR922529
LG11402026	ST23_35	23	1	clin	France	2011	None	ERR922530
LG12242012	ST23_36	23	1	clin	France	2012	None	ERR922531
LG12465006	ST23_37	23	1	clin	France	2012	None	ERR922532
H064240448	ST37_2	37	1	env	UK	2006	None	ERR363849
LC0731	ST37_3	37	1	clin	UK	1989	Related to ST37_4, ST37_5, ST37_59, ST37_61, ST37_63	ERR363882
LC0732	ST37_4	37	1	clin	UK	1989	Related to ST37_3, ST37_5, ST37_59, ST37_61, ST37_63	ERR363883
LC0763	ST37_5	37	1	env	UK	1989	Related to ST37_3,	ERR363884

							ST37_4, ST37_59, ST37_61, ST37_63	
LC5694	ST37_6	37	1	clin	UK	2000	None	ERR363891
LC5722	ST37_7	37	1	clin	UK	2000	None	ERR363892
LC5738	ST37_8	37	1	clin	UK	2000	None	ERR363893
LC5755	ST37_9	37	1	clin	UK	2000	None	ERR363894
LC5908	ST37_10	37	1	clin	UK	2001	None	ERR363895
LC6163	ST37_11	37	1	clin	UK	2002	None	ERR363897
LC6267	ST37_12	37	1	clin	UK	2002	None	ERR363899
LC6268	ST37_13	37	1	clin	UK	2002	None	ERR363900
LC6228	ST37_14	37	1	clin	UK	2002	None	ERR363898
H041380048	ST37_15	37	1	clin	UK	2004	Related to ST37_23	ERR363843
H042960010	ST37_16	37	1	clin	UK	2004	None	ERR363845
H061140013	ST37_17	37	1	clin	UK	2006	None	ERR363847
H071880001	ST37_18	37	1	clin	UK	2007	None	ERR363850
H073060003	ST37_19	37	1	clin	UK	2007	None	ERR363851
H080820009	ST37_20	37	1	clin	UK	2008	None	ERR363853
LC6058	ST37_21	37	1	clin	U/k (TA)	2001	None	ERR363896
LC6293	ST37_22	37	1	clin	U/k (TA)	2002	None	ERR363901
H041640791	ST37_23	37	1	env	UK	2004	Related to ST37_15	ERR363844
LC6788	ST37_24	37	1	clin	U/k (TA)	2003	None	ERR363902
H062660463	ST37_25	37	1	clin	U/k (TA)	2006	None	ERR363848
H073900557	ST37_26	37	1	clin	U/k (TA)	2007	None	ERR363852
LC1127	ST37_27	37	1	clin	UK	1989	None	ERR363890
H084760449	ST37_28	37	1	clin	UK	2008	None	ERR363857
H085020185	ST37_29	37	1	clin	UK	2008	None	ERR363858
H090320386	ST37_30	37	1	clin	UK	2009	None	ERR363859
H044260061	ST37_31	37	1	env	UK	2004	None	ERR363846
H093140322	ST37_32	37	1	clin	UK	2009	Related to ST37_33	ERR363861
H093160422	ST37_33	37	1	env	UK	2009	Related to ST37_32	ERR363862
H092760433	ST37_34	37	1	clin	U/k (TA)	2009	None	ERR363860
H100940111	ST37_35	37	1	clin	UK	2010	None	ERR363863
H101760092	ST37_36	37	1	clin	UK	2010	None	ERR363864
H101820190	ST37_37	37	1	clin	UK	2010	None	ERR363865
H102020414	ST37_38	37	1	clin	UK	2010	None	ERR363867
H101980130	ST37_39	37	1	clin	U/k (TA)	2010	None	ERR363866
H103820081	ST37_40	37	1	clin	UK	2010	None	ERR363868
H120240685	ST37_41	37	1	clin	Slovenia	2010	None	ERR363992

H104320293	ST37_42	37	1	env	UK	2010	None	ERR363869
H113180118	ST37_43	37	1	clin	UK	2011	Related to ST37_44	ERR363871
H113340664	ST37_44	37	1	env	UK	2011	Related to ST37_43	ERR363873
H113280076	ST37_45	37	1	clin	UK	2011	None	ERR363872
H113660550	ST37_46	37	1	clin	UK	2011	None	ERR363874
H114740454	ST37_47	37	1	clin	UK	2011	None	ERR363876
H115040456	ST37_48	37	1	clin	UK	2011	None	ERR363877
H111580389	ST37_49	37	1	clin	UK	2011	None	ERR363870
H113780240	ST37_50	37	1	clin	U/k (TA)	2011	None	ERR363875
H083920177	ST37_51	37	1	clin	UK	2008	Related to ST37_52	ERR363855
H084140691	ST37_52	37	1	env	UK	2008	Related to ST37_51	ERR363856
H081180019	ST37_53	37	1	env	UK	2008	None	ERR363854
H103260667	ST37_54	37	1	env	Greece	2010	None	ERR363938
LC464	ST37_55	37	1	clin	UK	1987	None	ERR363878
LC0512	ST37_56	37	1	clin	U/k (TA)	1988	None	ERR363879
LC0565	ST37_57	37	1	clin	UK	1988	Related to ST37_58, ST37_69, ST37_70, ST37_71	ERR363880
LC0583	ST37_58	37	1	clin	UK	1988	Related to ST37_57, ST37_69, ST37_70, ST37_71	ERR363881
LC0782	ST37_59	37	1	clin	UK	1989	Related to ST37_3, ST37_4, ST37_5, ST37_61, ST37_63	ERR363885
LC0794	ST37_60	37	1	clin	UK	1989	Related to ST37_62	ERR363886
LC0795	ST37_61	37	1	clin	UK	1989	Related to ST37_3, ST37_4, ST37_5, ST37_59, ST37_63	ERR363887
LC0798	ST37_62	37	1	clin	UK	1989	Related to ST37_60	ERR363888
LC0801	ST37_63	37	1	clin	UK	1989	Related to ST37_3, ST37_4, ST37_5, ST37_59, ST37_61	ERR363889
EUL 166 /LP056	ST37_64	37	1	env	UK	2003	Related to ST37_1	ERR364007
EUL 69	ST37_65	37	1	clin	UK	1995	None	ERR332155

EUL 73	ST37_66	37	1	clin	UK	1996	Related to ST37_67, ST37_68	ERR332159
EUL 78	ST37_67	37	1	clin	UK	1996	Related to ST37_66, ST37_68	ERR340955
EUL 79	ST37_68	37	1	clin	UK	1996	Related to ST37_66, ST37_67	ERR340956
EUL 132	ST37_69	37	1	clin	UK	1988	Related to ST37_57, ST37_58, ST37_70, ST37_71	ERR332168
EUL 133	ST37_70	37	1	clin	UK	1988	Related to ST37_57, ST37_58, ST37_69, ST37_71	ERR332169
EUL 134	ST37_71	37	1	clin	UK	1988	Related to ST37_57, ST37_58, ST37_69, ST37_70	ERR332170
EUL 131	ST37_72	37	1	clin	UK	1988	None	ERR332167
EUL 169	ST47_2	47	1	clin	UK	2006	Related to ST47_5, ST47_99	Underwood *et al.* (2013)
H034700617	ST47_3	47	1	clin	UK	2003	None	Reuter *et al.* (2013)
HL01313013	ST47_4	47	1	clin	France	2001	None	ERR1341919
H064160534	ST47_5	47	1	env	UK	2006	Related to ST47_2, ST47_99	ERR363994
H043580159	ST47_6	47	1	clin	UK	2004	None	ERR363943
H043580160	ST47_7	47	1	clin	UK	2004	None	ERR363959
H043660021	ST47_8	47	1	clin	UK	2004	None	ERR363946
H043680663	ST47_9	47	1	clin	UK	2004	None	ERR363949
H043700021	ST47_10	47	1	clin	UK	2004	None	ERR363944
H043790008	ST47_11	47	1	clin	UK	2004	None	ERR363945
H052920051	ST47_12	47	1	clin	UK	2005	None	ERR363961
H053540106	ST47_13	47	1	clin	UK	2005	None	ERR363948
H063660005	ST47_14	47	1	clin	UK	2006	Related to ST47_15, ST47_21	ERR363904
H063660006	ST47_15	47	1	clin	UK	2006	Related to ST47_14, ST47_21	ERR363922
H063660009	ST47_16	47	1	clin	UK	2006	None	ERR363911
H063680006	ST47_17	47	1	clin	UK	2006	Related to ST47_18	ERR363918
H063680007	ST47_18	47	1	clin	UK	2006	Related to ST47_17	ERR363913

H063740003	ST47_19	47	1	clin	UK	2006	None	ERR363929
H063740018	ST47_20	47	1	clin	UK	2006	None	ERR363906
H063760006	ST47_21	47	1	clin	UK	2006	Related to ST47_14, ST47_15	ERR363915
H063780007	ST47_22	47	1	clin	UK	2006	Related to ST47_23	ERR363934
H063780008	ST47_23	47	1	clin	UK	2006	Related to ST47_22	ERR363916
H063860003	ST47_24	47	1	clin	UK	2006	None	ERR363930
H063960001	ST47_25	47	1	clin	UK	2006	None	ERR363928
LC5759	ST47_26	47	1	clin	U/k (TA)	2000	None	ERR363995
H070420013	ST47_27	47	1	clin	UK	2007	None	ERR363968
LC5822	ST47_28	47	1	clin	UK	2001	None	ERR363996
H040260015	ST47_29	47	1	clin	UK	2004	None	ERR363903
H055140095	ST47_30	47	1	clin	UK	2006	None	ERR363947
H060780053	ST47_31	47	1	clin	UK	2006	None	ERR363907
H061120064	ST47_32	47	1	clin	UK	2006	None	ERR363914
H062840608	ST47_33	47	1	clin	UK	2006	None	ERR363917
H062940111	ST47_34	47	1	clin	UK	2006	None	ERR363919
H064320006	ST47_35	47	1	clin	UK	2006	None	ERR363923
H064280005	ST47_36	47	1	clin	UK	2006	None	ERR363924
H064380002	ST47_37	47	1	clin	UK	2006	None	ERR363926
H064380001	ST47_38	47	1	clin	UK	2006	None	ERR363921
H064560527	ST47_39	47	1	clin	UK	2006	None	ERR363925
H064660638	ST47_40	47	1	clin	UK	2006	None	ERR363964
H070160015	ST47_41	47	1	clin	UK	2007	None	ERR363970
H071120010	ST47_42	47	1	clin	UK	2007	None	ERR363931
H071360036	ST47_43	47	1	clin	UK	2007	None	ERR363908
H072740002	ST47_44	47	1	clin	UK	2007	None	ERR363935
H073000045	ST47_45	47	1	clin	UK	2007	None	ERR363932
H073380007	ST47_46	47	1	clin	UK	2007	None	ERR363940
H073600182	ST47_47	47	1	clin	UK	2007	None	ERR363976
H073640185	ST47_48	47	1	clin	UK	2007	None	ERR363933
H074960018	ST47_49	47	1	clin	UK	2008	None	ERR363920
H080780059	ST47_50	47	1	clin	UK	2008	None	ERR363910
H053840008	ST47_51	47	1	clin	UK	2004	None	ERR363954
H072520002	ST47_52	47	1	clin	UK	2007	None	ERR363927
H081340222	ST47_53	47	1	clin	UK	2007	None	ERR363909
H082520613	ST47_54	47	1	clin	UK	2008	None	ERR363912
H083120262	ST47_55	47	1	clin	UK	2008	None	ERR363941
H083620580	ST47_56	47	1	clin	UK	2008	None	ERR363936

H083960064	ST47_57	47	1	clin	UK	2008	None	ERR363937
H084620118	ST47_58	47	1	clin	UK	2008	None	ERR363939
H090140214	ST47_59	47	1	clin	UK	2009	None	ERR363963
H090440226	ST47_60	47	1	clin	UK	2009	None	ERR363966
H040960441	ST47_61	47	1	clin	UK	2004	None	ERR363953
H041120007	ST47_62	47	1	clin	UK	2004	None	ERR363942
H093480403	ST47_63	47	1	clin	U/k (TA)	2009	None	ERR363973
H094340202	ST47_64	47	1	clin	UK	2009	None	ERR363971
H095060125	ST47_65	47	1	clin	UK	2009	None	ERR363972
H100140151	ST47_66	47	1	clin	UK	2010	None	ERR363965
H100660110	ST47_67	47	1	clin	UK	2010	None	ERR363962
H100700025	ST47_68	47	1	clin	UK	2010	None	ERR363958
H103140121	ST47_69	47	1	clin	UK	2010	None	ERR363967
H103620160	ST47_70	47	1	clin	UK	2010	None	ERR363950
H103660126	ST47_71	47	1	clin	UK	2010	None	ERR363974
H103660121	ST47_72	47	1	clin	UK	2010	None	ERR363956
H104420240	ST47_73	47	1	clin	UK	2010	None	ERR363957
H110480273	ST47_74	47	1	clin	UK	2011	None	ERR363969
H112320437	ST47_75	47	1	clin	UK	2011	None	ERR363951
H112080616	ST47_76	47	1	clin	UK	2011	None	ERR363952
H112380374	ST47_77	47	1	clin	UK	2011	None	ERR363960
H120160499	ST47_78	47	1	clin	UK	2012	None	ERR363985
H120200371	ST47_79	47	1	clin	UK	2012	None	ERR363984
H105140391	ST47_80	47	1	clin	UK	2010	None	ERR363993
H121040204	ST47_81	47	1	clin	UK	2012	None	ERR363982
H121420445	ST47_82	47	1	clin	UK	2012	None	ERR363983
H102240357	ST47_83	47	1	clin	UK	2010	None	ERR363955
H122500497	ST47_84	47	1	clin	UK	2012	None	ERR363981
H122820408	ST47_85	47	1	clin	U/k (TA)	2012	None	ERR363980
H123620597	ST47_86	47	1	clin	UK	2012	None	ERR363979
H123840629	ST47_87	47	1	clin	UK	2012	None	ERR363978
H123940534	ST47_88	47	1	clin	UK	2012	None	ERR363975
H124920387	ST47_89	47	1	clin	UK	2012	None	ERR363991
H131340777	ST47_90	47	1	clin	UK	2013	Related to ST47_92, ST47_93, ST47_94	ERR363990
H131460248	ST47_91	47	1	clin	UK	2013	None	ERR363987
H131480353	ST47_92	47	1	env	UK	2013	Related to ST47_90, ST47_93, ST47_94	ERR363989
H131480354	ST47_93	47	1	env	UK	2013	Related to	ERR363988

							ST47_90, ST47_92, ST47_94	
H131840211	ST47_94	47	1	env	UK	2013	Related to ST47_90, ST47_92, ST47_93	ERR363986
H132140863	ST47_95	47	1	clin	UK	2013	None	ERR364031
EUL 31	ST47_96	47	1	clin	France	1994	None	ERR376656
EUL 70	ST47_97	47	1	clin	UK	1996	None	ERR376695
EUL 168	ST47_98	47	1	clin	UK	2005	None	ERR352161
EUL 170	ST47_99	47	1	env	UK	2006	Related to ST47_2, ST47_5	ERR376788
LG12084002	ST47_100	47	1	clin	France	2012	None	ERR922533
LG12034018	ST47_101	47	1	clin	France	2012	None	ERR1193351
LG11463009	ST47_102	47	1	clin	France	2011	None	ERR922534
LG11415002	ST47_103	47	1	clin	France	2011	None	ERR922535
LG11403003	ST47_104	47	1	clin	France	2011	None	ERR922536
LG10425016	ST47_105	47	1	clin	France	2010	None	ERR922537
LG10397001	ST47_106	47	1	clin	France	2010	None	ERR922538
LG09534017	ST47_107	47	1	clin	France	2009	None	ERR922539
LG09471012	ST47_108	47	1	clin	France	2009	None	ERR922540
LG08394013	ST47_109	47	1	clin	France	2008	None	ERR922541
LG08251002	ST47_110	47	1	clin	France	2008	None	ERR922542
HL07512016	ST47_111	47	1	clin	France	2007	None	ERR922543
HL07055011	ST47_112	47	1	clin	France	2007	None	ERR922544
HL06353025	ST47_113	47	1	clin	France	2006	None	ERR922545
HL05383032	ST47_114	47	1	clin	France	2005	None	ERR922546
HL05375017	ST47_115	47	1	clin	France	2005	None	ERR922547
HL04411050	ST47_116	47	1	env	France	2004	None	ERR922548
HL04284070	ST47_117	47	1	clin	France	2004	None	ERR922549
HL04075055	ST47_118	47	1	clin	France	2004	None	ERR922550
HL03503011	ST47_119	47	1	clin	France	2003	None	ERR922551
HL03443027	ST47_120	47	1	clin	France	2003	None	ERR922552
HL02392002	ST47_121	47	1	clin	France	2002	None	ERR922553
HL02274033	ST47_122	47	1	clin	France	2002	None	ERR922554
H043540106	ST62_2	62	1	clin	U/k (TA)	2004	None	ERR363997
H044120014	ST62_3	62	1	clin	Bulgaria	2004	None	ERR363999
H052780022	ST62_4	62	1	clin	UK	2005	None	ERR363998
H054280040	ST62_5	62	1	clin	UK	2005	None	ERR364028
H063680003	ST62_6	62	1	clin	UK	2006	None	ERR364002
H063840008	ST62_7	62	1	clin	UK	2006	None	ERR364001

H073660582	ST62_8	62	1	clin	UK	2007	None	ERR364008
LC5804	ST62_9	62	1	clin	UK	2000	None	ERR364029
H063760005	ST62_10	62	1	clin	UK	2006	None	ERR364000
H064240003	ST62_11	62	1	clin	UK	2006	None	ERR364005
H065040012	ST62_12	62	1	clin	UK	2007	None	ERR364012
H070140635	ST62_13	62	1	clin	UK	2007	None	ERR364011
H073020039	ST62_14	62	1	clin	UK	2007	None	ERR364022
H073320399	ST62_15	62	1	clin	UK	2007	None	ERR364010
H073440003	ST62_16	62	1	clin	UK	2007	None	ERR364009
LC6009	ST62_17	62	1	clin	U/k (TA)	2001	None	ERR364030
H083140015	ST62_18	62	1	clin	UK	2008	None	ERR364007
H064180019	ST62_19	62	1	env	UK	2006	Related to ST62_1	ERR364004
H093400182	ST62_20	62	1	clin	UK	2009	None	ERR364006
H094760070	ST62_21	62	1	clin	UK	2009	None	ERR364003
H094800237	ST62_22	62	1	clin	UK	2009	None	ERR364020
H110480715	ST62_23	62	1	clin	UK	2011	None	ERR364018
H112840293	ST62_24	62	1	clin	UK	2011	None	ERR364017
H114100406	ST62_25	62	1	clin	Greece	2011	None	ERR364016
H120240362	ST62_26	62	1	clin	UK	2012	None	ERR364025
H104640262	ST62_27	62	1	clin	U/k (TA)	2010	None	ERR364019
H123140428	ST62_28	62	1	env	UK	2012	None	ERR364015
H123460520	ST62_29	62	1	clin	UK	2012	None	ERR364014
H124360642	ST62_30	62	1	clin	UK	2012	None	ERR364013
EUL 54	ST62_31	62	1	clin	Spain	1994	Related to ST62_32	ERR332140
EUL 57	ST62_32	62	1	env	Spain	1995	Related to ST62_31	ERR332143
EUL 71	ST62_33	62	1	clin	UK	1996	Related to ST62_34, ST62_35	ERR332157
EUL 76	ST62_34	62	1	clin	UK	1996	Related to ST62_33, ST62_35	ERR332162
EUL 77	ST62_35	62	1	clin	UK	1996	Related to ST62_33, ST62_34	ERR332163
EUL 7		18	1	clin	Switzerland	1992	None	ERR376632
LG12482019		146	1	clin	France	2012	None	ERR923430

9.2 Chapter 4

Appendix Table 3. Additional isolates belonging to STs 1, 23, 42 and 578 used in Chapter 4. ST - sequence type; Sg - serogroup; clin - clinical; env - environmental; U/k – unknown

Isolate name	Other name	ST	Sg	Source	Country	Year	Known epidemiologic relatedness	Accession number/ Reference
ID_1688	ST1_60	1	1	env	Spain	2004	U/k	Sanchez-Buso *et al.* (2014)
ID_1690	ST1_61	1	1	env	Spain	2004	U/k	Sanchez-Buso *et al.* (2014)
ID_1828	ST1_62	1	1	env	Spain	2004	U/k	Sanchez-Buso *et al.* (2014)
ID_2041	ST1_63	1	1	env	Spain	2005	U/k	Sanchez-Buso *et al.* (2014)
ID_2947	ST1_64	1	1	env	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_2948	ST1_65	1	1	env	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_598	ST1_66	1	1	env	Spain	2002	U/k	Sanchez-Buso *et al.* (2014)
ID_6885	ST1_67	1	1	env	Spain	2011	U/k	Sanchez-Buso *et al.* (2014)
ID_747970	ST1_68	1	1	env	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_891	ST1_69	1	1	env	Spain	2002	U/k	Sanchez-Buso *et al.* (2014)
ID_125_BC1	ST23_38	23	1	clin	Spain	2012	U/k	Sanchez-Buso *et al.* (2016)
ID_192091_ BC52	ST23_39	23	1	clin	Spain	2012	U/k	Sanchez-Buso *et al.* (2016)
ID_4029_ BC37	ST23_40	23	1	env	Spain	2012	U/k	Sanchez-Buso *et al.* (2016)
ID_50291_ BC50	ST23_41	23	1	clin	Spain	2012	U/k	Sanchez-Buso *et al.* (2016)
ID_50726_ BC51	ST23_42	23	1	clin	Spain	2012	U/k	Sanchez-Buso *et al.* (2016)
ID_2680_ BC17	ST578_1	578	1	clin	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_2301_ BC14	ST578_2	578	1	clin	Spain	1999	U/k	Sanchez-Buso *et al.* (2014)
ID_2376_ BC15	ST578_3	578	1	clin	Spain	1999	U/k	Sanchez-Buso *et al.* (2014)
ID_3009_ BC21	ST578_4	578	1	clin	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_3108_ BC23	ST578_5	578	1	clin	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_3109_ BC24	ST578_6	578	1	clin	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_3110_ BC25	ST578_7	578	1	clin	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_3355_	ST578_8	578	1	clin	Spain	2000	U/k	Sanchez-Buso

BC32								*et al.* (2014)
ID_3785_ BC34	ST578_9	578	1	clin	Spain	2001	U/k	Sanchez-Buso *et al.* (2014)
ID_3908_ BC36	ST578_ 10	578	1	clin	Spain	2001	U/k	Sanchez-Buso *et al.* (2014)
ID_5856_ BC39	ST578_ 11	578	1	clin	Spain	2002	U/k	Sanchez-Buso *et al.* (2014)
ID_6536_ BC40	ST578_ 12	578	1	clin	Spain	2002	U/k	Sanchez-Buso *et al.* (2014)
ID_7147_ BC42	ST578_ 13	578	1	clin	Spain	2003	U/k	Sanchez-Buso *et al.* (2014)
ID_8141_ BC45	ST578_ 14	578	1	clin	Spain	2003	U/k	Sanchez-Buso *et al.* (2014)
ID_8189_ BC46	ST578_ 15	578	1	clin	Spain	2003	U/k	Sanchez-Buso *et al.* (2014)
ID_8190_ BC47	ST578_ 16	578	1	clin	Spain	2003	U/k	Sanchez-Buso *et al.* (2014)
ID_8227_ BC48	ST578_ 17	578	1	clin	Spain	2003	U/k	Sanchez-Buso *et al.* (2014)
ID_8228_ BC49	ST578_ 18	578	1	clin	Spain	2003	U/k	Sanchez-Buso *et al.* (2014)
ID_480203_ BC53	ST578_ 19	578	1	clin	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_480295_ BC55	ST578_ 20	578	1	clin	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_480372_ BC56	ST578_ 21	578	1	clin	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_480392_ BC57	ST578_ 22	578	1	clin	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_747968_ BC72	ST578_ 23	578	1	env	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_747969_ BC73	ST578_ 24	578	1	env	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_747973_ BC75	ST578_ 25	578	1	env	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_481107_ BC58	ST578_ 26	578	1	clin	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_481441_ BC59	ST578_ 27	578	1	clin	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_481707_ BC60	ST578_ 28	578	1	clin	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_481710_ BC61	ST578_ 29	578	1	clin	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_1190176_BC76	ST578_ 30	578	1	env	Spain	2010	U/k	Sanchez-Buso *et al.* (2014)
ID_489571_ BC65	ST578_ 31	578	1	clin	Spain	2010	U/k	Sanchez-Buso *et al.* (2014)
ID_489956_ BC66	ST578_ 32	578	1	clin	Spain	2010	U/k	Sanchez-Buso *et al.* (2014)
ID_490679_ BC68	ST578_ 33	578	1	clin	Spain	2010	U/k	Sanchez-Buso *et al.* (2014)
ID_490738_ BC69	ST578_ 34	578	1	clin	Spain	2010	U/k	Sanchez-Buso *et al.* (2014)
ID_1925_ BC12	ST578_ 35	578	1	env	Spain	2004	U/k	Sanchez-Buso *et al.* (2014)
ID_3499_ BC33	ST578_3 6	578	1	clin	Spain	2001	U/k	Sanchez-Buso *et al.* (2014)
ID_3786_	ST578_3	578	1	clin	Spain	2001	U/k	Sanchez-Buso

BC35	7							*et al.* (2014)
ID_480263_ BC54	ST578_ 38	578	1	clin	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_481898_ BC62	ST578_ 39	578	1	clin	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_481944_ BC63	ST578_ 40	578	1	clin	Spain	2009	U/k	Sanchez-Buso *et al.* (2014)
ID_489154_ BC64	ST578_ 41	578	1	clin	Spain	2010	U/k	Sanchez-Buso *et al.* (2014)
ID_490456_ BC67	ST578_ 42	578	1	clin	Spain	2010	U/k	Sanchez-Buso *et al.* (2014)
ID_5228_ BC38	ST578_ 43	578	1	clin	Spain	2002	U/k	Sanchez-Buso *et al.* (2014)
ID_7371_ BC43	ST578_ 44	578	1	clin	Spain	2003	U/k	Sanchez-Buso *et al.* (2014)
ID_8004_ BC44	ST578_ 45	578	1	clin	Spain	2003	U/k	Sanchez-Buso *et al.* (2014)
EUL 6	ST42_1	42	1	clin	Switzerland	1999	None	ERR376631
EUL 27	ST42_2	42	1	clin	France	2005	None	ERR376652
EUL 39	ST42_3	42	1	clin	Italy	1999	None	ERR376664
EUL 50	ST42_4	42	1	clin	Spain	1996	None	ERR376675
EUL 75	ST42_5	42	1	clin	UK	1995	None	ERR376700
EUL 105	ST42_6	42	1	clin	Sweden	1991	None	ERR376746
EUL 116	ST42_7	42	1	clin	Germany	1996	None	ERR376754
EUL 120	ST42_8	42	1	clin	Germany	1999	Related to EUL 121	ERR376758
EUL 121	ST42_9	42	1	clin	Germany	1999	Related to EUL 120	ERR376678
EUL 122	ST42_10	42	1	clin	Unknown	1987	Related to EUL 123	ERR376759
EUL 123	ST42_11	42	1	clin	Unknown	1987	Related to EUL 122	ERR332142
EUL 124	ST42_12	42	1	clin	UK	1987	Related to EUL 125	ERR332150
EUL 125	ST42_13	42	1	clin	UK	1987	Related to EUL 124	ERR376760

Appendix Table 4. An additional 100 isolates used in the inference of recombination donors that are not listed in Appendix Tables 1-3. ST - sequence type; Sg - serogroup; clin – clinical; env – environmental; U/k – unknown; NA – not applicable

Isolate name	**ST**	**Sg**	**Source**	**Country**	**Year**	**Known epidemiologic relatedness**	**Accession number/ Reference**
ATCC 43290	187	12	clin	USA	U/k	None	Amaro *et al.* (2012)
HL06041035	734	1	env	France	2006	None	Gomez-Valero *et al.* (2011)
Thunderbay	187	6	clin	Canada	U/k	None	Khan *et al.* (2013)

EUL 2	2	1	clin	Switzerland	1989	None	ERR376627
EUL 5	114	6	clin	Switzerland	U/k	None	ERR376630
EUL 7	18	1	clin	Switzerland	1992	None	ERR376632
EUL 18	26	1	clin	Scotland	1994	None	ERR376643
EUL 19	9	1	clin	Scotland	1994	Related to EUL 22, 23, 24	ERR376644
EUL 20	28	1	clin	Scotland	1995	None	ERR376645
EUL 22	9	1	clin	Scotland	1994	Related to EUL 19, 23, 24	ERR376647
EUL 23	9	1	clin	Scotland	1994	Related to EUL 19, 22, 24	ERR376648
EUL 24	9	1	env	Scotland	1994	Related to EUL 19, 22, 23	ERR332110
EUL 25	44	1	clin	France	1994	None	ERR376650
EUL 26	22	1	clin	France	U/k	None	ERR376651
EUL 30	38	1	clin	France	U/k	None	ERR376655
EUL 32	16	1	clin	France	1994	None	ERR376657
EUL 33	40	1	clin	France	U/k	Related to EUL 34, 35	ERR376658
EUL 34	40	1	env	France	U/k	Related to EUL 33, 35	ERR376659
EUL 35	40	1	env	France	1996	Related to EUL 33, 34	ERR376660
EUL 36	21	1	clin	Italy	1999	None	ERR332122
EUL 48	48	1	clin	Spain	1996	Related to EUL 56	ERR332134
EUL 56	48	1	clin	Spain	1996	Related to EUL 48	ERR376726
EUL 61	77	1	env	Greece	1989	None	ERR376686
EUL 64	77	1	env	Greece	1986	None	ERR376727
EUL 68	46	1	clin	UK	1995	None	ERR376693
EUL 72	4	1	clin	UK	1996	None	ERR332158
EUL 74	29	1	clin	UK	1995	None	ERR376729
EUL 81	53	1	env	Denmark	1994	Related to EUL 96	ERR376732
EUL 83	50	1	clin	Denmark	1995	None	ERR376734
EUL 86	46	1	clin	Denmark	1995	None	ERR332172
EUL 91	63	1	clin	Denmark	1995	None	ERR376737
EUL 92	53	1	clin	Denmark	1991	None	ERR376717
EUL 96	53	1	clin	Denmark	1994	Related to EUL 81	ERR376740
EUL 97	9	1	clin	Sweden	1994	Related to EUL 107	ERR376741
EUL 98	9	1	clin	Sweden	1996	None	ERR376629
EUL 99	34	1	clin	Sweden	1995	None	ERR376704
EUL 100	59	1	clin	Sweden	1995	None	ERR376742
EUL 101	60	1	clin	Sweden	1994	None	ERR376743
EUL 102	59	1	clin	Sweden	1993	None	ERR376714
EUL 103	45	1	clin	Sweden	1993	None	ERR376744
EUL 107	9	1	env	Sweden	1994	Related to EUL 97	ERR376747

EUL 111	25	1	clin	Germany	1981	None	ERR376749
EUL 118	36	1	clin	Germany	1989	None	ERR340981
EUL 126	27	1	clin	UK	1985	Related to EUL 127, 128	ERR376691
EUL 127	27	1	clin	UK	1985	Related to EUL 126, 128	ERR376761
EUL 128	27	1	clin	UK	1985	Related to EUL 126, 127	ERR376699
EUL 144	48	1	env	UK	2002	None	ERR376768
EUL 145	78	1	env	UK	2002	Barrow outbreak	ERR376769
EUL 148	1321	8	env	Australia	2003	None	ERR376772
EUL 149	83	1	clin	UK	2004	None	ERR376773
EUL 150	79	1	clin	UK	2003	None	ERR376774
EUL 152	80	5	env	UK	2004	None	ERR352157
EUL 153	68	6	clin	UK	1986	Related to EUL 158	ERR376775
EUL 154	1326	8	clin	UK	1988	Related to EUL 155	ERR376776
EUL 155	1326	8	env	UK	1988	Related to EUL 154	ERR376777
EUL 158	68	6	env	UK	1986	Related to EUL 153	ERR376780
EUL 161	75	1	clin	UK	U/k	None	ERR376781
EUL 162	85	1	clin	UK	U/k	None	ERR376782
EUL 163	73	U/k	clin	Austria	U/k	None	ERR376783
EUL 167	82	1	clin	UK	U/k	None	ERR352160
H073240536	1327	5	clin	NA (cruise ship)	2007	Related to H073280012, H073340034, H073340594	ERR364024
H073280012	1327	5	env	NA (cruise ship)	2007	Related to H073240536, H073340034, H073340594	ERR364026
H073340034	1327	5	clin	NA (cruise ship)	2007	Related to H073240536, H073280012, H073340594	ERR364027
H073340594	1327	5	clin	NA (cruise ship)	2007	Related to H073240536, H073280012, H073340034	ERR364023
H092380261	109	U/k	clin	UK	2009	Related to H092400768	ERR434063
H092400768	109	U/k	env	UK	2009	Related to H092380261	ERR434064
H123640643	71	11	clin	U/k	U/k	None	ERR332166
LC6376	78	1	env	UK	2002	Barrow outbreak	ERR376790
LC6382	78	1	env	UK	2002	Barrow outbreak	ERR376792
LC6385	78	1	env	UK	2002	Barrow outbreak	ERR352162
LC6388	78	1	env	UK	2002	Barrow outbreak	ERR352163
LC6391	78	1	env	UK	2002	Barrow outbreak	ERR376793
LC6394	78	1	env	UK	2002	Barrow outbreak	ERR376794

LC6397	78	1	clin	UK	2002	Barrow outbreak	ERR376795
LC6406	78	1	clin	UK	2002	Barrow outbreak	ERR376796
LC6407	78	1	clin	UK	2002	Barrow outbreak	ERR376797
LC6408	78	1	clin	UK	2002	Barrow outbreak	ERR341023
LC6409	78	1	clin	UK	2002	Barrow outbreak	ERR352164
LC6410	78	1	clin	UK	2002	Barrow outbreak	ERR352165
LC6411	78	1	clin	UK	2002	Barrow outbreak	ERR376799
LC6412	78	1	clin	UK	2002	Barrow outbreak	ERR376800
LC6413	78	1	clin	UK	2002	Barrow outbreak	ERR376801
LC6416	78	1	clin	UK	2002	Barrow outbreak	ERR376802
LC6417	78	1	clin	UK	2002	Barrow outbreak	ERR376803
LC6418	78	1	clin	UK	2002	Barrow outbreak	ERR376804
ID_1885	1037	U/k	env	Spain	2004	U/k	Sanchez-Buso *et al.* (2014)
ID_2423	1037	U/k	env	Spain	1999	U/k	Sanchez-Buso *et al.* (2014)
ID_496053	1106	U/k	clin	Spain	2011	U/k	Sanchez-Buso *et al.* (2014)
ID_918	1236	U/k	U/k	Spain	U/k	U/k	Sanchez-Buso *et al.* (2014)
ID_3019	15	U/k	clin	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_479	171	U/k	env	Spain	2001	U/k	Sanchez-Buso *et al.* (2014)
ID_2949	328	U/k	env	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_3164	51	U/k	env	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_3201	637	U/k	env	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_3215	637	U/k	clin	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_3216	637	U/k	clin	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_3238	637	U/k	clin	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_3334	637	U/k	clin	Spain	2000	U/k	Sanchez-Buso *et al.* (2014)
ID_505237	637	U/k	clin	Spain	2011	U/k	Sanchez-Buso *et al.* (2014)
ID_482	804	U/k	env	Spain	2001	U/k	Sanchez-Buso *et al.* (2014)

Appendix Table 5. Sequencing statistics for four isolates sequenced using the Pacific Biosciences (PacBio) RSII sequencer.

Isolate	No. SMRT cells	No. contigs	Extra-chromo-somal plasmid	Mean cover-age	Total mapped reads	Mapped subread N50	Accession numbers
EUL 28 (ST23)	2	2 (3,514,605bp and 149,271bp)	Yes	79.3x	67,032	2.97kb	ERR660551 ERR663930
EUL 120 (ST42)	4	2 (2,732,926bp and 697,556bp)	No	121.8x	100,833	3.46kb	ERR663926 ERR663929 ERR671908 ERR690961
EUL 165 (ST37)	3	1 (3,486,389bp)	No	101.3x	74,378	3.47kb	ERR663927 ERR676880 ERR676882
H044120014 (ST62)	2	1 (3,541,412bp)	No	62x	45,460	4.59kb	ERR663928 ERR676881

Appendix Table 6. Genomic positions of repetitive regions and predicted mobile genetic elements (MGEs) in the six reference genomes (Paris/ST1; EUL 28/ST23; EUL 165/ST37; EUL 120/ST42; H044120014/ST62; Alcoy/ST578).

Reference genome	MGEs/repetitive regions	Start (bp)	End (bp)
Paris (ST1)	MGEs	46159	47563
		66766	85124
		135178	136374
		183008	234034
		237675	238718
		324918	332469
		376713	377346
		793632	795036
		862645	871886
		954348	954884
		961746	964139
		991147	992551
		1160427	1219613
		1733202	1746135
		1757883	1760206
		2046691	2048095
		2096462	2097866
		2112357	2114389
		2154197	2159177
		2205207	2206611

		2302584	2306608
		2311680	2312429
		2315629	2318863
		2322799	2323173
		2408503	2419758
		2581017	2582223
		2654264	2776774
		3280794	3299038
		3307993	3309397
		3370307	3374926
	Repetitive regions	46017	47645
		84055	84184
		84186	84417
		84432	84549
		84551	84784
		84786	84925
		135135	135551
		135553	136387
		136389	136532
		192590	192762
		225557	225727
		225729	225960
		225977	226094
		226189	226420
		226580	226757
		226759	227116
		227118	228316
		228318	228573
		229902	230631
		230659	230888
		233038	233618
		233620	233790
		233858	234049
		237451	238747
		376427	376528
		376530	376754
		376931	377236
		377238	377379
		432728	437887
		438476	439233
		439235	439668
		454759	455516
		455518	455951
		672085	672248
		675276	678229

		678231	678473
		678475	680279
		680305	680448
		753773	776812
		793490	795118
		866982	867892
		867894	869122
		908097	908302
		908696	909294
		941258	941606
		941708	941966
		944115	944463
		944565	944823
		991005	992637
		1160691	1162835
		1214103	1214280
		1214282	1214639
		1214641	1215839
		1215841	1216096
		1217551	1217819
		1400722	1400944
		1612592	1612744
		1612856	1613008
		1741749	1745218
		1758907	1759195
		1759197	1759527
		1759529	1759705
		1759745	1759864
		1759896	1760229
		1798998	1799103
		2046607	2049869
		2096347	2098442
		2205123	2208381
		2302555	2304718
		2311797	2311936
		2312032	2312268
		2312397	2312657
		2317698	2319146
		2373166	2373308
		2375190	2375335
		2375455	2375567
		2382059	2382201
		2384080	2384225
		2384345	2384457
		2400599	2400704

		2565424	2565773
		2580974	2581390
		2581392	2582226
		2582228	2582371
		2654235	2656381
		2664412	2664542
		2746607	2748753
		2754267	2755340
		2755342	2756858
		3020965	3021084
		3184953	3185096
		3186783	3191125
		3198943	3199106
		3290982	3292977
		3307851	3310581
		3315696	3315883
		3370184	3370472
		3370474	3370804
		3370806	3370982
		3371018	3371137
		3371168	3371501
		3453430	3453635
		3453637	3454166
		3454465	3454670
		3454672	3455201
EUL 28 (ST23)	MGEs	68459	87081
		184525	201357
		540430	541401
		889104	889640
		1090127	1209082
		1262908	1264312
		1293776	1296768
		1779836	1781240
		2265319	2390833
		2628381	2629577
		2701634	2766320
		2990966	2999528
		3374539	3378950
		3388959	3390363
	Repetitive regions	43294	43620
		43680	44257
		44259	44373
		280639	280870
		280920	281161
		326366	327641

		343575	344233
		344239	344583
		381793	383676
		383726	386952
		387541	387740
		387841	388241
		388243	388637
		403824	404023
		404124	404524
		404526	404920
		540176	540312
		540446	540581
		540583	540762
		540764	541546
		622714	622879
		625907	626170
		626172	627790
		627805	631091
		704322	721519
		846937	847044
		847251	847358
		897056	897212
		948671	948995
		1262765	1264598
		1397268	1397598
		1604321	1604454
		1730876	1730989
		1779800	1781421
		1822149	1823203
		2065597	2066568
		2085857	2087011
		2119394	2119507
		2165562	2166726
		2222017	2223151
		2382957	2384245
		2403947	2404150
		2427445	2428388
		2434402	2434604
		2435853	2436257
		2436737	2436898
		2437165	2437295
		2443299	2443501
		2444750	2445154
		2445631	2445792
		2446059	2446189

		2449041	2449151
		2451145	2451257
		2451978	2452086
		2452088	2452421
		2452430	2453097
		2453206	2453549
		2453847	2453957
		2455957	2456069
		2456790	2456898
		2456900	2457233
		2457242	2457909
		2458018	2458361
		2466814	2466913
		2508190	2508297
		2629018	2629800
		2695209	2695485
		2701604	2702892
		2757054	2758339
		2800215	2800322
		2827913	2828229
		2971986	2972085
		3015972	3017357
		3029597	3029929
		3196343	3196590
		3197873	3200048
		3200972	3202414
		3210094	3210419
		3388818	3390719
		3459212	3459417
		3459435	3459762
		3459764	3460020
		3460247	3460452
		3460470	3460797
		3460799	3461055
EUL 165 (ST37)	MGEs	172790	183163
		527549	528334
		740653	742059
		867654	868190
		904453	905857
		1073685	1221017
		1407332	1417368
		1969813	1972066
		2263685	2443144
		2515121	2516296
		2534145	2535551

		2744776	2770703
		2836001	2836438
		2967699	2980324
	Repetitive regions	172706	175499
		308419	308619
		359051	364205
		364794	365353
		365355	365984
		381076	381635
		381637	382266
		527360	527563
		527580	527776
		528165	528537
		607410	607520
		607525	612590
		683377	699638
		740593	740875
		740908	742082
		750489	750597
		821439	821644
		822038	822636
		854601	854949
		855051	855309
		857459	857807
		857909	858167
		1182049	1182400
		1182436	1183287
		1183299	1183410
		1196434	1197949
		1206656	1206795
		1206896	1206999
		1207289	1207407
		1207571	1207704
		2344092	2344225
		2344374	2344492
		2344767	2344870
		2345007	2345146
		2437777	2440572
		2442713	2442928
		2442930	2443156
		2478740	2478953
		2479018	2480344
		2480418	2480866
		2480907	2481070
		2481072	2481184

		2481792	2481895
		2487631	2487844
		2487909	2489235
		2489309	2489757
		2489798	2489961
		2489963	2490075
		2490683	2490786
		2515149	2515306
		2516012	2517460
		2535296	2536470
		2625664	2625772
		2835783	2836158
		2836407	2836603
		2836627	2836830
		3173105	3173327
		3173899	3174666
		3175676	3177451
		3177962	3179836
		3426189	3426989
		3427221	3428021
EUL 120 (ST42)	MGEs	1	60757
		355812	358377
		457092	457967
		697482	698935
		730314	742107
		759142	759678
		827201	828769
		964006	1053111
		1456188	1457399
		1654156	1655267
		1798125	1800378
		1952654	1965442
		1987708	1990686
		2113816	2267685
		2492938	2514904
		2690115	2691186
		2733011	2760276
		3013090	3013479
		3060467	3061538
		3426014	3429537
	Repetitive regions	1	409
		615	823
		871	990
		1054	1187
		1259	14010

		35494	35902
		36104	36312
		36358	36477
		36539	36672
		36742	44913
		53337	53572
		53574	53754
		53756	54701
		239400	242626
		242628	242906
		242908	244560
		245149	246339
		261431	262621
		355783	358499
		456965	457991
		487118	492291
		565861	585285
		731316	732342
		742215	742439
		1456166	1457514
		1654158	1654306
		1654308	1654654
		1654700	1654940
		1654942	1655077
		1655091	1655225
		1655227	1655419
		1800387	1800490
		1883141	1883286
		1957099	1958044
		1987686	1989034
		1989067	1990799
		2495762	2497487
		2498696	2498799
		2500430	2502167
		2690022	2690375
		2690377	2691336
		2736620	2736756
		2736920	2737052
		2748283	2748419
		2748583	2748715
		2776718	2776823
		2777025	2777130
		2953422	2953578
		2954540	2955659
		2956539	2960108

		3060374	3061058
		3061060	3061690
		3076741	3076995
		3151745	3151934
		3155573	3155762
		3167974	3168119
		3210132	3210940
		3211167	3211975
		3425957	3430562
H044120014 (ST62)	MGEs	171309	236885
		500116	501210
		578399	579595
		936356	936892
		1139075	1265997
		1306592	1307845
		1326070	1327474
		1356934	1359926
		1368895	1369989
		1840061	1841465
		2031174	2032575
		2181330	2187802
		2337326	2420261
		2531094	2532498
		2714961	2800869
		3025389	3030927
		3412308	3413712
	Repetitive regions	43308	43659
		43749	44192
		44263	44377
		221637	221748
		316178	316285
		316299	316409
		316459	316631
		361237	362492
		378427	379527
		418745	420510
		420645	423918
		424507	424726
		424728	425039
		425041	425207
		425209	425375
		425377	425699
		440790	441009
		441011	441322
		441324	441490

		441492	441658
		441660	441982
		500096	501314
		663932	669104
		742331	765181
		891347	891461
		891844	892246
		945355	945466
		946062	946177
		997717	997943
		1297304	1298479
		1306451	1306753
		1306756	1308131
		1325927	1326704
		1326758	1327813
		1368875	1370093
		1458408	1458635
		1661021	1661157
		1730085	1731315
		1839975	1842085
		1882338	1883397
		2031141	2032825
		2127264	2128547
		2147695	2149152
		2181298	2181558
		2181560	2181712
		2181714	2181929
		2181931	2182589
		2237462	2238859
		2293902	2294918
		2405818	2405952
		2406054	2406314
		2406316	2406468
		2406470	2406685
		2406687	2407345
		2442174	2443374
		2443469	2444469
		2450488	2450654
		2452057	2452189
		2452787	2452948
		2453215	2453345
		2459385	2459551
		2460954	2461086
		2461681	2461842
		2462109	2462239

		2468028	2468136
		2468138	2468471
		2468480	2469302
		2469304	2469410
		2469412	2469566
		2472840	2472948
		2472950	2473283
		2473292	2474114
		2474116	2474222
		2474224	2474378
		2482888	2482987
		2530991	2532845
		2714931	2715044
		2766175	2766288
		2787145	2787279
		2790313	2791513
		2861756	2862031
		3006312	3006411
		3027839	3027952
		3028949	3029064
		3030402	3031601
		3046532	3047812
		3060384	3060509
		3117702	3117917
		3224584	3224819
		3227635	3234158
		3412167	3413037
		3413067	3414198
		3482481	3482698
		3482700	3483289
		3483516	3483733
		3483735	3484324
Alcoy (ST578)	MGEs	68252	87883
		181578	230836
		609680	648761
		713383	714258
		957527	958980
		1163221	1331608
		1366871	1369020
		1404437	1405972
		1566257	1567378
		1951210	1953250
		1975809	1976930
		2077930	2080183
		2206613	2207192

		2240194	2253489
		2275132	2276259
		2401979	2404670
		2415705	2416739
		2421247	2421714
		2446243	2447118
		2486005	2509743
		2684076	2685217
		2756697	2786025
		3003084	3016236
		3166939	3168060
		3196856	3197731
		3327199	3328339
		3380374	3387940
		3411607	3416013
		3441670	3442791
	Repetitive regions	408773	413725
		413782	413927
		414516	414676
		414678	415612
		430799	430959
		430961	431895
		690387	690497
		690500	692117
		692252	693057
		693059	695557
		713359	714383
		772045	786054
		774047	774161
		774214	780482
		780487	782171
		966340	966496
		1299798	1301846
		1315695	1315938
		1316687	1316792
		1566162	1567389
		1919285	1919391
		1975799	1978249
		1992976	1993850
		1997483	1998357
		2249330	2250620
		2262945	2263070
		2275108	2277151
		2446116	2448162
		2470939	2471621

		2471623	2472333
		2472335	2472529
		2472531	2473605
		2473607	2473902
		2473904	2474091
		2474290	2474511
		2479839	2480521
		2480523	2481233
		2481235	2481429
		2481431	2482505
		2482507	2482802
		2482804	2482991
		2483190	2483411
		2510167	2510274
		2668380	2668580
		2668631	2668831
		2683982	2684989
		2684991	2685177
		2756963	2757078
		2757687	2757812
		2757814	2758321
		2758323	2758704
		2781518	2781623
		2781795	2782038
		2975051	2975175
		2985013	2985112
		3014335	3014448
		3015445	3016793
		3044504	3044660
		3100292	3100400
		3166844	3169295
		3196730	3198777
		3211309	3211454
		3212855	3217033
		3327105	3328111
		3328113	3328299
		3441576	3444026
		3466052	3466794
		3467087	3467829

Appendix Table 7. Genes in recombination hotspots in the six major disease-associated STs.

Gene	Number of recombination events	Product/function
lpp0019	4	hypothetical protein. Similar to Legionella zinc metalloproteinase precursor
lpp0020	4	hypothetical protein. Putative integral membrane protein
lpp0021	4	hypothetical protein. Similar to conserved hypothetical protein
lpp0022	5	hypothetical protein. Similar to conserved hypothetical protein
lpp0023	4	hypothetical protein. Putative membrane protein
lpp0024	4	hemin binding protein
lpp0356	4	hypothetical protein. Protein with ankyrin motif
lpp0819	4	N-acylglucosamine 2-epimerase
lpp0820	4	hypothetical protein. Similar to acetyl transferase
lpp0821	4	hypothetical protein. Similar to polysaccharide biosynthesis protein
lpp0822	4	dTDP-4-dehydrorhamnose 3,5-epimerase
lpp0823	4	dTDP-4-keto-L-rhamnose reductase
lpp0824	4	dTDP-D-glucose 4,6-dehydratase
lpp0825	5	glucose-6-phosphate isomerase
lpp0826	5	glucose-1-phosphate thymidylyltransferase
lpp0827	5	hypothetical protein. Similar to NAD dependent epimerase/dehydratase family protein
lpp0828	7	alpha-N-acetylglucosaminyltransferase
lpp0829	7	hypothetical protein
lpp0830	6	hypothetical protein
lpp0961	4	hypothetical protein. Similar to conserved hypothetical protein
lpp0962	4	hypothetical protein
lpp0963	4	hypothetical protein
lpp1640	4	hypothetical protein
lpp1641	4	hypothetical protein, alpha-amylase
lpp1642	3	hypothetical protein
lpp1643	4	hypothetical protein, alpha-amylase
lpp1644	4	Phosphoribosylglycinamide formyltransferase
lpp1645	4	Phosphoribosylamine-glycine ligase
lpp1761	7	hypothetical protein
lpp1762	8	hypothetical protein
lpp1763	13	alanyl-tRNA synthetase
lpp1764	12	Regulatory protein RecX
lpp1765	13	RecA protein
lpp1766	14	hypothetical protein
lpp1767	15	hypothetical protein
lpp1768	17	DNA mismatch repair protein MutS

lpp1769	18	hypothetical protein
lpp1770	25	hypothetical protein
lpp1771	27	hypothetical protein. Similar to delta-aminolevulinic acid dehydratases (porphobilinogen synthase)
lpp1772	25	hypothetical protein
lpp1773	25	hypothetical protein. Similar to long-chain fatty acid transport protein
lpp1774	24	hypothetical protein. Similar to diaminopimelate decarboxylase, aspartate kinase (fusion of lysA and lysC)
lpp1775	19	hypothetical protein. Similar to UvrD/REP helicase family protein
lpp1776	15	hypothetical protein. Similar to unknown protein
lpp1777	14	hypothetical protein. Similar to conserved hypothetical protein
lpp1778	14	Hydrogen peroxide-inducible genes activator
lpp1779	13	hypothetical protein. Similar to major facilitator family transporter
lpp1780	13	hypothetical protein
lpp1781	9	hypothetical protein. Similar to tetraacyldisaccharide 4'-kinase
lpp1782	9	lipid A export ATP-binding/permease protein MsbA
lpp1783	8	hypothetical protein
lpp1784	9	dihydroorotate dehydrogenase
lpp1785	8	hypothetical protein. Predicted transmembrane protein
lpp1786	7	hypothetical protein. Similar to conserved hypothetical protein
lpp1787	6	hypothetical protein. Similar to acyl-CoA dehydrogenase
lpp1788	5	hypothetical protein. Similar to acetyl-CoA acetyltransferase
lpp1789	5	hypothetical protein.
lpp1790	5	hypothetical protein. Similar to Acetyl/propionyl-CoA carboxylase, beta subunit
lpp1791	5	hypothetical protein. Similar to enoyl-CoA hydratase/isomerase
lpp1792	5	hypothetical protein. Similar to Acetyl/propionyl-CoA carboxylase, alpha subunit
lpp1793	5	hypothetical protein. Similar to hydroxymethylglutaryl-CoA lyase
lpp1794	4	hypothetical protein. Similar to acetyl-coenzyme A synthetase
lpp2198	4	hypothetical protein
lpp2543	4	hypothetical protein. Similar to glycosyl transferase
lpp2544	5	hypothetical protein. Similar to conserved hypothetical protein
lpp2545	4	hypothetical protein. Integral membrane protein, similar to metabolite efflux pump
lpp2546	4	SdbB protein (putative substrate of the Dot/Icm system).
lpp2547	4	hypothetical protein. Similar to hypothetical protein
lpp2548	4	hypothetical protein. Similar to conserved hypothetical protein
lpp2549	4	hypothetical protein. Protein with TPR motifs (protein-protein interaction motif)
lpp2550	4	phosphomannomutase
lpp2595	5	phospho-2-dehydro-3-deoxyheptonate aldolase

lpp2596	5	hypothetical protein. Similar to chorismate mutase (N-terminal part)
lpp2597	5	hypothetical protein. Similar to chorismate mutase (C-terminal part)
lpp2598	5	hypothetical protein. Similar to aspartate aminotransferase
lpp2599	6	hypothetical protein. Similar to tellurite resistance protein TehB
lpp2600	5	hypothetical protein
lpp2601	5	hypothetical protein. Similar to hemoglobin (protozoan/cyanobacterial globin family)
lpp2602	5	hypothetical protein. Similar to xylene monooxygenase
lpp2603	5	hypothetical protein. Similar to conserved hypothetical protein
lpp2604	5	hypothetical protein
lpp2977	4	hypothetical protein. Highly similar to peptide methionine sulfoxide reductase
lpp2978	4	hypothetical protein. Similar to hypothetical protein
lpp2979	4	hypothetical protein. Similar to copper amine oxidase
ST23_00399	2	protease HtpX homolog, heat shock protein HtpX, putative Zn-dependent protease, contains TPR repeats, peptidase family M48.
ST23_00400	2	inner membrane transport permease yadH, ,daunorubicin resistance, ABC transporter membrane protein
ST23_00401	2	daunorubicin/doxorubicin resistance ATP-binding protein DrrA, nodulation ABC transporter NodI, daunorubicin resistance ABC transporter
ST23_00402	2	hypothetical protein
ST23_00403	2	predicted proline hydroxylase
ST23_00404	2	protein of unknown function DUF45
ST23_00405	2	hypothetical protein
ST23_00406	2	hypothetical protein
ST23_00407	2	Methylated-DNA--protein-cysteine methyltransferase
ST23_00408	2	50S ribosomal protein L19
ST23_00409	2	tRNA (guanine-N(1)-)-methyltransferase
ST23_00410	2	21K,16S rRNA-processing protein RimM
ST23_00411	2	30S ribosomal protein S16
ST23_00412	2	p48, signal recognition particle protein
ST23_00413	2	hypothetical protein
ST23_00414	2	hypothetical protein
ST23_00415	2	ribulose-5-phosphate 4-epimerase and related epimerases and aldolases, ankyrin repeats (3 copies)
ST23_00416	2	glutamate/gamma-aminobutyrate antiporter
ST23_00417	2	hypothetical protein
ST23_00625	3	Carboxylate-amine ligase YbdK
ST23_00626	3	acetyl coenzyme A synthetase (ADP forming)
ST23_00647	2	thymidine kinase
ST23_00648	2	D-glucarate permease, regulatory protein UhpC, major facilitator superfamily
ST23_00703	2	hypothetical protein
ST23_00704	2	Proline--tRNA ligase
ST23_00705	2	ribulose-5-phosphate 4-epimerase and related epimerases

		and aldolases, transient-receptor-potential calcium channel protein, ankyrin repeats (3 copies)
ST23_00706	2	hypothetical protein
ST23_00707	2	carbonic anhydrase, sulfate transporter family
ST23_00708	2	tRNA 2-thiocytidine biosynthesis protein TtcA, predicted ATPase of the PP-loop superfamily implicated in cell cycle control
ST23_00709	2	outer membrane protein tolC precursor, outer membrane efflux protein.
ST23_00710	2	protein-L-isoaspartate O-methyltransferase
ST23_00711	2	2-amino-3-ketobutyrate coenzyme A ligase
ST23_00712	2	L-threonine 3-dehydrogenase
ST23_00713	2	uncharacterized ABC transporter, ChvD family
ST23_01779	2	DNA mismatch repair protein mutS
ST23_01780	2	outer membrane protein assembly factor YaeT
ST23_01781	3	hypothetical protein (DUF490)
ST23_01931	2	patatin-like phospholipase
ST23_01932	2	uncharacterized protein conserved in bacteria
ST23_01933	2	low-affinity cAMP phosphodiesterase
ST23_01934	2	hypothetical protein
ST23_01935	2	hypothetical protein
ST23_01936	2	hypothetical protein
ST23_01937	2	hypothetical protein
ST23_01938	2	Sec7 domain-containing protein
ST23_01939	2	putative lipid kinase BmrU
ST23_01940	2	cyclic 3',5'-adenosine monophosphate phosphodiesterase
ST23_01941	2	hypothetical protein
ST23_01942	2	hypothetical protein
ST23_01943	2	H+/gluconate symporter and related permeases
ST23_01944	2	L-Ala-D/L-Glu epimerase
ST23_01945	2	hypothetical protein
ST23_01946	2	hypothetical protein
ST23_01947	2	hypothetical protein
ST23_01990	2	Tfp pilus assembly protein PilW
ST23_02606	2	ankyrin repeats (3 copies)
ST23_02607	2	hypothetical protein
ST23_02608	2	hypothetical protein
ST23_02609	2	putative acyltransferase, GNAT family
ST23_02610	2	hypothetical protein
ST23_02611	2	hypothetical protein
ST23_02612	2	Response regulator rcp1
ST23_02613	2	phytochrome-like protein cph1, sensory histidine kinase AtoS, predicted periplasmic ligand-binding sensor domain, phosphate regulon sensor kinase PhoR, histidine kinase-, DNA gyrase B-, and HSP90-like ATPase
ST23_02614	2	heme NO binding
ST23_02615	2	hypothetical protein
ST23_02616	2	hypothetical protein

ST23_02617	2	methyltransferase domain
ST23_03044	2	hypothetical protein
ST23_03045	2	7-cyano-7-deazaguanine synthase, queuosine biosynthesis protein QueC, asparagine synthase (glutamine-hydrolyzing)
ST23_03046	2	alginate biosynthesis protein AlgA, mannose-1-phosphate guanyltransferase
ST37_01205	2	recombination-associated protein rdgC
ST37_01206	2	potassium transport protein Kup
ST42_02559	2	cytosol aminopeptidase, multifunctional aminopeptidase A
ST42_02560	3	hypothetical protein, integral membrane protein MviN
ST42_02561	3	30S ribosomal protein S20
ST42_02562	3	hypothetical protein
ST42_02563	3	hypothetical protein
ST42_02564	3	hypothetical protein, contains Sel1 repeat (EnhC)
ST42_02565	4	hypothetical protein
ST42_02566	3	hypothetical protein, L,D-transpeptidase catalytic domain
ST42_02567	3	Cyclic di-GMP phosphodiesterase Gmr, RNase II stability modulator, MHYT domain (predicted integral membrane sensor domain)
ST62_00255	2	protein of unknown function (DUF2878)
ST62_00256	4	hypothetical protein, EDD domain protein, DegV family
ST62_00257	4	2-(S)-hydroxypropyl-CoM dehydrogenase, 3-ketoacyl-(acyl-carrier-protein) reductase
ST62_00258	4	deoxyribodipyrimidine photo-lyase-related protein
ST62_00259	2	predicted membrane protein (DUF2177)
ST62_00260	2	serine/threonine-protein kinase PrkC
ST62_00261	2	hypothetical protein
ST62_00262	2	hypothetical protein
ST62_00263	2	TspO/MBR family
ST62_00264	2	deoxyribodipyrimidine photo-lyase
ST62_00265	2	inner membrane protein yohK, cytidylyltransferase, LrgB-like family
ST62_00266	2	antiholin-like protein LrgA
ST62_00267	2	Cyn operon transcriptional activator, DNA-binding transcriptional regulator CynR
ST62_00277	2	outer membrane efflux protein
ST62_00287	2	glutathione-dependent formaldehyde-activating enzyme
ST62_00288	2	hypothetical protein
ST62_00289	2	putative non-heme bromoperoxidase BpoC, acetoin dehydrogenase E2 subunit dihydrolipoyllysine-residue acetyltransferase, esterase/lipase,3-oxoadipate enol-lactonase, alpha/beta hydrolase family
ST62_00290	2	betaine aldehyde dehydrogenase
ST62_00291	2	4-aminobutyrate aminotransferase GabT
ST62_00292	2	hypothetical protein
ST62_00754	2	2-amino-3-ketobutyrate coenzyme A ligase
ST62_00755	2	L-threonine 3-dehydrogenase
ST62_00756	2	Uncharacterized ABC transporter, ChvD family
ST62_00757	2	hypothetical protein, L,D-transpeptidase catalytic domain

ST62_00758	2	predicted transporter component, YeeE/YedE family (DUF395)
ST62_00759	2	predicted transporter component, YeeE/YedE family (DUF395)
ST62_00760	2	outer membrane protein transport protein (OMPP1/FadL/TodX)
ST62_00761	2	macrophage killing protein with similarity to conjugation protein
ST62_00762	2	formimidoylglutamase
ST62_00763	2	benzil reductase, short chain dehydrogenase
ST62_00764	2	Imidazolonepropionase,imidazolonepropionase, cytosine deaminase and related metal-dependent hydrolases
ST62_00817	3	hypothetical protein
ST62_00823	2	hypothetical protein
ST62_01733	2	hypothetical protein
ST62_01734	2	DNA polymerase V subunit UmuC, nucleotidyltransferase/DNA polymerase involved in DNA repair
ST62_01735	2	DNA polymerase V subunit UmuD, repressor LexA, peptidase S24-like.
ST62_01736	2	carboxypeptidase G2 precursor, ArgE/DapE family, peptidase family M20/M25/M40
lpa_01248	2	ATP binding protease component
lpa_01249	2	lipopolysaccharide biosynthesis glycosyltransferase
lpa_01251	2	O-antigen biosynthesis protein
lpa_01252	2	hypothetical protein
lpa_01253	2	romboid family protein
lpa_01254	2	peptidase, M23/M37 family
lpa_01255	2	exodeoxyribonuclease VII large subunit
lpa_01256	2	agglutination protein
lpa_01258	2	predicted periplasmic protein
lpa_01261	2	two component histidine kinase
lpa_01262	2	hypothetical protein
lpa_01264	2	flavin containing monooxygenase
lpa_01265	2	short-chain dehydrogenase of various substrate specificities
lpa_01266	2	indole-3-glycerol phosphate synthase
lpa_01267	2	anthranilate phosphoribosyltransferase
lpa_01268	2	anthranilate synthase component II
lpa_01269	2	ABC-type transport system protein involved in lipoprotein release
lpa_01270	2	putative protein conserved in bacteria
lpa_01271	2	putative protein conserved in bacteria
lpa_01272	2	hydrolase, HAD superfamily, low specificity phosphatase
lpa_01273	2	polysialic acid capsule expression protein, predicted sugar phosphate isomerase involved in capsule formation
lpa_01289	2	putative conserved protein
lpa_02154	2	potassium efflux system protein KefA
lpa_04035	2	glucose/sorbosone dehydrogenase
lpa_04036	2	polyribonucleotide nucleotidyltransferase
lpa_04037	2	small subunit ribosomal protein S15

lpa_04038	2	tRNA pseudouridine synthase B
lpa_04039	2	ribosome-binding factor A
lpa_04041	2	translation initiation factor 2 (GTPase)
lpa_04042	2	N utilization substance protein A
lpa_04043	2	putative protein conserved in bacteria
lpa_04044	2	NADH dehydrogenase I chain N
lpa_04046	2	NADH dehydrogenase I chain M
lpa_04047	2	NADH dehydrogenase I chain L
lpa_04048	2	NADH dehydrogenase I chain K
lpa_04049	2	NADH dehydrogenase I chain J
lpa_04050	2	NADH dehydrogenase I chain I
lpa_04051	2	NADH dehydrogenase I chain H
lpa_04052	2	NADH dehydrogenase I chain G
lpa_04053	2	NADH dehydrogenase I chain F
lpa_04055	2	NADH dehydrogenase I chain E
lpa_04056	2	NADH dehydrogenase I chain D
lpa_04057	2	NADH dehydrogenase I chain C
lpa_04058	2	NADH dehydrogenase I chain B
lpa_04060	2	NADH dehydrogenase I chain A
lpa_04061	2	preprotein translocase SecG subunit
lpa_04062	2	triosephosphate isomerase (TIM)
lpa_04063	2	interaptin

9.3 Chapter 5

Appendix Table 8. The ESGLI standard typing panel of 106 isolates of *L. pneumophila* sg1 from 10 European countries, comprising epidemiologically "unrelated" and "related" panels. ST – sequence type; U/k - unknown

EUL no.	Country of origin	Isolation date	ST	Related strain	Evidence of relatedness	Accession number
Epidemiologically "unrelated" panel (*n*=79)						
1	Switzerland	01/02/1998	1			ERR376626
2	Switzerland	01/12/1989	2			ERR376627
3	Switzerland	01/10/1989	1			ERR376628
4	Switzerland	01/01/1991	23			ERR376721
6	Switzerland	01/01/1999	42*			ERR376631
7	Switzerland	01/05/1992	18			ERR376632
8	Switzerland	01/08/1993	23			ERR376633
13	Scotland	01/01/1983	5			ERR376646

14	Scotland	06/06/1984	5			ERR376639
16	Scotland	06/06/1984	5			ERR376641
17	Scotland	01/01/1993	7			ERR376642
18	Scotland	01/01/1994	26			ERR376643
19	Scotland	01/01/1994	9			ERR376644
20	Scotland	01/01/1995	28			ERR376645
25	France	01/01/1994	44			ERR376650
26	France	U/k	22			ERR376651
27	France	U/k	42			ERR376652
28	France	01/01/1994	23			ERR376722
29	France	01/01/1994	20*			ERR376654
30	France	U/k	38			ERR376655
31	France	01/01/1994	47			ERR376656
32	France	01/01/1994	16			ERR376657
33	France	U/k	40			ERR376658
36	Italy	01/01/1999	21			ERR332122
37	Italy	01/01/1999	1			ERR376723
38	Italy	01/01/1999	1*			ERR376663
39	Italy	01/01/1999	42			ERR376664
40	Italy	01/01/1999	12			ERR376665
41	Italy	01/01/1999	23			ERR376666
42	Italy	01/01/1999	1			ERR376667
43	Italy	01/01/1999	1			ERR376668
48	Spain	01/03/1996	48			ERR332134
49	Spain	01/02/1996	20			ERR376724
50	Spain	01/03/1996	42*			ERR376675
51	Spain	01/11/1995	1156*			ERR376676
52	Spain	01/09/1995	107*			ERR376677
53	Spain	01/05/1995	1*			ERR376725
54	Spain	01/02/1994	62			ERR376679
55	Spain	01/04/1994	1*			ERR332141
60	Greece	01/01/1992	1			ERR376685
63	Greece	01/01/1993	77*			ERR332149
66	Greece	01/01/1986	77*			ERR376728
67	Greece	01/01/1995	1			ERR376692
68	England and Wales	01/09/1995	46			ERR376693
69	England and Wales	21/11/1995	37			ERR376694
70	England and Wales	09/01/1996	47			ERR376695
71	England and Wales	10/05/1996	62			ERR332157
72	England and Wales	05/02/1996	4			ERR332158
73	England and Wales	01/04/1996	37			ERR376698
74	England and	14/03/1995	29			ERR376729

	Wales					
75	England and Wales	09/01/1995	42			ERR376700
81	Denmark	28/03/1994	53			ERR376732
82	Denmark	29/08/1994	1			ERR376733
83	Denmark	01/02/1995	50			ERR376734
84	Denmark	03/04/1995	1			ERR376735
85	Denmark	01/05/1995	1			ERR376710
86	Denmark	01/09/1995	46			ERR332172
87	Denmark	02/10/1995	2122*			ERR376712
88	Denmark	11/10/1995	1			ERR332174
91	Denmark	01/10/1995	63			ERR376737
92	Denmark	11/06/1991	53			ERR376717
93	Denmark	19/10/1992	1			ERR332179
97	Sweden	16/06/1994	9			ERR376741
98	Sweden	01/01/1996	9			ERR376629
99	Sweden	01/01/1995	34			ERR376704
100	Sweden	01/01/1995	59			ERR376742
101	Sweden	01/01/1994	60			ERR376743
102	Sweden	01/01/1993	59			ERR376714
103	Sweden	01/01/1993	45			ERR376744
104	Sweden	01/01/1992	1*			ERR376745
105	Sweden	01/01/1991	42			ERR376746
110	Germany	01/01/1993	10			ERR376674
111	Germany	01/01/1981	25			ERR376749
114**	Germany	27/02/1995	7			ERR376752
116	Germany	01/05/1996	42			ERR376754
117	Germany	U/k	6			ERR376755
118	Germany	01/10/1989	36			ERR340981
119	Germany	U/k	1			ERR376757
120	Germany	01/01/1999	42			ERR376758
Epidemiologically "related" panel (*n*=44)						
Subdivision I ("definitely related")						
48	Spain	01/03/1996	48		Clinical isolate from patient	ERR332134
56	Spain	01/03/1996	48	EUL 48	Clinical isolate from same patient (15 days later)	ERR376726
71	England and Wales	10/05/1996	62		Clinical isolate from patient (sputum *via* direct culture)	ERR332157
76	England and Wales	10/05/1996	62	EUL 71	Clinical isolate from same patient (isolated *via* amoebae)	ERR376701
77	England and Wales	10/05/1996	62	EUL 71	Clinical isolate from same patient (isolated *via* faeces)	ERR376702

73	England and Wales	01/04/1996	37		Clinical isolates from the same patient - each is a single colony picked from the isolation plate	ERR376698
78	England and Wales	01/04/1996	37	EUL 73		ERR376730
79	England and Wales	01/03/1996	37	EUL 73		ERR376731
120	Germany	01/01/1999	42		Clinical isolate from patient	ERR376758
121	Germany	01/01/1999	42	EUL 120	Duplicate of EUL120	ERR376678
Subdivision II ("probably related")						
3	Switzerland	01/10/1989	1		Clinical isolate from patient	ERR376628
9	Switzerland	01/10/1989	1	EUL 3	Environmental isolate from water (spa-pool)	ERR376634
10	Switzerland	01/10/1989	1	EUL 3	Environmental isolate from water (spa-pool)	ERR376635
8	Switzerland	01/08/1993	23		Clinical isolate from patient	ERR376633
11	Switzerland	01/08/1993	23	EUL 8	Environmental isolate from water (rest-home)	ERR376636
12	Switzerland	01/03/1993	23	EUL 8	Environmental isolate from water (rest-home)	ERR376637
19	Scotland	01/01/1994	9		Clinical isolate from patient 1	ERR376644
22	Scotland	01/01/1994	9	EUL 19	Clinical isolate from patient 2 (same outbreak)	ERR376647
23	Scotland	01/01/1994	9	EUL 19	Clinical isolate from patient 3 (same outbreak)	ERR376648
24	Scotland	01/01/1994	9	EUL 19	Related environmental isolate	ERR332110
33	France	Unknown	40		Clinical isolate from patient	ERR376658
34	France	Unknown	40	EUL 33	Related environmental isolate	ERR376659
35	France	01/01/1996	40	EUL 33	Related environmental isolate	ERR376660
37	Italy	01/01/1999	1		Clinical isolate from patient 1	ERR376723
44	Italy	01/01/1999	1	EUL 37	Related environmental isolate	ERR376669
45	Italy	01/01/1999	72	EUL 37	Clinical isolate from patient 2	ERR376670
38	Italy	01/01/1999	1*		Clinical isolate from patient	ERR376663
46	Italy	01/01/1999	1	EUL 38	Related environmental isolate	ERR376671

40	Italy	01/01/1999	12		Clinical isolate from patient	ERR376665
47	Italy	01/01/1999	12	EUL 40	Related environmental isolate	ERR376672
54	Spain	01/02/1994	62		Clinical isolate from patient (hotel-associated)	ERR376679
57	Spain	01/01/1995	62	EUL 54	Environmental isolate from shower water of hotel	ERR376682
55	Spain	01/04/1994	1*		Clinical isolate from patient (nosocomial)	ERR332141
58	Spain	01/01/1994	1	EUL 55	Environmental isolate from shower water of hospital	ERR376683
51	Spain	01/11/1995	1156*		Clinical isolate from patient (hotel-associated)	ERR376676
59	Spain	01/01/1993	1156*	EUL 51	Environmental isolate from shower water of hotel	ERR376684
93	Denmark	19/10/1992	1		Clinical isolate from patient (hotel-associated)	ERR332179
94	Denmark	08/12/1992	1	EUL 93	Clinical isolate from a related patient	ERR376738
95	Denmark	21/01/1993	1	EUL 93	Related environmental isolate	ERR376739
81	Denmark	28/03/1994	53		Environmental isolate	ERR376732
96	Denmark	01/01/1994	53	EUL 81	Clinical isolate from patient (community-acquired)	ERR376740
97	Sweden	16/06/1994	9		Clinical isolate from patient (community-acquired)	ERR376741
106	Sweden	01/01/1994	9	EUL 97	Clinical isolate from same patient	ERR332181
107	Sweden	01/01/1994	9	EUL 97	Related environmental isolate	ERR376747

*The STs of all typing panel isolates were re-called using the latest SBT protocol (version 5.0) and from the whole genome assemblies. While all are concordant using these two methods, those marked with an asterisk are discordant with the originally designated ST, as assigned by older SBT protocols prior to the introduction of the sequence quality tool and using less optimal primers. In most of these isolates, just one allele has changed, although in some, up to three alleles have been re-designated.

**EUL 114 was used as a substitute for EUL 112, which yielded a different ST to that recorded (both *in silico* and *via* traditional SBT).

Appendix Table 9. An additional 229 clinical and environmental isolates used in the evaluation of the WGS-based methods. ST - sequence type; Sg - serogroup; TA - travel-associated; U/k - unknown

Isolate number	Country of origin	Date of isolation	ST	Sg	Related isolate	Evidence of relatedness	Accession number/ Reference
3 pairs of epidemiologically related non-sg1 isolates							
LC 202/ EUL 153	UK	17/12/1986	68	6		Clinical isolate from patient (nosocomial)	ERR376775
LC 206/ EUL 158	UK	01/12/1986	68	6	EUL 153	Related environment-al isolate	ERR376780
LC 569/ EUL 154	UK	21/05/1988	1326	8		Clinical isolate from patient (nosocomial)	ERR376776
LC 606/ EUL 155	UK	01/06/1988	1326	8	EUL 154	Related environment-al isolate	ERR376777
LC 384/ EUL 156	Belgium	13/07/1987	1362	10		Clinical isolate from patient (nosocomial)	ERR376778
LC 395/ EUL 159	Belgium	01/07/1984	1362	10	EUL 156	Related environment-al isolate	ERR352158
Point-source outbreak (Barrow-in-Furness, 2002)							
LC6379-1/ EUL 145	UK	09/08/2002	78	1		Environment -al isolate from cooling tower 2 pond (recently working)	ERR376769
LC6376	UK	09/08/2002	78	1	LC6379-1/ EUL 145	Environment -al isolate from cooling tower 1 pond (not recently working)	ERR376790
LC6382	UK	09/08/2002	78	1	LC6379-1/ EUL 145	Environment -al isolate from tower 2 water cascade (working)	ERR376792
LC6391	UK	09/08/2002	78	1	LC6379-1/ EUL 145	Environment -al isolate from tower 1 water cascade (not working)	ERR376793

LC6394	UK	09/08/2002	78	1	LC6379-1/ EUL 145	Environment -al isolate from tower 1 water cascade (not working) (different sample to one above)	ERR376794
LC6397	UK	12/08/2002	78	1	LC6379-1/ EUL 145	Clinical isolate from patient 1	ERR376795
LC6406	UK	01/08/2002	78	1	LC6379-1/ EUL 145	Clinical isolate from patient 2	ERR376796
LC6407	UK	15/08/2002	78	1	LC6379-1/ EUL 145	Clinical isolate from patient 3	ERR376797
LC6408	UK	12/08/2002	78	1	LC6379-1/ EUL 145	Clinical isolate from patient 4	ERR341023
LC6411	UK	01/08/2002	78	1	LC6379-1/ EUL 145	Clinical isolate from patient 5	ERR376799
LC6412	UK	15/08/2002	78	1	LC6379-1/ EUL 145	Clinical isolate from patient 6	ERR376800
LC6413	UK	15/08/2002	78	1	LC6379-1/ EUL 145	Clinical isolate from patient 7	ERR376801
LC6416	UK	01/08/2002	78	1	LC6379-1/ EUL 145	Clinical isolate from patient 8	ERR376802
LC6418	UK	15/08/2002	78	1	LC6379-1/ EUL 145	Clinical isolate from patient 9	ERR376804
LC6385	UK	09/08/2002	78	1	LC6379-1/ EUL 145	Environment -al isolate from tower 2 water cascade (working)	ERR352162
LC6388	UK	09/08/2002	78	1	LC6379-1/ EUL 145	Environment -al isolate from tower 2 water cascade (working)	ERR352163
LC6409	UK	15/08/2002	78	1	LC6379-1/ EUL 145	Clinical isolate from patient 10	ERR352164
LC6410	UK	15/08/2002	78	1	LC6379-1/ EUL 145	Clinical isolate from patient 11	ERR352165
Point-source outbreak (BBC, Portland Place, 1988)							
LC0537/ EUL 132	UK	01/05/1988	37	1		Clinical isolate from patient 1	ERR332168
LC0539/ EUL 133	UK	01/05/1988	37	1	LC0537/ EUL 132	Clinical isolate from patient 2	ERR332169
LC0540/ EUL 134	UK	01/05/1998	37	1	LC0537/ EUL 132	Clinical isolate from patient 3	ERR332170
LC0565	UK	01/05/1988	37	1	LC0537/	Clinical	ERR363880

					EUL 132	isolate from patient 4	
LC0583	UK	01/05/1988	37	1	LC0537/ EUL 132	Clinical isolate from patient 5	ERR363881
Point-source outbreak (Hereford, 2003)							
H034680033	UK	01/11/2003	37	1		Clinical isolate from patient 1	ERR1232479
H034680035/ EUL 165	UK	01/11/2003	37	1	H034680033	Clinical isolate from patient 2	ERR376785
H034690056/ EUL 166	UK	01/11/2003	37	1	H034680033	Environment -al isolate from site A cooling tower 1	ERR376786
H034800427	UK	01/11/2003	37	1	H034680033	Environment -al isolate from site A cooling tower 2	ERR1232480
H034980467	UK	01/11/2003	37	1	H034680033	Environment -al isolate from domestic spa pool	ERR1232481
Additional ST1 isolates							
Paris	France	U/k	1	1			Cazalet *et al.* (2004)
H034800423	UK	01/11/2003	1	1			Reuter *et al.* (2013)
OLDA1 (NCTC12008)	USA	01/01/1947	1	1			ERR434061
EUL 109	Sweden	01/01/1992	1	1			ERR376662
Additional ST37 isolates							
H064240448	UK	12/10/2006	37	1			ERR363849
LC0731	UK	01/02/1989	37	1		Clinical isolate from patient 1	ERR363882
LC0732	UK	01/02/1989	37	1	LC0731	Clinical isolate from patient 2	ERR363883
LC0763	UK	01/02/1989	37	1	LC0731	Related environment- al isolate	ERR363884
LC0782	UK	01/02/1989	37	1	LC0731	Clinical isolate from patient 3	ERR363885
LC0795	UK	01/02/1989	37	1	LC0731	Clinical isolate from patient 4	ERR363887
LC0801	UK	01/02/1989	37	1	LC0731	Clinical isolate from patient 5	ERR363889
LC5694	UK	12/07/2000	37	1			ERR363891
LC5722	UK	31/08/2000	37	1			ERR363892
LC5738	UK	05/10/2000	37	1			ERR363893
LC5755	UK	01/11/2000	37	1			ERR363894
LC6163	UK	15/02/2002	37	1			ERR363897
LC6267	UK	10/07/2002	37	1			ERR363899

LC6268	UK	05/07/2002	37	1			ERR363900
LC6228	UK	10/04/2002	37	1			ERR363898
H041380048	UK	30/04/2004	37	1		Clinical isolate from patient	ERR363843
H041640791	UK	12/04/2004	37	1	H041380048	Related environment-al isolate	ERR363844
H042960010	UK	10/08/2004	37	1			ERR363845
H061140013	UK	19/04/2006	37	1			ERR363847
H071880001	UK	08/06/2007	37	1			ERR363850
H073060003	UK	30/08/2007	37	1			ERR363851
H080820009	UK	15/03/2008	37	1			ERR363853
LC6058	U/k (TA)	19/10/2001	37	1			ERR363896
LC6293	U/k (TA)	24/07/2002	37	1			ERR363901
LC6788	U/k (TA)	30/07/2003	37	1			ERR363902
H062660463	U/k (TA)	03/07/2006	37	1			ERR363848
H073900557	U/k (TA)	21/09/2007	37	1			ERR363852
LC1127	UK	26/12/1989	37	1			ERR363890
H084760449	UK	17/11/2008	37	1			ERR363857
H085020185	UK	15/12/2008	37	1			ERR363858
H090320386	UK	12/01/2009	37	1			ERR363859
H044260061	UK	11/10/2004	37	1			ERR363846
H093140322	UK	01/07/2009	37	1		Clinical isolate from patient	ERR363861
H093160422	UK	17/07/2009	37	1	H093140322	Related environment-al isolate	ERR363862
H092760433	U/k (TA)	06/07/2009	37	1			ERR363860
H100940111	UK	08/03/2010	37	1			ERR363863
H101760092	UK	03/05/2010	37	1			ERR363864
H101820190	UK	11/05/2010	37	1			ERR363865
H102020414	UK	24/05/2010	37	1			ERR363867
H101980130	U/k (TA)	17/05/2010	37	1			ERR363866
H103820081	UK	24/09/2010	37	1			ERR363868
H120240685	Slovenia	15/09/2010	37	1			ERR363992
H104320293	UK	26/10/2010	37	1			ERR363869
H113180118	UK	01/08/2011	37	1		Clinical isolate from patient	ERR363871
H113340664	UK	05/08/2011	37	1	H113180118	Related environment-al isolate	ERR363873
H113280076	UK	05/08/2011	37	1			ERR363872
H113660550	UK	12/09/2011	37	1			ERR363874
H114740454	UK	20/11/2011	37	1			ERR363876
H115040456	UK	11/12/2011	37	1			ERR363877
H111580389	UK	18/04/2011	37	1			ERR363870
H113780240	U/k (TA)	19/08/2011	37	1			ERR363875
H083920177	UK	26/09/2008	37	1		Clinical isolate from patient	ERR363855
H084140691	UK	03/10/2008	37	1	H083920177	Related	ERR363856

						environment-al isolate	
H081180019	UK	11/03/2008	37	1			ERR363854
H103260667	Greece	16/08/2010	37	1			ERR363938
LC464	UK	01/11/1987	37	1			ERR363878
LC0512	U/k (TA)	01/01/1988	37	1			ERR363879
LC0794	UK	01/02/1989	37	1		Clinical isolate from patient 1	ERR363886
LC0798	UK	01/02/1989	37	1	LC0794	Clinical isolate from patient 2	ERR363888
LC0536/EUL 131	UK	01/05/1988	37*	1			ERR332167
Additional ST42 isolates							
LC230/ EUL 122	U/k	01/03/1987	42	1		Clinical isolate from patient, isolated *via* direct plating	ERR376759
LC231/ EUL 123	U/k	01/03/1987	42	1	LC230/ EUL 122	Isolate from same patient, isolated *via* amoebal enrichment	ERR332142
LC0462/ EUL 124	UK	01/11/1987	42	1		Clinical isolate from patient, isolated *via* direct plating	ERR332150
LC0463/ EUL 125	UK	01/11/1987	42	1	LC0462/ EUL 124	Isolate from same patient, isolated *via* amoebal enrichment	ERR376760
Additional ST47 isolates							
Lorraine	France	20/08/2004	47	1			Gomez-Valero *et al.* (2011)
H063920004/ EUL 169	UK	25/09/2006	47	1		Clinical isolate from patient	Underwood *et al.* (2013)
H064160534/ EULV0410	UK	10/10/2006	47	1	H063920004 /EUL 169	Environment-al isolate from swimming pool	ERR363994
H064160538/ EUL 170	UK	10/10/2006	47	1	H063920004 /EUL 169	Environment-al isolate from spa pool (attached to swimming pool)	ERR376788
H034700617	UK	20/11/2003	47	1			Reuter *et al.* (2013)
H043580159	UK	01/09/2004	47	1			ERR363943
H043580160	UK	01/09/2004	47	1			ERR363959
H043660021	UK	01/09/2004	47	1			ERR363946
H043680663	UK	01/09/2004	47	1			ERR363949
H043700021	UK	01/09/2004	47	1			ERR363944
H043790008	UK	01/09/2004	47	1			ERR363945

H052920051	UK	01/07/2005	47	1			ERR363961
H053540106	UK	01/08/2005	47	1			ERR363948
H063660005	UK	01/09/2006	47	1		Clinical isolate from patient 1	ERR363904
H063660006	UK	09/09/2006	47	1	H063660005	Clinical isolate from patient 2, clustered in time and space with patient 1	ERR363922
H063760006	UK	14/09/2006	47	1	H063660005	Clinical isolate from patient 3, clustered in time and space with patient 1	ERR363915
H063660009	UK	01/09/2006	47	1			ERR363911
H063680006	UK	10/09/2006	47	1		Clinical isolate from patient 1	ERR363918
H063680007	UK	10/09/2006	47	1	H063680006	Clinical isolate from patient 2, clustered in time and space with patient 1	ERR363913
H063740003	UK	01/09/2006	47	1			ERR363929
H063740018	UK	01/09/2006	47	1			ERR363906
H063780007	UK	01/09/2006	47	1		Clinical isolate from patient 1	ERR363934
H063780008	UK	01/09/2006	47	1	H063780007	Clinical isolate from patient 2, clustered in time and space with patient 1	ERR363916
H063860003	UK	21/09/2006	47	1			ERR363930
H063960001	UK	01/09/2006	47	1			ERR363928
LC5759	U/k (TA)	23/10/2000	47	1			ERR363995
H070420013	UK	26/02/2007	47	1			ERR363968
LC5822	UK	07/02/2001	47	1			ERR363996
H040260015	UK	10/02/2004	47	1			ERR363903
H055140095	UK	15/01/2006	47	1			ERR363947
H060780053	UK	12/03/2006	47	1			ERR363907
H061120064	UK	10/04/2006	47	1			ERR363914
H062840608	UK	15/08/2006	47	1			ERR363917
H062940111	UK	22/08/2006	47	1			ERR363919
H064320006	UK	20/11/2006	47	1			ERR363923
H064280005	UK	24/11/2006	47	1			ERR363924
H064380002	UK	22/11/2006	47	1			ERR363926
H064380001	UK	30/11/2006	47	1			ERR363921
H064560527	UK	12/12/2006	47	1			ERR363925

H064660638	UK	20/12/2006	47	1			ERR363964
H070160015	UK	07/02/2007	47	1			ERR363970
H071120010	UK	16/04/2007	47	1			ERR363931
H071360036	UK	02/05/2007	47	1			ERR363908
H072740002	UK	08/08/2007	47	1			ERR363935
H073000045	UK	26/08/2007	47	1			ERR363932
H073380007	UK	13/09/2007	47	1			ERR363940
H073600182	UK	06/10/2007	47	1			ERR363976
H073640185	UK	09/10/2007	47	1			ERR363933
H074960018	UK	02/01/2008	47	1			ERR363920
H080780059	UK	13/03/2008	47	1			ERR363910
H053840008	UK	01/10/2004	47	1			ERR363954
H072520002	UK	22/06/2007	47	1			ERR363927
H081340222	UK	29/03/2007	47	1			ERR363909
H082520613	UK	20/06/2008	47	1			ERR363912
H083120262	UK	01/08/2008	47	1			ERR363941
H083620580	UK	05/09/2008	47	1			ERR363936
H083960064	UK	29/09/2008	47	1			ERR363937
H084620118	UK	17/11/2008	47	1			ERR363939
H090140214	UK	05/01/2009	47	1			ERR363963
H090440226	UK	26/01/2009	47	1			ERR363966
H040960441	UK	19/02/2004	47	1			ERR363953
H041120007	UK	05/03/2004	47	1			ERR363942
H093480403	U/k (TA)	24/08/2009	47	1			ERR363973
H094340202	UK	26/10/2009	47	1			ERR363971
H095060125	UK	14/12/2009	47	1			ERR363972
H100140151	UK	18/01/2010	47	1			ERR363965
H100660110	UK	15/02/2010	47	1			ERR363962
H100700025	UK	19/02/2010	47	1			ERR363958
H103140121	UK	02/08/2010	47	1			ERR363967
H103620160	UK	10/09/2010	47	1			ERR363950
H103660126	UK	23/09/2010	47	1			ERR363974
H103660121	UK	11/09/2010	47	1			ERR363956
H104420240	UK	04/11/2010	47	1			ERR363957
H110480273	UK	03/01/2011	47	1			ERR363969
H112320437	UK	06/06/2011	47	1			ERR363951
H112080616	UK	23/05/2011	47	1			ERR363952
H112380374	UK	13/06/2011	47	1			ERR363960
H120160499	UK	09/01/2012	47	1			ERR363985
H120200371	UK	12/01/2012	47	1			ERR363984
H105140391	UK	28/12/2010	47	1			ERR363993
H121040204	UK	09/03/2012	47	1			ERR363982
H121420445	UK	03/04/2012	47	1			ERR363983
H102240357	UK	19/05/2010	47	1			ERR363955
H122500497	UK	21/06/2012	47	1			ERR363981
H122820408	U/k (TA)	06/07/2012	47	1			ERR363980
H123620597	UK	04/09/2012	47	1			ERR363979
H123840629	UK	21/09/2012	47	1			ERR363978

H123940534	UK	28/09/2012	47	1			ERR363975
H124920387	UK	06/12/2012	47	1			ERR363991
H131340777	UK	29/03/2013	47	1		Clinical isolate from patient	ERR363990
H131480353	UK	01/04/2013	47	1	H131340777	Related environment-al isolate	ERR363989
H131480354	UK	01/04/2013	47	1	H131340777	Related environment-al isolate	ERR363988
H131840211	UK	01/04/2013	47	1	H131340777	Related environment-al isolate	ERR363986
H131460248	UK	06/04/2013	47	1			ERR363987
H132140863	UK	24/05/2013	47	1			ERR364031
H053640534/ EUL 168	UK	01/09/2005	47	1			ERR352161
Additional ST62 isolates							
H064180002	UK	01/10/2006	62	1		Clinical isolate from patient	Underwood *et al.* (2013)
H064180019	UK	09/10/2006	62	1	H064180002	Related environment-al isolate	ERR364004
H043540106	U/k (TA)	01/08/2004	62	1			ERR363997
H044120014	Bulgaria	01/10/2004	62	1			ERR363999
H052780022	UK	01/07/2005	62	1			ERR363998
H054280040	UK	01/11/2005	62	1			ERR364028
H063680003	UK	01/09/2006	62	1			ERR364002
H063840008	UK	04/09/2006	62	1			ERR364001
H073660582	UK	01/09/2007	62	1			ERR364008
LC5804	UK	01/11/2000	62	1			ERR364029
H063760005	UK	10/10/2006	62	1			ERR364000
H064240003	UK	14/11/2006	62	1			ERR364005
H065040012	UK	07/01/2007	62	1			ERR364012
H070140635	UK	06/02/2007	62	1			ERR364011
H073020039	UK	28/08/2007	62	1			ERR364022
H073320399	UK	10/09/2007	62	1			ERR364010
H073440003	UK	18/09/2007	62	1			ERR364009
LC6009	U/k (TA)	26/07/2001	62	1			ERR364030
H083140015	UK	25/07/2008	62	1			ERR364007
H093400182	UK	31/07/2009	62	1			ERR364006
H094760070	UK	23/11/2009	62	1			ERR364003
H094800237	UK	26/11/2009	62	1			ERR364020
H110480715	UK	21/01/2011	62	1			ERR364018
H112840293	UK	13/07/2011	62	1			ERR364017
H114100406	Greece	13/10/2011	62	1			ERR364016
H120240362	UK	16/01/2012	62	1			ERR364025
H104640262	U/k (TA)	19/11/2010	62	1			ERR364019
H123140428	UK	31/07/2012	62	1			ERR364015
H123460520	UK	27/08/2012	62	1			ERR364014
H124360642	UK	27/10/2012	62	1			ERR364013
Pontiac-1	USA	01/07/1968	62	1			ERR1232478

*The ST of LC0536/EUL 131 has been re-designated as 37 (from 13), as determined using the latest SBT protocol (v. 5.0) and using the whole genome assembly.

Appendix Table 10. Quality metrics and accession numbers for all *de novo* assemblies (derived from Illumina data) used in this chapter.

EUL/isolate number	Assembly length (bp)	No. of contigs	N50 (bp)	Accession number
Typing panel				
1	3582272	43	221291	FJAR01000001-FJAR01000043
2	3467814	21	441390	FJAF01000001-FJAF01000021
3	3584140	42	168231	FJAN01000001-FJAN01000042
4	3682698	37	180190	FJBD01000001-FJBD01000037
6	3387307	32	248780	FJBM01000001-FJBM01000032
7	3516217	36	198712	FJAI01000001-FJAI01000036
8	3489430	37	188012	FJBU01000001-FJBU01000037
13	3606063	38	221291	FJBF01000001-FJBF01000038
14	3606338	42	168264	FJAG01000001-FJAG01000042
16	3605510	44	168255	FJBH01000001-FJBH01000044
17	3440178	39	219509	FJBJ01000001-FJBJ01000039
18	3229839	38	143054	FJAW01000001-FJAW01000038
19	3422384	46	197885	FJAL01000001-FJAL01000046
20	3348748	18	336934	FJBO01000001-FJBO01000018
25	3353388	19	298032	FJAO01000001-FJAO01000019
26	3330854	36	164985	FJAB01000001-FJAB01000036
27	3493198	24	250279	FJAY01000001-FJAY01000024
27 (replicate)	3493923	30	250250	FJNG01000001-FJNG01000030
28	3624059	37	204558	FJBP01000001-FJBP01000037
29	3547570	32	199980	FJBR01000001-FJBR01000032
30	3295143	19	401585	FJAE01000001-FJAE01000019
31	3541152	62	103738	FJAT01000001-FJAT01000062
32	3545771	38	242920	FJAD01000001-FJAD01000038
33	3294148	23	333325	FJAH01000001-FJAH01000023
33 (replicate)	3294679	23	338768	FJNK01000001-FJNK01000023
36	3507348	36	200628	FJAM01000001-FJAM01000036
37	3446273	42	235781	FJBN01000001-FJBN01000042
38	3570332	33	235792	FJBQ01000001-FJBQ01000033
39	3347795	20	264182	FJAU01000001-FJAU01000020
40	3438629	22	275955	FJAV01000001-FJAV01000022
41	3488391	34	188086	FJBW01000001-FJBW01000034
42	3581764	43	168249	FJAP01000001-FJAP01000043
43	3575213	37	168244	FJBL01000001-FJBL01000037
48	3503195	65	101284	FJBT01000001-FJBT01000065
49	3517591	57	237958	FJBE01000001-FJBE01000057
50	3385882	24	214368	FJBG01000001-FJBG01000024

51	3405363	15	726453	FJBB01000001-FJBB01000015
52	3318517	28	639347	FJBX01000001-FJBX01000028
53	3578730	41	168217	FJBY01000001-FJBY01000041
54	3453348	34	183458	FJBZ01000001-FJBZ01000034
55	3579917	37	168229	FJCA01000001-FJCA01000037
60	3583463	47	167804	FJCF01000001-FJCF01000047
63	3461713	24	281315	FJCG01000001-FJCG01000024
66	3461261	30	263172	FJCH01000001-FJCH01000030
67	3580420	38	168281	FJCI01000001-FJCI01000038
68	3353004	64	86373	FJCJ01000001-FJCJ01000064
69	3348891	19	525181	FJCK01000001-FJCK01000019
69 (replicate)	3348937	22	357199	FJNJ01000001-FJNJ01000022
70	3598684	68	103511	FJCL01000001-FJCL01000068
71	3485244	38	196616	FJCM01000001-FJCM01000038
72	3346824	19	437858	FJCN01000001-FJCN01000019
73	3349062	17	336718	FJCP01000001-FJCP01000017
74	3459400	27	713253	FJCQ01000001-FJCQ01000027
75	3386255	27	196971	FJCO01000001-FJCO01000027
75 (replicate)	3385427	26	214349	FJNI01000001-FJNI01000026
81	3514412	36	223875	FJCV01000001-FJCV01000036
82	3576637	38	262463	FJCX01000001-FJCX01000038
83	3515373	40	228374	FJCW01000001-FJCW01000040
84	3446240	36	263208	FJCY01000001-FJCY01000036
85	3577453	33	262466	FJCZ01000001-FJCZ01000033
86	3547658	72	101986	FJDC01000001-FJDC01000072
87	3422196	20	246513	FJDB01000001-FJDB01000020
88	3576785	34	262392	FJDA01000001-FJDA01000034
91	3266471	34	223981	FJDD01000001-FJDD01000034
92	3514819	42	216313	FJDE01000001-FJDE01000042
92 (replicate)	3513289	39	225203	FJNL01000001-FJNL01000039
93	3641343	43	262032	FJDF01000001-FJDF01000043
97	3465455	41	250633	FJDJ01000001-FJDJ01000041
98	3465680	42	248414	FJDK01000001-FJDK01000042
99	3259617	19	657238	FJDL01000001-FJDL01000019
100	3351397	34	183216	FJDO01000001-FJDO01000034
101	3429712	48	151004	FJDM01000001-FJDM01000048
102	3369225	39	176020	FJDN01000001-FJDN01000039
103	3669856	70	103529	FJDP01000001-FJDP01000070
104	3607517	47	155184	FJDQ01000001-FJDQ01000047
105	3383263	21	445776	FJDR01000001-FJDR01000021
110	3624710	43	168239	FJDU01000001-FJDU01000043
111	3298695	41	124718	FJDW01000001-FJDW01000041
111 (replicate)	3299082	47	126247	FJNH01000001-FJNH01000047
114	3440457	40	221541	FJDV01000001-FJDV01000040
116	3300129	16	444178	FJDX01000001-FJDX01000016
117	3437615	47	121106	FJDY01000001-FJDY01000047
118	3410805	25	356081	FJDZ01000001-FJDZ01000025
119	3571941	36	220433	FJEA01000001-FJEA01000036

120	3384999	21	264302	FJEB01000001-FJEB01000021
9	3583104	44	168236	FJBC01000001-FJBC01000044
10	3583674	37	262030	FJAZ01000001-FJAZ01000037
11	3488698	31	242931	FJBI01000001-FJBI01000031
12	3487861	31	215445	FJAJ01000001-FJAJ01000031
22	3436716	19	320105	FJBS01000001-FJBS01000019
23	3438368	21	336944	FJAC01000001-FJAC01000021
24	3436497	24	337270	FJAS01000001-FJAS01000024
34	3294395	21	334734	FJAX01000001-FJAX01000021
35	3294934	18	336503	FJBK01000001-FJBK01000018
44	3580987	41	235794	FJBV01000001-FJBV01000041
45	3453726	40	235793	FJBA01000001-FJBA01000040
46	3570783	37	168238	FJAQ01000001-FJAQ01000037
47	3438617	26	275143	FJAK01000001-FJAK01000026
56	3499528	63	103579	FJCC01000001-FJCC01000063
57	3452138	34	183315	FJCB01000001-FJCB01000034
58	3581148	43	169426	FJCE01000001-FJCE01000043
59	3406404	20	244116	FJCD01000001-FJCD01000020
76	3485883	39	162984	FJCR01000001-FJCR01000039
77	3486431	36	164472	FJCS01000001-FJCS01000036
78	3348065	25	324151	FJCT01000001-FJCT01000025
79	3349415	28	336667	FJCU01000001-FJCU01000028
94	3640935	37	262076	FJDH01000001-FJDH01000037
95	3641028	37	262384	FJDG01000001-FJDG01000037
96	3513757	37	223869	FJDI01000001-FJDI01000037
106	3464579	35	250633	FJDS01000001-FJDS01000035
107	3464340	39	249913	FJDT01000001-FJDT01000039
121	3385179	21	300906	FJEC01000001-FJEC01000021
Additional isolates				
LC 202/EUL 153	3370172	21	332679	FJED01000001-FJED01000021
LC 206/EUL 158	3369964	17	485867	FJEF01000001-FJEF01000017
LC 569/EUL 154	3416299	12	2134649	FJEG01000001-FJEG01000012
LC 606/EUL 155	3416417	12	1881974	FJEE01000001-FJEE01000012
LC 384/EUL 156	3487522	22	403021	FJEH01000001-FJEH01000022
LC 395/EUL 159	3482177	20	413838	FJEI01000001-FJEI01000020
LC6379-1/EUL 145	3365082	30	184547	FJEK01000001-FJEK01000030
LC6376	3365099	30	391540	FJEJ01000001-FJEJ01000030
LC6382	3364319	27	335444	FJEL01000001-FJEL01000027
LC6391	3363339	27	243038	FJEN01000001-FJEN01000027
LC6394	3364625	32	184559	FJEM01000001-FJEM01000032
LC6397	3363172	29	184544	FJEQ01000001-FJEQ01000029
LC6406	3363398	30	242960	FJEO01000001-FJEO01000030
LC6407	3362234	29	242960	FJEP01000001-FJEP01000029
LC6408	3362492	35	242960	FJER01000001-FJER01000035
LC6411	3363467	26	251479	FJEU01000001-FJEU01000026
LC6412	3362944	34	184542	FJES01000001-FJES01000034
LC6413	3362923	32	184530	FJET01000001-FJET01000032
LC6416	3362654	30	242960	FJEV01000001-FJEV01000030

LC6418	3362330	30	184783	FJEW01000001-FJEW01000030
LC6385	3364576	27	242794	FJEX01000001-FJEX01000027
LC6388	3363943	27	242791	FJEY01000001-FJEY01000027
LC6409	3364426	28	336094	FJEZ01000001-FJEZ01000028
LC6410	3365323	25	251329	FJFA01000001-FJFA01000025
LC0537/EUL 132	3412033	24	324159	FJFB01000001-FJFB01000024
LC0539/EUL 133	3413672	25	638011	FJFC01000001-FJFC01000025
LC0540/EUL 134	3412957	27	327654	FJFD01000001-FJFD01000027
LC0565	3413362	24	324339	FJFE01000001-FJFE01000024
LC0583	3414048	26	324530	FJFF01000001-FJFF01000026
H034680033	3446924	30	294935	FJOB01000001-FJOB01000030
H034680035/EUL 165	3444436	25	336672	FJFG01000001-FJFG01000025
H034690056/EUL 166	3445871	22	739607	FJFH01000001-FJFH01000022
H034800427	3446329	20	409792	FJNZ01000001-FJNZ01000020
H034980467	3446486	17	584207	FJNY01000001-FJNY01000017
H034800423	3557791	48	129964	FJOE01000001-FJOE01000323
OLDA1 (NCTC12008)	3586509	42	241020	FJFJ01000001-FJFJ01000042
EUL 109	3609634	42	155185	FJFI01000001-FJFI01000042
H064240448	3412426	18	788251	FJFK01000001-FJFK01000018
LC0731	3389456	23	356594	FJFM01000001-FJFM01000023
LC0732	3388952	20	435223	FJFL01000001-FJFL01000020
LC0763	3388509	19	337610	FJFN01000001-FJFN01000019
LC0782	3388588	22	474242	FJHM01000001-FJHM01000022
LC0795	3388538	22	434228	FJHO01000001-FJHO01000022
LC0801	3389651	20	337731	FJHR01000001-FJHR01000020
LC5694	3340104	20	308508	FJFO01000001-FJFO01000020
LC5722	3339668	27	304117	FJFP01000001-FJFP01000027
LC5738	3339758	22	248375	FJFQ01000001-FJFQ01000022
LC5755	3415172	26	336687	FJFR01000001-FJFR01000026
LC6163	3350195	23	248366	FJFS01000001-FJFS01000023
LC6267	3413801	24	324534	FJFT01000001-FJFT01000024
LC6268	3413910	22	414101	FJFU01000001-FJFU01000022
LC6228	3446506	21	324341	FJFV01000001-FJFV01000021
H041380048	3411003	22	336628	FJFW01000001-FJFW01000022
H041640791	3411081	20	336634	FJGD01000001-FJGD01000020
H042960010	3411388	19	417334	FJFX01000001-FJFX01000019
H061140013	3414012	22	337363	FJFY01000001-FJFY01000022
H071880001	3351157	26	317885	FJFZ01000001-FJFZ01000026
H073060003	3413611	23	336671	FJGB01000001-FJGB01000023
H080820009	3581505	27	656985	FJGA01000001-FJGA01000027
LC6058	3412804	25	549630	FJGC01000001-FJGC01000025
LC6293	3414379	22	1076411	FJGE01000001-FJGE01000022
LC6788	3443537	20	549813	FJGF01000001-FJGF01000020
H062660463	3351247	23	330279	FJGG01000001-FJGG01000023
H073900557	3494788	27	324154	FJGH01000001-FJGH01000027
LC1127	3389579	23	337256	FJGJ01000001-FJGJ01000023

H084760449	3445602	22	324339	FJGI01000001-FJGI01000022
H085020185	3446838	18	770488	FJGK01000001-FJGK01000018
H090320386	3409856	24	336675	FJGL01000001-FJGL01000024
H044260061	3445323	22	969495	FJGM01000001-FJGM01000022
H093140322	3445924	25	351298	FJGN01000001-FJGN01000025
H093160422	3445791	21	548795	FJGO01000001-FJGO01000021
H092760433	3461547	34	304114	FJGP01000001-FJGP01000034
H100940111	3411026	27	248357	FJGQ01000001-FJGQ01000027
H101760092	3350043	23	324616	FJGR01000001-FJGR01000023
H101820190	3411782	19	416839	FJGS01000001-FJGS01000019
H102020414	3445616	26	336678	FJGT01000001-FJGT01000026
H101980130	3351235	25	311133	FJGU01000001-FJGU01000025
H103820081	3482669	31	219507	FJGV01000001-FJGV01000031
H120240685	3431770	27	330426	FJGW01000001-FJGW01000027
H104320293	3413966	29	355130	FJGY01000001-FJGY01000029
H113180118	3341309	22	301691	FJGX01000001-FJGX01000022
H113340664	3341936	20	388532	FJHA01000001-FJHA01000020
H113280076	3380585	20	590371	FJGZ01000001-FJGZ01000020
H113660550	3423689	20	509386	FJHB01000001-FJHB01000020
H114740454	3413960	24	336679	FJHC01000001-FJHC01000024
H115040456	3413149	26	336680	FJHD01000001-FJHD01000026
H111580389	3446413	20	484400	FJHE01000001-FJHE01000020
H113780240	3414602	25	482710	FJHF01000001-FJHF01000025
H083920177	3441071	22	304113	FJHG01000001-FJHG01000022
H084140691	3410070	21	416599	FJHH01000001-FJHH01000021
H081180019	3348323	25	249077	FJHI01000001-FJHI01000025
H103260667	3453245	18	498080	FJHJ01000001-FJHJ01000018
LC464	3350265	22	324172	FJHK01000001-FJHK01000022
LC0512	3413673	22	638021	FJHL01000001-FJHL01000022
LC0794	3413957	26	324532	FJHN01000001-FJHN01000026
LC0798	3415031	24	364734	FJHP01000001-FJHP01000024
LC0536/EUL 131	3422513	35	468672	FJHQ01000001-FJHQ01000035
LC230/EUL 122	3443140	31	259364	FJHS01000001-FJHS01000031
LC231/EUL 123	3442071	28	264302	FJHU01000001-FJHU01000028
LC0462/EUL 124	3386214	18	323630	FJHT01000001-FJHT01000018
LC0463/EUL 125	3385613	28	304205	FJHV01000001-FJHV01000028
H063920004/ EUL 169	3540658	65	104090	FJHW01000001-FJHW01000065
H064160534/ EULV0410	3542236	61	103515	FJHY01000001-FJHY01000061
H064160538/ EUL 170	3542584	66	103231	FJLL01000001-FJLL01000066
H034700617	3535096	74	82988	FJOC01000001-FJOC01000074
H043580159	3540112	63	102126	FJHX01000001-FJHX01000063
H043580160	3543495	68	102118	FJHZ01000001-FJHZ01000068
H043660021	3539680	58	86360	FJIB01000001-FJIB01000058
H043680663	3542231	60	103521	FJIC01000001-FJIC01000060
H043700021	3572737	63	94275	FJIA01000001-FJIA01000063
H043790008	3540216	58	103518	FJIE01000001-FJIE01000058

H052920051	3538542	57	104627	FJID01000001-FJID01000057
H053540106	3541070	64	94274	FJIF01000001-FJIF01000064
H063660005	3540882	65	94015	FJIG01000001-FJIG01000065
H063660006	3540846	62	94015	FJIH01000001-FJIH01000062
H063760006	3542859	72	81300	FJIN01000001-FJIN01000072
H063660009	3540749	61	103554	FJII01000001-FJII01000061
H063680006	3595917	73	81272	FJIJ01000001-FJIJ01000073
H063680007	3539075	66	81274	FJIK01000001-FJIK01000066
H063740003	3538993	62	94742	FJIL01000001-FJIL01000062
H063740018	3540596	59	103556	FJIM01000001-FJIM01000059
H063780007	3540451	63	102072	FJIO01000001-FJIO01000063
H063780008	3544143	62	104160	FJIP01000001-FJIP01000062
H063860003	3526491	140	103528	FJIR01000001-FJIR01000140
H063960001	3538795	69	103540	FJIQ01000001-FJIQ01000069
LC5759	3537671	59	102058	FJIS01000001-FJIS01000059
H070420013	3541477	60	104720	FJIT01000001-FJIT01000060
LC5822	3540057	62	102059	FJIU01000001-FJIU01000062
H040260015	3541261	61	94015	FJIW01000001-FJIW01000061
H055140095	3539663	58	113585	FJIV01000001-FJIV01000058
H060780053	3541600	61	103528	FJIY01000001-FJIY01000061
H061120064	3542809	63	113625	FJIX01000001-FJIX01000063
H062840608	3540451	63	110289	FJIZ01000001-FJIZ01000063
H062940111	3538952	60	102071	FJJA01000001-FJJA01000060
H064320006	3538778	60	103523	FJJB01000001-FJJB01000060
H064280005	3540049	64	93845	FJJC01000001-FJJC01000064
H064380002	3541557	63	113617	FJJD01000001-FJJD01000063
H064380001	3540188	61	103511	FJJE01000001-FJJE01000061
H064560527	3539696	65	103511	FJJF01000001-FJJF01000065
H064660638	3541541	59	110131	FJJG01000001-FJJG01000059
H070160015	3541486	63	102059	FJJH01000001-FJJH01000063
H071120010	3540618	62	103558	FJJI01000001-FJJI01000062
H071360036	3541845	61	103511	FJJJ01000001-FJJJ01000061
H072740002	3540414	61	103550	FJJK01000001-FJJK01000061
H073000045	3538163	59	103505	FJJL01000001-FJJL01000059
H073380007	3540390	58	103525	FJJM01000001-FJJM01000058
H073600182	3542113	63	103521	FJJO01000001-FJJO01000063
H073640185	3540979	65	103550	FJJN01000001-FJJN01000065
H074960018	3540833	62	99193	FJJP01000001-FJJP01000062
H080780059	3540582	61	103556	FJJR01000001-FJJR01000061
H053840008	3541342	56	103521	FJJQ01000001-FJJQ01000056
H072520002	3541042	65	94271	FJJS01000001-FJJS01000065
H081340222	3541677	62	93140	FJJU01000001-FJJU01000062
H082520613	3541454	62	103525	FJJT01000001-FJJT01000062
H083120262	3540437	64	103519	FJJV01000001-FJJV01000064
H083620580	3540741	62	113625	FJJW01000001-FJJW01000062
H083960064	3539747	60	94015	FJJX01000001-FJJX01000060
H084620118	3537979	60	102059	FJJY01000001-FJJY01000060
H090140214	3541703	60	113601	FJJZ01000001-FJJZ01000060

H090440226	3540261	62	103514	FJKA01000001-FJKA01000062
H040960441	3540558	64	94267	FJKB01000001-FJKB01000064
H041120007	3541166	58	103178	FJKC01000001-FJKC01000058
H093480403	3541510	64	113613	FJKD01000001-FJKD01000064
H094340202	3538499	59	103523	FJKE01000001-FJKE01000059
H095060125	3539739	57	102062	FJKF01000001-FJKF01000057
H100140151	3539368	61	103525	FJKG01000001-FJKG01000061
H100660110	3540568	63	103524	FJKH01000001-FJKH01000063
H100700025	3541027	56	103510	FJKI01000001-FJKI01000056
H103140121	3540509	57	103518	FJKJ01000001-FJKJ01000057
H103620160	3536019	60	94248	FJKK01000001-FJKK01000060
H103660126	3540470	60	110131	FJKL01000001-FJKL01000060
H103660121	3541153	58	113585	FJKM01000001-FJKM01000058
H104420240	3541670	63	103514	FJKN01000001-FJKN01000063
H110480273	3540984	62	102059	FJKO01000001-FJKO01000062
H112320437	3540480	64	102060	FJKP01000001-FJKP01000064
H112080616	3541936	58	106775	FJKQ01000001-FJKQ01000058
H112380374	3539899	61	110282	FJKR01000001-FJKR01000061
H120160499	3542582	58	102061	FJKS01000001-FJKS01000058
H120200371	3541348	61	103519	FJKT01000001-FJKT01000061
H105140391	3540891	56	113609	FJKU01000001-FJKU01000056
H121040204	3540160	62	106993	FJKV01000001-FJKV01000062
H121420445	3540539	60	103511	FJKW01000001-FJKW01000060
H102240357	3540020	60	102182	FJKZ01000001-FJKZ01000060
H122500497	3541272	61	103526	FJKY01000001-FJKY01000061
H122820408	3540631	61	110131	FJKX01000001-FJKX01000061
H123620597	3390280	61	102055	FJLA01000001-FJLA01000061
H123840629	3539581	60	103523	FJLB01000001-FJLB01000060
H123940534	3542623	63	103518	FJLC01000001-FJLC01000063
H124920387	3541567	62	103510	FJLD01000001-FJLD01000062
H131340777	3540459	62	103526	FJLF01000001-FJLF01000062
H131480353	3659340	64	95680	FJLH01000001-FJLH01000064
H131480354	3710927	70	102060	FJLE01000001-FJLE01000070
H131840211	3660511	71	86362	FJLI01000001-FJLI01000071
H131460248	3541332	61	103521	FJLG01000001-FJLG01000061
H132140863	3538262	58	113601	FJLJ01000001-FJLJ01000058
H053640534/ EUL 168	3542142	63	103562	FJLM01000001-FJLM01000063
H064180002	3435971	33	129546	FJLK01000001-FJLK01000033
H064180019	3448814	33	182977	FJMF01000001-FJMF01000033
H043540106	3464056	39	182977	FJLN01000001-FJLN01000039
H044120014	3483235	32	182989	FJLP01000001-FJLP01000032
H052780022	3487895	39	172799	FJLQ01000001-FJLQ01000039
H054280040	3486918	37	176908	FJLO01000001-FJLO01000037
H063680003	3629395	35	182973	FJLR01000001-FJLR01000035
H063840008	3486102	33	238970	FJLS01000001-FJLS01000033
H073660582	3624857	31	255070	FJLT01000001-FJLT01000031
LC5804	3537086	31	183150	FJLU01000001-FJLU01000031
H063760005	3438575	33	177023	FJLV01000001-FJLV01000033

H064240003	3450470	39	183482	FJLW01000001-FJLW01000039
H065040012	3520886	39	176911	FJLY01000001-FJLY01000039
H070140635	3485846	34	183151	FJLZ01000001-FJLZ01000034
H073020039	3486953	39	176940	FJLX01000001-FJLX01000039
H073320399	3485633	31	182851	FJMA01000001-FJMA01000031
H073440003	3489140	34	182973	FJMB01000001-FJMB01000034
LC6009	3452667	37	176933	FJMC01000001-FJMC01000037
H083140015	3542056	34	199082	FJMD01000001-FJMD01000034
H093400182	3476832	33	199435	FJME01000001-FJME01000033
H094760070	3543856	34	185121	FJMJ01000001-FJMJ01000034
H094800237	3545764	33	236407	FJMG01000001-FJMG01000033
H110480715	3560000	38	225261	FJMH01000001-FJMH01000038
H112840293	3553030	35	247920	FJMI01000001-FJMI01000035
H114100406	3452106	33	199452	FJMK01000001-FJMK01000033
H120240362	3544187	37	227709	FJML01000001-FJML01000037
H104640262	3437410	31	183517	FJMM01000001-FJMM01000031
H123140428	3558699	36	181172	FJMN01000001-FJMN01000036
H123460520	3537604	35	183495	FJMO01000001-FJMO01000035
H124360642	3543070	36	176909	FJMP01000001-FJMP01000036
Pontiac-1	3473661	31	204700	FJOA01000001-FJOA01000031

Appendix Table 11. Reference genomes used in the SNP-based analysis. ST - sequence type; U/k - unknown

Reference name	**ST**	**Length of chromosome (bp)**	**Complete genome/PacBio assembly**	**Reference**
Paris	1	3503610	Complete	Cazalet *et al.* 2004
Lorraine	47	3467254	Complete	Gomez-Valero *et al.* 2011
Alcoy	678	3516334	Complete	D'Auria *et al.* 2010
Philadelphia	36	3397754	Complete	Chien *et al.* 2004
Lens	15	3345687	Complete	Cazalet *et al.* 2004
Corby	51	3576470	Complete	Glockner *et al.* 2008
LPE509	U/k*	3434224	Complete	Ma *et al.* 2013
ATCC 43290	187	3359001	Complete	Amaro *et al.* 2012
HL 0604 1035	734	3492535	Complete	Gomez-Valero *et al.* 2011
EUL 28	23	3509586	PacBio assembly (2 contigs: 1 chromosome, 1 plasmid)	This study
EUL 120	42	3430562	PacBio assembly (2 contigs: both chromosomal, 0 plasmids)	This study
EUL 165	37	3474638	PacBio assembly (1 contig: 1 chromosome, 0 plasmids)	This study
H044120014	62	3530817	PacBio assembly (1 contig: 1 chromosome, 0 plasmids)	This study

*The following allele numbers are called in LPE509: 3 (*flaA*), 10 (*pilE*), 1 (*asd*), 1 (*mip*), 9 (*proA*), and 1 (*neuA*). However, due to the presence of multiple copies of the *mompS* gene, this allele number cannot be determined *in silico*.

Appendix Table 12. Reference genomes used for the mapping of all isolates in this study and the coverage achieved.

EUL/isolate number	Reference	% reference length mapped	Mean depth of coverage	Standard deviation of depth of coverage
Typing panel				
1	Paris	98.3	136.4	14.9
2	EUL 2	99.9	136.1	26.4
3	Paris	98.4	138.7	15.4
4	EUL 28	97.0	95.9	20.2
6	EUL 120	98.0	135.7	25.8
7	EUL 7	100.0	145.6	39.7
8	EUL 28	97.0	151.4	24.6
13	Paris	97.2	150.1	23.6
14	Paris	97.2	142.9	21.5
16	Paris	97.2	162.3	23.5
17	Paris	97.8	161.8	21.6
18	EUL 18	100.0	164.2	48.2
19	Philadelphia	94.4	149.2	36.2
20	Philadelphia	94.1	152.1	34.6
25	EUL 25	100.0	156.1	43.1
26	EUL 26	100.0	154.3	44.3
27	EUL 120	98.2	144.8	17.9
27 (replicate)	EUL 120	98.2	98.8	13.6
28	EUL 28	99.0	99.2	12.3
29	EUL 49	98.8	150.8	53.8
30	Philadelphia	93.5	145.4	34.3
31	Lorraine	98.0	144.8	18.9
32	EUL 32	100.0	145.5	44.0
33	Philadelphia	93.4	137.1	35.0
33 (replicate)	Philadelphia	93.2	80.8	21.4
36	EUL 36	99.9	86.5	20.9
37	Paris	97.2	94.9	13.2
38	Paris	97.4	142.4	19.4
39	EUL 120	95.9	154.5	27.1
40	EUL 40	100.0	158.7	45.1
41	EUL 28	97.0	130.1	21.5
42	Paris	98.0	129.4	16.8
43	Paris	98.0	139.1	17.0

48	EUL 48	99.9	126.4	82.1
49	EUL 49	99.9	99.0	36.0
50	EUL 120	98.0	144.8	18.0
51	EUL 51	100.0	143.4	43.3
52	ATCC43290	94.3	140.5	34.3
53	Paris	98.1	79.5	11.1
54	H044120014	94.6	147.4	51.0
55	Paris	98.3	97.8	14.1
60	Paris	98.4	137.1	17.2
63	EUL 63	99.8	95.9	22.8
66	EUL 63	99.8	99.0	24.4
67	Paris	98.3	130.0	15.2
68	Lorraine	94.8	148.4	31.4
69	EUL 165	95.7	143.2	29.1
69 (replicate)	EUL 165	95.7	96.0	20.4
70	Lorraine	98.0	142.4	19.2
71	H044120014	95.9	97.9	21.1
72	EUL 2	94.8	89.3	26.7
73	EUL 165	95.7	145.3	31.4
74	Philadelphia	97.0	98.2	19.5
75	EUL 120	98.0	133.0	19.5
75 (replicate)	EUL 120	98.0	91.8	14.5
81	Corby	94.0	95.3	20.7
82	Paris	94.2	93.6	20.4
83	Corby	94.5	93.8	19.7
84	Paris	94.2	101.6	21.7
85	Paris	94.3	131.3	27.8
86	Lorraine	95.4	78.5	14.6
87	EUL 36	96.4	150.2	54.6
88	Paris	94.2	111.8	23.7
91	EUL 91	100.0	110.1	29.0
92	Corby	94.2	126.1	27.9
92 (replicate)	Corby	93.8	67.8	16.1
93	Paris	94.1	77.1	17.0
97	ATCC43290	92.8	95.7	24.7
98	ATCC43290	92.9	123.5	31.7
99	EUL 99	100.0	153.7	38.0
100	EUL 100	100.0	104.0	24.8
101	EUL 100	97.7	104.8	28.3
102	EUL 100	98.2	132.2	42.8
103	Lorraine	96.2	98.4	17.1
104	Paris	98.1	97.4	13.2
105	EUL 120	97.7	103.6	15.1
110	Paris	97.0	139.1	20.8
111	EUL 111	100.0	107.1	35.8

111 (replicate)	EUL 111	100.0	142.6	47.0
114	Paris	97.4	106.2	15.9
116	EUL 120	95.7	93.2	19.8
117	Paris	97.0	98.8	15.9
118	Philadelphia	98.7	105.7	14.4
119	Paris	97.8	99.6	13.8
120	EUL 120	98.2	112.4	94.0
9	Paris	98.4	140.6	16.6
10	Paris	98.4	133.4	16.4
11	EUL 28	97.0	142.4	38.9
12	EUL 28	97.0	144.8	23.5
22	Philadelphia	92.5	148.6	40.5
23	Philadelphia	92.5	156.4	43.8
24	Philadelphia	92.3	93.2	26.8
34	Philadelphia	93.4	132.4	34.4
35	Philadelphia	93.4	140.6	35.3
44	Paris	98.1	140.4	15.8
45	Paris	98.3	137.0	14.8
46	Paris	97.4	139.5	18.8
47	EUL 40	100.0	157.6	44.5
56	EUL 48	99.9	94.2	59.1
57	H044120014	94.6	140.7	36.1
58	Paris	98.3	128.3	16.5
59	EUL 51	100.0	143.9	44.6
76	H044120014	95.9	127.2	25.7
77	H044120014	95.9	126.2	25.3
78	EUL 165	95.7	92.3	20.5
79	EUL 165	95.8	127.8	26.4
94	Paris	94.2	96.9	20.8
95	Paris	94.2	95.9	20.5
96	Corby	94.0	97.2	21.0
106	ATCC43290	92.7	71.4	19.1
107	ATCC43290	92.8	90.4	23.5
121	EUL 120	98.2	149.1	101.8
Additional isolates				
LC 202/EUL 153	ATCC43290	95.2	120.9	26.0
LC 206/EUL 158	ATCC43290	95.3	131.7	27.9
LC 569/EUL 154	ATCC43290	92.7	165.3	60.9
LC 606/EUL 155	ATCC43290	92.7	168.8	62.7
LC 384/EUL 156	EUL 156	99.8	180.5	37.5
LC 395/EUL 159	EUL 156	99.8	86.2	19.1
LC6379-1/ EUL 145	EUL 145	99.9	157.3	39.3
LC6376	EUL 145	99.9	128.5	33.4
LC6382	EUL 145	99.9	135.7	34.4

LC6391	EUL 145	99.9	114.4	29.8
LC6394	EUL 145	99.9	130.8	33.4
LC6397	EUL 145	99.9	101.9	26.6
LC6406	EUL 145	99.9	98.7	25.7
LC6407	EUL 145	99.9	111.5	28.5
LC6408	EUL 145	99.9	118.5	29.3
LC6411	EUL 145	99.9	93.8	24.1
LC6412	EUL 145	99.9	115.0	29.8
LC6413	EUL 145	99.9	105.2	26.9
LC6416	EUL 145	99.9	100.1	25.9
LC6418	EUL 145	99.9	94.6	24.1
LC6385	EUL 145	99.9	147.7	38.0
LC6388	EUL 145	99.9	160.3	40.9
LC6409	EUL 145	99.9	159.0	40.3
LC6410	EUL 145	99.9	186.8	46.4
LC0537/EUL 132	EUL 165	97.5	89.2	14.2
LC0539/EUL 133	EUL 165	97.5	145.9	45.0
LC0540/EUL 134	EUL 165	97.5	87.5	15.4
LC0565	EUL 165	97.6	150.3	22.0
LC0583	EUL 165	97.6	155.5	24.0
H034680033	EUL 165	98.8	49.4	13.5
H034680035/ EUL 165	EUL 165	99.0	136.0	13.9
H034690056/ EUL 166	EUL 165	99.0	145.0	14.7
H034800427	EUL 165	98.9	99.8	35.2
H034980467	EUL 165	99.0	116.7	33.1
H034800423	Paris	97.5	82.7	28.4
OLDA1 (NCTC12008)	Paris	98.4	99.4	14.9
EUL 109	Paris	98.1	132.2	14.8
H064240448	EUL 165	97.9	148.1	24.2
LC0731	EUL 165	94.4	154.9	34.0
LC0732	EUL 165	94.4	153.3	34.7
LC0763	EUL 165	94.4	151.8	32.3
LC0782	EUL 165	94.4	156.9	33.4
LC0795	EUL 165	94.4	139.3	30.5
LC0801	EUL 165	94.3	97.6	21.9
LC5694	EUL 165	95.5	171.6	35.7
LC5722	EUL 165	95.4	159.7	34.6
LC5738	EUL 165	95.5	163.6	36.1
LC5755	EUL 165	97.6	161.5	23.3
LC6163	EUL 165	95.7	155.6	31.9
LC6267	EUL 165	97.6	180.4	25.4
LC6268	EUL 165	97.6	167.7	24.0
LC6228	EUL 165	98.9	141.5	15.4
H041380048	EUL 165	97.9	97.4	17.0

H041640791	EUL 165	97.9	136.7	20.3
H042960010	EUL 165	97.9	165.2	23.6
H061140013	EUL 165	97.6	142.5	23.3
H071880001	EUL 165	95.7	154.1	31.3
H073060003	EUL 165	97.6	137.9	20.8
H080820009	EUL 165	95.8	137.0	27.7
LC6058	EUL 165	97.6	152.0	23.7
LC6293	EUL 165	97.6	163.3	24.8
LC6788	EUL 165	97.9	159.5	25.4
H062660463	EUL 165	95.7	145.7	29.7
H073900557	EUL 165	97.6	130.3	21.1
LC1127	EUL 165	94.4	136.1	30.9
H084760449	EUL 165	98.9	141.1	23.9
H085020185	EUL 165	98.9	140.6	16.0
H090320386	EUL 165	97.9	138.0	22.9
H044260061	EUL 165	98.9	154.8	26.7
H093140322	EUL 165	98.9	138.2	27.2
H093160422	EUL 165	98.9	138.5	16.0
H092760433	EUL 165	97.6	123.0	19.2
H100940111	EUL 165	97.9	107.4	18.7
H101760092	EUL 165	95.7	142.4	29.4
H101820190	EUL 165	97.6	139.2	21.5
H102020414	EUL 165	98.9	149.1	21.5
H101980130	EUL 165	95.7	140.9	29.8
H103820081	EUL 165	97.6	141.5	22.2
H120240685	EUL 165	95.7	113.1	23.5
H104320293	EUL 165	97.6	140.8	24.8
H113180118	EUL 165	95.5	148.6	31.9
H113340664	EUL 165	95.5	141.2	32.6
H113280076	EUL 165	95.8	148.5	28.1
H113660550	EUL 165	95.7	145.5	29.9
H114740454	EUL 165	97.6	135.1	20.3
H115040456	EUL 165	97.6	140.2	20.7
H111580389	EUL 165	98.9	152.0	15.8
H113780240	EUL 165	97.6	135.8	22.6
H083920177	EUL 165	98.8	138.6	15.8
H084140691	EUL 165	97.9	145.8	22.1
H081180019	EUL 165	95.7	150.2	31.9
H103260667	EUL 165	95.7	101.8	21.8
LC464	EUL 165	95.7	137.4	28.2
LC0512	EUL 165	97.6	155.6	35.3
LC0794	EUL 165	97.6	142.8	21.8
LC0798	EUL 165	97.6	156.5	23.2
LC0536/EUL 131	EUL 165	97.6	96.2	15.2
LC230/EUL 122	EUL 120	98.2	112.9	15.6

LC231/EUL 123	EUL 120	98.2	78.7	12.2
LC0462/EUL 124	EUL 120	98.1	90.1	12.6
LC0463/EUL 125	EUL 120	98.0	114.0	14.3
H063920004/ EUL 169	Lorraine	98.0	138.5	17.7
H064160534/ EULV0410	Lorraine	98.1	102.1	15.2
H064160538/ EUL 170	Lorraine	98.0	118.1	15.3
H034700617	Lorraine	97.9	112.4	36.6
H043580159	Lorraine	98.0	96.1	13.8
H043580160	Lorraine	98.0	114.4	15.6
H043660021	Lorraine	98.1	95.5	13.6
H043680663	Lorraine	98.0	96.2	14.4
H043700021	Lorraine	98.1	100.5	14.7
H043790008	Lorraine	98.0	108.2	15.3
H052920051	Lorraine	98.1	104.6	14.3
H053540106	Lorraine	98.0	106.0	17.5
H063660005	Lorraine	97.9	169.9	22.3
H063660006	Lorraine	97.9	125.7	19.7
H063760006	Lorraine	97.9	196.0	24.4
H063660009	Lorraine	97.9	137.2	18.2
H063680006	Lorraine	97.9	149.7	20.1
H063680007	Lorraine	97.9	128.1	17.6
H063740003	Lorraine	97.9	125.6	23.9
H063740018	Lorraine	97.9	153.9	20.7
H063780007	Lorraine	97.9	133.7	20.8
H063780008	Lorraine	97.9	191.3	26.4
H063860003	Lorraine	97.9	123.8	20.5
H063960001	Lorraine	97.9	134.5	20.1
LC5759	Lorraine	98.0	99.1	13.9
H070420013	Lorraine	98.1	104.7	14.8
LC5822	Lorraine	98.1	106.2	15.3
H040260015	Lorraine	97.9	153.5	19.8
H055140095	Lorraine	98.0	86.3	12.7
H060780053	Lorraine	97.9	164.0	21.3
H061120064	Lorraine	97.9	139.1	19.8
H062840608	Lorraine	97.9	160.0	19.7
H062940111	Lorraine	97.9	143.6	50.8
H064320006	Lorraine	97.9	158.9	33.5
H064280005	Lorraine	97.9	136.6	26.4
H064380002	Lorraine	97.9	139.5	18.4
H064380001	Lorraine	97.9	138.9	24.0
H064560527	Lorraine	97.9	141.5	25.7
H064660638	Lorraine	98.1	94.7	15.5
H070160015	Lorraine	98.1	94.9	15.3
H071120010	Lorraine	97.9	134.1	19.6

H071360036	Lorraine	97.9	211.3	29.9
H072740002	Lorraine	97.9	123.9	28.7
H073000045	Lorraine	97.9	124.7	17.8
H073380007	Lorraine	98.1	102.4	14.4
H073600182	Lorraine	98.1	112.4	18.0
H073640185	Lorraine	97.9	129.7	19.1
H074960018	Lorraine	97.9	141.7	21.5
H080780059	Lorraine	97.9	152.6	19.9
H053840008	Lorraine	98.1	90.4	13.1
H072520002	Lorraine	97.9	135.0	18.6
H081340222	Lorraine	97.9	177.0	22.9
H082520613	Lorraine	97.9	141.4	23.3
H083120262	Lorraine	98.1	96.9	13.8
H083620580	Lorraine	97.9	122.3	20.8
H083960064	Lorraine	97.9	106.1	16.7
H084620118	Lorraine	98.1	92.3	13.1
H090140214	Lorraine	98.0	90.1	13.2
H090440226	Lorraine	98.0	99.7	15.6
H040960441	Lorraine	98.0	98.9	14.0
H041120007	Lorraine	98.1	96.5	14.6
H093480403	Lorraine	98.1	97.0	14.1
H094340202	Lorraine	98.1	91.1	12.5
H095060125	Lorraine	98.1	108.6	14.9
H100140151	Lorraine	98.0	94.2	13.6
H100660110	Lorraine	98.0	91.7	13.4
H100700025	Lorraine	98.0	95.3	16.8
H103140121	Lorraine	98.1	102.4	15.6
H103620160	Lorraine	98.1	91.6	13.2
H103660126	Lorraine	98.1	99.1	14.3
H103660121	Lorraine	98.0	105.4	14.6
H104420240	Lorraine	98.0	101.6	14.5
H110480273	Lorraine	98.0	102.2	14.6
H112320437	Lorraine	98.0	98.5	15.1
H112080616	Lorraine	98.1	104.6	14.6
H112380374	Lorraine	98.0	109.6	14.8
H120160499	Lorraine	98.1	108.4	15.3
H120200371	Lorraine	98.1	104.0	17.1
H105140391	Lorraine	98.1	116.2	17.0
H121040204	Lorraine	98.0	95.1	13.7
H121420445	Lorraine	98.0	99.7	14.3
H102240357	Lorraine	98.1	88.6	13.4
H122500497	Lorraine	98.0	103.4	14.8
H122820408	Lorraine	98.1	103.5	14.8
H123620597	Lorraine	98.1	95.9	13.7
H123840629	Lorraine	98.1	92.6	13.2

H123940534	Lorraine	98.1	105.6	15.9
H124920387	Lorraine	98.1	105.0	14.4
H131340777	Lorraine	98.0	99.7	15.0
H131480353	Lorraine	98.0	89.3	13.9
H131480354	Lorraine	98.1	94.6	14.6
H131840211	Lorraine	98.0	88.6	14.6
H131460248	Lorraine	98.1	97.3	14.1
H132140863	Lorraine	98.0	93.4	18.3
H053640534/ EUL 168	Lorraine	98.0	150.8	18.6
H064180002	H044120014	92.4	98.1	27.9
H064180019	H044120014	92.4	109.0	29.7
H043540106	H044120014	95.8	108.2	21.5
H044120014	H044120014	98.8	110.1	15.4
H052780022	H044120014	95.9	92.3	19.9
H054280040	H044120014	95.3	88.2	20.2
H063680003	H044120014	98.2	91.5	13.1
H063840008	H044120014	95.3	107.5	23.5
H073660582	H044120014	98.2	106.7	16.7
LC5804	H044120014	95.5	82.8	17.4
H063760005	H044120014	94.8	103.8	27.1
H064240003	H044120014	94.6	100.6	28.6
H065040012	H044120014	94.6	84.7	22.2
H070140635	H044120014	95.3	90.0	19.8
H073020039	H044120014	95.3	92.9	22.2
H073320399	H044120014	95.3	97.7	27.5
H073440003	H044120014	95.3	114.0	26.9
LC6009	H044120014	94.5	80.1	21.8
H083140015	H044120014	94.6	97.4	24.4
H093400182	H044120014	94.9	100.3	34.8
H094760070	H044120014	95.4	96.4	21.9
H094800237	H044120014	95.4	95.6	22.3
H110480715	H044120014	95.8	86.5	29.9
H112840293	H044120014	96.1	92.5	20.5
H114100406	H044120014	94.6	102.7	27.9
H120240362	H044120014	95.4	86.1	18.0
H104640262	H044120014	95.5	89.4	18.7
H123140428	H044120014	95.3	95.9	22.2
H123460520	H044120014	95.5	96.3	22.4
H124360642	H044120014	95.4	88.5	18.7
Pontiac-1	H044120014	95.4	101.4	33.7

Appendix Table 13. 370 *L. pneumophila* isolates used to define the total core gene content of the species.

Isolate name	Reference/ Accession number		Isolate name	Reference/ Accession number		Isolate name	Reference/ Accession number
EUL 1	ERR376626		EUL 145	ERR376769		H122820408	ERR363980
EUL 2	ERR376627		EUL 148	ERR376772		H122500497	ERR363981
EUL 3	ERR376628		EUL 149	ERR376773		H121040204	ERR363982
EUL 4	ERR376721		EUL 150	ERR376774		H121420445	ERR363983
EUL 5	ERR376630		EUL 153	ERR376775		H120200371	ERR363984
EUL6	ERR376631		EUL 154	ERR376776		H120160499	ERR363985
EUL 7	ERR376632		EUL 155	ERR376777		H131840211	ERR363986
EUL 8	ERR376633		EUL 156	ERR376778		H131460248	ERR363987
EUL 9	ERR376634		EUL 157	ERR376779		H131480354	ERR363988
EUL 10	ERR376635		EUL 158	ERR376780		H131480353	ERR363989
EUL 11	ERR376636		EUL 159	ERR352158		H124920387	ERR363991
EUL 12	ERR376637		EUL 161	ERR376781		H120240685	ERR363992
EUL 13	ERR376646		EUL 162	ERR376782		H105140391	ERR363993
EUL 14	ERR376639		EUL 163	ERR376783		H064160534	ERR363994
EUL 16	ERR376641		EUL 164	ERR376784		H043540106	ERR363997
EUL 17	ERR376642		EUL 165	ERR376785		H052780022	ERR363998
EUL 18	ERR376643		EUL 166	ERR376786		H044120014	ERR363999
EUL 19	ERR376644		EUL 167	ERR352160		H063760005	ERR364000
EUL 20	ERR376645		EUL 168	ERR352161		H063840008	ERR364001
EUL 21	ERR376638		EUL 169	ERR376787		H063680003	ERR364002
EUL 22	ERR376647		EUL 170	ERR376788		H094760070	ERR364003
EUL 23	ERR376648		H123640643	ERR332166		H064180019	ERR364004
EUL 24	ERR332110		H041380048	ERR363843		H064240003	ERR364005
EUL 25	ERR376650		H041640791	ERR363844		H093400182	ERR364006
EUL 26	ERR376651		H042960010	ERR363845		H083140015	ERR364007
EUL 27	ERR376652		H044260061	ERR363846		H073660582	ERR364008
EUL 28	ERR376722		H061140013	ERR363847		H073320399	ERR364010
EUL 30	ERR376655		H062660463	ERR363848		H070140635	ERR364011
EUL 31	ERR376656		H064240448	ERR363849		H124360642	ERR364013
EUL 32	ERR376657		H071880001	ERR363850		H123460520	ERR364014
EUL 33	ERR376658		H073060003	ERR363851		H123140428	ERR364015
EUL 34	ERR376659		H073900557	ERR363852		H114100406	ERR364016
EUL 35	ERR376660		H080820009	ERR363853		H112840293	ERR364017
EUL 36	ERR332122		H081180019	ERR363854		H110480715	ERR364018
EUL 37	ERR376723		H083920177	ERR363855		H104640262	ERR364019
EUL 38	ERR376663		H084140691	ERR363856		H094800237	ERR364020
EUL 40	ERR376665		H084760449	ERR363857		H064180002	ERR364021
EUL 41	ERR376666		H085020185	ERR363858		H073020039	ERR364022
EUL 42	ERR376667		H090320386	ERR363859		H073340594	ERR364023

EUL 43	ERR376668		H092760433	ERR363860		H073240536	ERR364024
EUL 44	ERR376669		H093140322	ERR363861		H120240362	ERR364025
EUL 45	ERR376670		H093160422	ERR363862		H073280012	ERR364026
EUL 46	ERR376671		H100940111	ERR363863		H073340034	ERR364027
EUL 47	ERR376672		H101760092	ERR363864		H054280040	ERR364028
EUL 48	ERR332134		H101820190	ERR363865		H132140863	ERR364031
EUL 50	ERR376675		H101980130	ERR363866		H092380261	ERR434063
EUL 51	ERR376676		H102020414	ERR363867		H092400768	ERR434064
EUL 52	ERR376677		H103820081	ERR363868		LC6385	ERR352162
H041120007	ERR363942		H104320293	ERR363869		LC6388	ERR352163
EUL 53	ERR376725		H111580389	ERR363870		LC6409	ERR352164
EUL 54	ERR376679		H113180118	ERR363871		LC6410	ERR352165
EUL 55	ERR332141		H113280076	ERR363872		LC464	ERR363878
EUL 56	ERR376726		H113340664	ERR363873		LC0512	ERR363879
EUL 57	ERR376682		H113660550	ERR363874		LC0565	ERR363880
EUL 58	ERR376683		H113780240	ERR363875		LC0583	ERR363881
EUL 60	ERR376685		H114740454	ERR363876		LC0731	ERR363882
EUL 61	ERR376686		H115040456	ERR363877		LC0732	ERR363883
EUL 62	ERR376687		H040260015	ERR363903		LC0763	ERR363884
EUL 63	ERR332149		H063660005	ERR363904		LC0782	ERR363885
EUL 64	ERR376727		H063740018	ERR363906		LC0794	ERR363886
EUL 66	ERR376728		H060780053	ERR363907		LC0795	ERR363887
EUL 67	ERR376692		H071360036	ERR363908		LC0801	ERR363889
EUL 68	ERR376693		H081340222	ERR363909		LC1127	ERR363890
EUL 69	ERR376694		H080780059	ERR363910		LC5694	ERR363891
EUL 70	ERR376695		H082520613	ERR363912		LC5722	ERR363892
EUL 71	ERR332157		H063680007	ERR363913		LC5738	ERR363893
EUL 72	ERR332158		H061120064	ERR363914		LC5755	ERR363894
EUL 73	ERR376698		H063760006	ERR363915		LC6058	ERR363896
EUL 74	ERR376729		H063780008	ERR363916		LC6163	ERR363897
EUL 75	ERR376700		H062840608	ERR363917		LC6228	ERR363898
EUL 76	ERR376701		H063680006	ERR363918		LC6267	ERR363899
EUL 77	ERR376702		H062940111	ERR363919		LC6268	ERR363900
EUL 78	ERR376730		H074960018	ERR363920		LC6293	ERR363901
EUL 81	ERR376732		H064380001	ERR363921		LC6788	ERR363902
EUL 82	ERR376733		H063660006	ERR363922		LC5759	ERR363995
EUL 83	ERR376734		H064320006	ERR363923		LC5822	ERR363996
EUL 84	ERR376735		H064280005	ERR363924		LC5804	ERR364029
EUL 85	ERR376710		H064560527	ERR363925		LC6009	ERR364030
EUL 86	ERR332172		H064380002	ERR363926		LC6376	ERR376790
EUL 87	ERR376712		H072520002	ERR363927		LC6382	ERR376792
EUL 88	ERR332174		H063960001	ERR363928		LC6391	ERR376793
EUL 90	ERR376736		H063740003	ERR363929		LC6394	ERR376794
EUL 91	ERR376737		H063860003	ERR363930		LC6397	ERR376795
EUL 92	ERR376717		H071120010	ERR363931		LC6406	ERR376796

EUL 93	ERR332179		H073000045	ERR363932		LC6407	ERR376797
EUL 94	ERR376738		H073640185	ERR363933		LC6408	ERR376798
EUL 95	ERR376739		H063780007	ERR363934		LC6411	ERR376799
EUL 96	ERR376740		H083620580	ERR363936		LC6412	ERR376800
EUL 97	ERR376741		H083960064	ERR363937		LC6413	ERR376801
EUL 98	ERR376629		H103260667	ERR363938		LC6416	ERR376802
EUL 100	ERR376742		H084620118	ERR363939		LC6418	ERR376804
EUL 101	ERR376743		H073380007	ERR363940		OLDA1	ERR434061
EUL 102	ERR376714		H083120262	ERR363941		Alcoy	D'Auria *et al*. 2010
EUL 103	ERR376744		H043700021	ERR363944		ATCC43290	Amaro *et al*. 2012
EUL 104	ERR376745		H043790008	ERR363945		Corby	Glockner *et al*. 2008
EUL 105	ERR376746		H043660021	ERR363946		HL06041035	Gomez-Valero *et al.* 2011
EUL 107	ERR376747		H055140095	ERR363947		Lorraine	Gomez-Valero *et al*. 2011
EUL 108	ERR376748		H053540106	ERR363948		Thunderbay	Khan *et al.* 2013
EUL 109	ERR376662		H043680663	ERR363949		Lens	Cazalet *et al.* 2004
EUL 110	ERR376674		H103620160	ERR363950		Paris	Cazalet *et al.* 2004
EUL 111	ERR376749		H112320437	ERR363951		Philadelphia	Chien *et al*. 2004
EUL 113	ERR363968		H112080616	ERR363952		H043940028	Underwood *et al.* 2013
EUL 114	ERR363969		H040960441	ERR363953		H044500045	Underwood *et al.* 2013
EUL 115	ERR376753		H053840008	ERR363954		H044540088	Underwood *et al.* 2013
EUL 116	ERR376754		H102240357	ERR363955		H063280001	Underwood *et al.* 2013
EUL 117	ERR376755		H104420240	ERR363957		H065000139	Underwood *et al.* 2013
EUL 118	ERR340981		H100700025	ERR363958		H070840415	Underwood *et al.* 2013
EUL 119	ERR376757		H043580160	ERR363959		H071260094	Underwood *et al.* 2013
EUL 120	ERR376758		H112380374	ERR363960		H074360702	Underwood *et al.* 2013
EUL 121	ERR376678		H052920051	ERR363961		H074360710	Underwood *et al.* 2013
EUL 122	ERR376759		H100660110	ERR363962		H075160080	Underwood *et al.* 2013
EUL 123	ERR332142		H090140214	ERR363963		H090500162	Underwood *et al.* 2013
EUL 124	ERR332150		H064660638	ERR363964		H091960009	Underwood *et al.* 2013
EUL 125	ERR376760		H100140151	ERR363965		H091960011	Underwood *et al.* 2013
EUL 129	ERR376762		H090440226	ERR363966		H093380153	Underwood *et al.* 2013

EUL 130	ERR376703		H103140121	ERR363967		H093620212	Underwood *et al.* 2013
EUL 131	ERR332167		H070160015	ERR363970		H100260089	Underwood *et al.* 2013
EUL 132	ERR332168		H094340202	ERR363971		LC3330	Underwood *et al.* 2013
EUL 133	ERR332169		H093480403	ERR363973		LC6451	Underwood *et al.* 2013
EUL 134	ERR332170		H103660126	ERR363974		LC6774	Underwood *et al.* 2013
EUL 140	ERR376653		H123940534	ERR363975		RR08000517	Underwood *et al.* 2013
EUL 141	ERR376765		H073600182	ERR363976		RR08000760	Underwood *et al.* 2013
EUL 142	ERR376766		H123840629	ERR363978			
EUL 143	ERR376767		H123620597	ERR363979			

Appendix Table 14. Genes used in the cgMLST schemes with 50, 100, 500 or 1455 core genes.

Gene	Product	Length (bp)	Scheme (no. of genes)
lpg0085	hypothetical protein	528	50, 100, 500, 1455
lpg0104	peptide methionine sulfoxide reductase	576	50, 100, 500, 1455
lpg0131	dihydropicolinate reductase	732	50, 100, 500, 1455
lpg0136	pyruvate kinase II	1425	50, 100, 500, 1455
lpg0189	hypothetical protein	867	50, 100, 500, 1455
lpg0245	NAD-glutamate dehydrogenase	3381	50, 100, 500, 1455
lpg0329	50S ribosomal protein L3	651	50, 100, 500, 1455
lpg0331	50S ribosomal protein L23	279	50, 100, 500, 1455
lpg0409	hypothetical, SURF1 family	729	50, 100, 500, 1455
lpg0419	glucokinase	1008	50, 100, 500, 1455
lpg0525	hypothetical virulence protein	627	50, 100, 500, 1455
lpg0596	hypothetical protein	696	50, 100, 500, 1455
lpg0601	ABC transporter, permease	1449	50, 100, 500, 1455
lpg0607	lysyl tRNA synthetase	954	50, 100, 500, 1455
lpg0622	transmembrane protein	1944	50, 100, 500, 1455
lpg0664	D-ribulose-5-phosphate-3-epimerase	654	50, 100, 500, 1455
lpg0689	DNA binding stress protein	441	50, 100, 500, 1455
lpg0700	protein-L-isoaspartate-O-methyltransferase	675	50, 100, 500, 1455
lpg0812	rod shape determining protein MreC	909	50, 100, 500, 1455
lpg0866	3-methyladenine DNA glycosylase	552	50, 100, 500, 1455
lpg0871	hypothetical protein	867	50, 100, 500, 1455
lpg0890	cystathionine beta-lyase	1152	50, 100, 500, 1455
lpg0957	hypothetical protein	906	50, 100, 500, 1455
lpg1323	drug resistance transporter, Bcr/CflA	1161	50, 100, 500, 1455

lpg1503	pyruvate dehydrogenase E2 component	1653	50, 100, 500, 1455
lpg1534	glutamate-1-semialdehyde-2,1-aminomutase	1302	50, 100, 500, 1455
lpg1543	transmembrane protein	675	50, 100, 500, 1455
lpg1586	hypothetical protein	378	50, 100, 500, 1455
lpg1737	glutamyl/tRNA (Gln) amidotransferase, B subunit	1434	50, 100, 500, 1455
lpg1744	HesB family protein	369	50, 100, 500, 1455
lpg1759	flagellar motor switch protein FliG	990	50, 100, 500, 1455
lpg1811	aspartokinase	2580	50, 100, 500, 1455
lpg1869	ribonuclease III	675	50, 100, 500, 1455
lpg1909	hypothetical protein	1005	50, 100, 500, 1455
lpg2229	saframycin Mx1 synthetase B	1746	50, 100, 500, 1455
lpg2264	hypothetical protein	315	50, 100, 500, 1455
lpg2331	biotin synthase BioC	1005	50, 100, 500, 1455
lpg2349	alkylhydroperoxidase AhpD family core domain protein	564	50, 100, 500, 1455
lpg2387	plasminogen activator	927	50, 100, 500, 1455
lpg2494	hypothetical protein	693	50, 100, 500, 1455
lpg2528	alpha-amylase, putative	1551	50, 100, 500, 1455
lpg2597	DNA processing enzyme DprA (SMF family)	1086	50, 100, 500, 1455
lpg2633	hypothetical protein	318	50, 100, 500, 1455
lpg2654	GTP binding protein	1092	50, 100, 500, 1455
lpg2691	cation transporting ATPase PacS	2544	50, 100, 500, 1455
lpg2699	ATPase or kinase	483	50, 100, 500, 1455
lpg2864	hypothetical protein	1119	50, 100, 500, 1455
lpg2878	cobalt/magnesium uptake transporter	1065	50, 100, 500, 1455
lpg2882	methionyl tRNA synthetase	2082	50, 100, 500, 1455
lpg2902	hypothetical protein	426	50, 100, 500, 1455
lpg0011	thiol-disulfide oxidoreductase ResA	471	100, 500, 1455
lpg0014	transmembrane protein	1212	100, 500, 1455
lpg0033	hypothetical protein	1026	100, 500, 1455
lpg0079	2-polyprenyl-6-methoxyphenol hydroxylase	1164	100, 500, 1455
lpg0127	acetyl-coenzyme A synthetase	1884	100, 500, 1455
lpg0287	translation elongation factor P (EF-P)	618	100, 500, 1455
lpg0415	hypothetical protein	246	100, 500, 1455
lpg0531	succinate dehydrogenase iron-sulfur protein subunit B	723	100, 500, 1455
lpg0540	major facilitator family transporter	1284	100, 500, 1455
lpg0551	1-acyl-sn-glycerol-3-phosphate acetyltransferase	774	100, 500, 1455
lpg0581	hypothetical protein	186	100, 500, 1455
lpg0606	metal-sulfur cluster biosynthetic enzyme	372	100, 500, 1455
lpg0650	50S ribosomal protein L31	228	100, 500, 1455
lpg0785	acetyl CoA carboxylase, carboxyltransferase, alpha subunit	954	100, 500, 1455
lpg0880	hypothetical protein	642	100, 500, 1455

lpg0963	hypothetical protein	1242	100, 500, 1455
lpg1202	cytochrome D ubiquinol oxidase, subunit I	1533	100, 500, 1455
lpg1225	flagellar hook associated protein 1 FlgK	1950	100, 500, 1455
lpg1298	hypothetical protein	201	100, 500, 1455
lpg1302	tRNA pseudouridine synthase A	789	100, 500, 1455
lpg1366	hypothetical protein	774	100, 500, 1455
lpg1386	enhanced entry protein EnhA	474	100, 500, 1455
lpg1396	acyl carrier protein	249	100, 500, 1455
lpg1457	GTP pyrophosphokinase ((p)ppGpp synthetase I) stringent stress response RelA	2205	100, 500, 1455
lpg1565	thiamine biosynthesis protein NMT-1	951	100, 500, 1455
lpg1576	Holliday junction DNA helicase RuvB	1032	100, 500, 1455
lpg1690	aconitate hydratase	2676	100, 500, 1455
lpg1772	hypothetical protein	777	100, 500, 1455
lpg1844	D-tyrosyl-tRNA	438	100, 500, 1455
lpg1916	possible regulator of murein genes BolA	318	100, 500, 1455
lpg2008	endoribonuclease L-PSP	387	100, 500, 1455
lpg2053	hypothetical protein	879	100, 500, 1455
lpg2191	global stress protein GspA	528	100, 500, 1455
lpg2209	hypothetical protein	531	100, 500, 1455
lpg2299	ATP-dependent RNA helicase	2877	100, 500, 1455
lpg2317	transmembrane protein	1161	100, 500, 1455
lpg2333	membrane associated zinc metalloprotease	1092	100, 500, 1455
lpg2337	protein methyltransferase HemK	864	100, 500, 1455
lpg2345	ATP-dependent RNA helicase	1770	100, 500, 1455
lpg2481	integral membrane protein	903	100, 500, 1455
lpg2594	methionyl tRNA formyltransferase	945	100, 500, 1455
lpg2620	chromosome segregation SMC protein	3495	100, 500, 1455
lpg2623	transmembrane protein	813	100, 500, 1455
lpg2627	hypothetical protein	1182	100, 500, 1455
lpg2657	ferrous iron transporter B	2256	100, 500, 1455
lpg2674	DotD	492	100, 500, 1455
lpg2764	inorganic pyrophosphatase	537	100, 500, 1455
lpg2843	inosine 5'-monophosphate dehydrogenase	1014	100, 500, 1455
lpg2930	sec-independent (periplasmic) protein translocase protein TatC	726	100, 500, 1455
lpg3005	50S ribosomal protein L34	135	100, 500, 1455
lpg0001	chromosomal replication initiator protein DnaA	1359	500, 1455
lpg0009	host factor-I protein for bacteriophage Q beta replication	258	500, 1455
lpg0010	GTP binding protein HflX	1260	500, 1455
lpg0018	outer membrane efflux protein	1386	500, 1455
lpg0024	hemin binding protein Hbp	426	500, 1455
lpg0027	low affinity inorganic phosphate transporter	996	500, 1455
lpg0059	hypothetical protein	1107	500, 1455

lpg0078	2-octaprenyl-6-methoxyphenol hydroxylase	1203	500, 1455
lpg0083	glutathione synthase/ribosomal protein S6 modification enzyme	948	500, 1455
lpg0084	hypothetical protein	1524	500, 1455
lpg0098	two component sensor and regulator, histidine kinase response regulator	453	500, 1455
lpg0099	DNA polymerase I	2691	500, 1455
lpg0101	hypothetical protein	783	500, 1455
lpg0103	N-terminal acetyltransferase, GNAT family	861	500, 1455
lpg0116	glycine cleavage system protein P	1371	500, 1455
lpg0118	glycine cleavage system T protein	1104	500, 1455
lpg0120	IcmL-like	531	500, 1455
lpg0128	3-hydroxyisobutyrate dehydrogenase	918	500, 1455
lpg0137	phosphoglycerate kinase	1191	500, 1455
lpg0165	hypothetical protein	456	500, 1455
lpg0175	pyoverdine biosynthesis protein PvcB	837	500, 1455
lpg0212	deoxyribodipyrimidine photolyase	1416	500, 1455
lpg0218	phosphoribosylaminoimidazole carboxylase, catalytic subunit PurE	501	500, 1455
lpg0232	transcriptional regulator np20, Fur family	534	500, 1455
lpg0241	glutaminase	933	500, 1455
lpg0248	arsenate reductase	342	500, 1455
lpg0252	membrane protein	696	500, 1455
lpg0257	multidrug resistance secretion protein	993	500, 1455
lpg0260	hypothetical protein	399	500, 1455
lpg0268	hypothetical protein	597	500, 1455
lpg0288	L-lysine 2,3-aminomutase, radical SAM domain protein	981	500, 1455
lpg0289	polyphosphate kinase	2085	500, 1455
lpg0290	lipoprotein	1206	500, 1455
lpg0291	chromate transport protein	534	500, 1455
lpg0293	long chain acyl-CoA dehydrogenase	2439	500, 1455
lpg0294	hypothetical protein	693	500, 1455
lpg0296	hypothetical phosphotransferase	1047	500, 1455
lpg0317	transcription antitermination protein NusG	555	500, 1455
lpg0318	50S ribosomal protein L11	435	500, 1455
lpg0342	30S ribosomal protein S14	288	500, 1455
lpg0346	30S ribosomal protein S5	507	500, 1455
lpg0352	30S ribosomal protein S11	399	500, 1455
lpg0354	DNA-directed RNA polymerase alpha subunit RpoA	993	500, 1455
lpg0362	3-oxoacyl-(acyl carrier protein) synthase II, N-terminal	1278	500, 1455
lpg0376	SdhA, GRIP coiled-coil protein GCC185	4290	500, 1455
lpg0383	hypothetical protein	483	500, 1455
lpg0385	LemA protein	582	500, 1455
lpg0386	heat shock protein HtpX	1020	500, 1455

lpg0388	ABC transporter, ATP binding component	915	500, 1455
lpg0408	inner (transmembrane) protein	549	500, 1455
lpg0410	hypothetical protein	546	500, 1455
lpg0411	cytochrome c oxidase assembly protein	1032	500, 1455
lpg0414	glutathione synthase, ribosomal protein S6 modification protein	909	500, 1455
lpg0439	hypothetical protein	1050	500, 1455
lpg0453	IcmC (DotE)	585	500, 1455
lpg0456	IcmB (DotO)	3030	500, 1455
lpg0461	ribosomal protein L11 methyltransferase	870	500, 1455
lpg0463	acetyl CoA carboxylase, biotin carboxyl carrier protein	483	500, 1455
lpg0464	3-dehydroquinate dehydratase type II	438	500, 1455
lpg0474	CDP-diacylglycerol-serine-O-phosphatidyltransferase	744	500, 1455
lpg0477	RNA polymerase signma-54 factor RpoN	1395	500, 1455
lpg0479	50S ribosomal protein L28	237	500, 1455
lpg0481	S-adenosylmethionine-dependent methyltransferase	681	500, 1455
lpg0483	ankyrin repeat-containing protein	1488	500, 1455
lpg0485	HflC protein	921	500, 1455
lpg0493	amino acid (glutamine) ABC transporter, ATP binding component	669	500, 1455
lpg0497	adenosine deaminase	1476	500, 1455
lpg0529	succinate dehydrogenase hydrophobic membrane anchor protein subunit D	348	500, 1455
lpg0532	2-oxoglutarate dehydrogenase E1 component)	2835	500, 1455
lpg0533	dihydrolipoamide succinyltransferase	1230	500, 1455
lpg0536	pyridoxamine 5'-phosphate oxidase	648	500, 1455
lpg0541	probable membrane protein YdgA-like	1485	500, 1455
lpg0557	formamidopyrimidine DNA glycosylase	825	500, 1455
lpg0559	hypothetical protein	399	500, 1455
lpg0564	hypothetical protein	1071	500, 1455
lpg0565	spore maturation protein A	618	500, 1455
lpg0580	adenosine deaminase	981	500, 1455
lpg0586	transcriptional regulator	564	500, 1455
lpg0593	5-formyltetrahydrofolate cyclo-ligase	582	500, 1455
lpg0595	4-amino-4-deoxychorismate lyase	816	500, 1455
lpg0598	hypothetical protein	420	500, 1455
lpg0599	poly-beta-hydroxybutyrate polymerase	1761	500, 1455
lpg0600	rrf2 family protein	462	500, 1455
lpg0603	ABC transporter, permease component	1287	500, 1455
lpg0611	metal ion transporter	1314	500, 1455
lpg0618	3-methyladenine DNA glycosylase	573	500, 1455
lpg0623	hypothetical protein	393	500, 1455
lpg0629	Tfp pilus assembly protein PilX	513	500, 1455
lpg0631	type IV fimbrial biogenesis protein PilV	540	500, 1455

lpg0633	polysaccharide deacetylase	906	500, 1455
lpg0634	hypothetical protein	1350	500, 1455
lpg0662	multidrug efflux MFS outer membrane protein (RND family)	1440	500, 1455
lpg0667	hypothetical protein	870	500, 1455
lpg0670	hypothetical protein	1083	500, 1455
lpg0686	thiol:disulfide interchange protein DsbD	1791	500, 1455
lpg0697	sulfate transporter	2307	500, 1455
lpg0701	2-amino-3-ketobutyrate coenzyme A ligase	1248	500, 1455
lpg0726	ATP cone and Zn ribbon domains protein	468	500, 1455
lpg0729	phosphatidylglycerophosphatase A (PgpA)	483	500, 1455
lpg0745	lipoic acid synthetase	990	500, 1455
lpg0748	LPS biosynthesis protein, PseA-like	1383	500, 1455
lpg0760	glucose-1-phosphate thymidylyltransferase RmlA	918	500, 1455
lpg0800	L-aspartate oxidase	1650	500, 1455
lpg0801	adenylsuccinate lyase	1371	500, 1455
lpg0803	acyl CoA dehydrogenase, short chain specific	1704	500, 1455
lpg0805	phosphoenolpyruvate synthase	2388	500, 1455
lpg0808	UDP-N-acetylglucosamine-N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase	1092	500, 1455
lpg0817	hypothetical protein	336	500, 1455
lpg0821	lipopolysaccharide biosynthesis glycosyltransferase	780	500, 1455
lpg0822	O-antigen biosynthesis protein	1251	500, 1455
lpg0823	neurogenic locus notch protein homolog precursor	375	500, 1455
lpg0824	rhomboid family protein	600	500, 1455
lpg0825	peptidase, M23/M37 family	915	500, 1455
lpg0834	anthranilate phosphoribosyltransferase	1035	500, 1455
lpg0839	hydrolase, HAD superfamily, subfamily III A	552	500, 1455
lpg0840	polysialic acid capsule expression protein	963	500, 1455
lpg0841	toluene tolerance ABC transporter, ATP binding protein Ttg2A	798	500, 1455
lpg0849	ABC transporter, ATP binding protein	678	500, 1455
lpg0853	transcriptional regulator FleQ	1416	500, 1455
lpg0854	hypothetical protein	282	500, 1455
lpg0865	cytochrome c type biogenesis protein CycH	405	500, 1455
lpg0870	3-hydroxyisobutyryl Coenzyme A hydrolase	1059	500, 1455
lpg0873	hypothetical protein	426	500, 1455
lpg0879	two component response regulator with GGDEF domain	1143	500, 1455
lpg0885	glycosyl hydrolase	1077	500, 1455
lpg0886	sodium:dicarboxylate symporter	1281	500, 1455
lpg0891	sensory box protein/GGDEF/EAL domains	2316	500, 1455
lpg0895	hypothetical protein	510	500, 1455

lpg0901	hypothetical protein NMA0899	657	500, 1455
lpg0911	Bvg accessory factor	771	500, 1455
lpg0919	transmembrane protein	660	500, 1455
lpg0921	hypothetical protein	1245	500, 1455
lpg0922	electron transfer flavoprotein, beta subunit	750	500, 1455
lpg0923	electron transfer flavoprotein, alpha subunit	939	500, 1455
lpg0924	alanine dehydrogenase	1122	500, 1455
lpg0930	type IV pilus biogenesis protein PilP	588	500, 1455
lpg0936	riboflavin biosynthesis RibF	987	500, 1455
lpg0950	nitrilase	807	500, 1455
lpg0954	transcription repair coupling factor	3462	500, 1455
lpg0958	DNA ligase	2052	500, 1455
lpg0971	ecto-ATP diphosphohydrolase II	1146	500, 1455
lpg1121	hypothetical protein	771	500, 1455
lpg1139	spermidine/putrescine ABC transporter permease protein PotC	768	500, 1455
lpg1140	spermidine/putrescine ABC transporter permease protein PotB	825	500, 1455
lpg1141	spermidine/putrescine ABC transporter, ATP-binding protein PotA	1149	500, 1455
lpg1143	short chain type dehydrogenase/reductase	885	500, 1455
lpg1154	hypothetical protein	1083	500, 1455
lpg1157	lipase B	765	500, 1455
lpg1159	permeases of drug/transporter	1038	500, 1455
lpg1162	OmpA-like transmembrane domain protein	732	500, 1455
lpg1179	riboflavin biosynthesis protein RibA	1209	500, 1455
lpg1180	riboflavin synthase, beta subunit	468	500, 1455
lpg1198	histidinol-phosphate aminotransferase	1095	500, 1455
lpg1218	flagellar basal body rod modification protein FlgD	678	500, 1455
lpg1219	flagellar hook protein FlgE	1314	500, 1455
lpg1226	flagellar hook associated protein type 3 FlgL	1236	500, 1455
lpg1276	electron transferring flavoprotein dehydrogenase	1632	500, 1455
lpg1280	malate oxidoreductase	1725	500, 1455
lpg1283	lipoprotein NlpD	744	500, 1455
lpg1285	homogentisate 1,2-dioxygenase	1251	500, 1455
lpg1287	crossover junction endodeoxyribonuclease RuvC	525	500, 1455
lpg1296	protein involved in catabolism of external DNA	864	500, 1455
lpg1304	tryptophan synthetase, beta subunit	1200	500, 1455
lpg1319	type II secretory pathway protein E	1485	500, 1455
lpg1338	flagellar hook associated protein 2 FliD	1626	500, 1455
lpg1340	flagellin	1428	500, 1455
lpg1347	rare lipoprotein B	492	500, 1455
lpg1348	leucyl tRNA synthetase	2472	500, 1455

lpg1351	piperidine-6-carboxylate dehydrogenase	1521	500, 1455
lpg1358	general secretion pathway protein LspK	969	500, 1455
lpg1360	general secretion pathway protein LspI	378	500, 1455
lpg1364	glutamine synthetase, type I	1410	500, 1455
lpg1365	hypothetical protein	561	500, 1455
lpg1367	1-aminocyclopropane-1-carboxylate deaminase	903	500, 1455
lpg1374	rod shape determining protein RodA	1119	500, 1455
lpg1394	S-malonyl transferase	948	500, 1455
lpg1395	3-oxoacyl-(acyl carrier protein) reductase	747	500, 1455
lpg1399	thymidylate kinase	639	500, 1455
lpg1400	DNA polymerase III, delta prime subunit	906	500, 1455
lpg1402	deoxyribonuclease TatD	813	500, 1455
lpg1414	glycerol kinase	1476	500, 1455
lpg1420	cytidylate kinase	696	500, 1455
lpg1422	hypothetical membrane protein	291	500, 1455
lpg1424	aminotransferase	1116	500, 1455
lpg1425	orotidine 5'-phosphate decarboxylase PyrF	690	500, 1455
lpg1431	hypothetical protein	399	500, 1455
lpg1434	xanthosine phosphorylase	840	500, 1455
lpg1452	lipase A	888	500, 1455
lpg1464	hypothetical protein	159	500, 1455
lpg1469	Rtn protein	1599	500, 1455
lpg1485	hypothetical protein	402	500, 1455
lpg1486	AsnC family transcription regulator protein	522	500, 1455
lpg1507	sodium/hydrogen antiporter	1173	500, 1455
lpg1508	rare lipoprotein A	822	500, 1455
lpg1509	D-alanyl-D-alanine carboxypeptidase	1293	500, 1455
lpg1512	DedA/PAP2 domain protein	2037	500, 1455
lpg1514	lipoprotein	732	500, 1455
lpg1526	hypothetical protein	594	500, 1455
lpg1531	phenazine biosynthesis PhzF	792	500, 1455
lpg1537	transport protein	732	500, 1455
lpg1540	universal stress protein A	423	500, 1455
lpg1548	nucleoside diphosphate kinase	477	500, 1455
lpg1558	pyruvate dehydrogenase E1 alpha subunit	1095	500, 1455
lpg1562	mercuric reductase	2145	500, 1455
lpg1597	thiolase	1320	500, 1455
lpg1604	hypothetical protein	678	500, 1455
lpg1620	hypothetical protein	441	500, 1455
lpg1639	hypothetical protein	1317	500, 1455
lpg1641	acylaminoacyl peptidase	1980	500, 1455
lpg1644	hypothetical protein	765	500, 1455
lpg1646	cytochrome b561 transmembrane protein	531	500, 1455
lpg1659	membrane protein	1044	500, 1455
lpg1666	hypothetical protein	1404	500, 1455

lpg1669	alpha-amylase, putative	2226	500, 1455
lpg1672	phosphoribosylglycinamide formyltransferase	579	500, 1455
lpg1674	amidophosphoribosyltransferase	1500	500, 1455
lpg1680	thiol:disulfide interchange protein DsbD	1380	500, 1455
lpg1700	uracil DNA glycosylase	720	500, 1455
lpg1721	deaminase	426	500, 1455
lpg1722	GMP synthetase	1578	500, 1455
lpg1730	sn-glycerol-3-phosphate transmembrane ABC transporter	771	500, 1455
lpg1735	glutamyl/tRNA (Gln) amidotransferase, C subunit	303	500, 1455
lpg1736	glutamyl/tRNA (Gln) amidotransferase, A subunit	1452	500, 1455
lpg1746	cysteine desulfurase NifS	1164	500, 1455
lpg1748	inositol-1-monophosphatase	786	500, 1455
lpg1756	flagellar protein FliJ	456	500, 1455
lpg1761	flagellar hook-basal body protein FliE	315	500, 1455
lpg1763	sensor kinase HydH	1032	500, 1455
lpg1771	peptide maturation protein PmbA	1416	500, 1455
lpg1779	hypothetical protein	402	500, 1455
lpg1782	flagellar biosynthesis sigma factor FliA	789	500, 1455
lpg1785	flagellar biosynthetic protein FlhA	2079	500, 1455
lpg1789	flagellar biosynthetic protein FliP	750	500, 1455
lpg1791	flagellar motor switch protein FliN	330	500, 1455
lpg1798	hypothetical protein	1197	500, 1455
lpg1800	regulatory protein RecX	396	500, 1455
lpg1805	DNA mismatch repair protein MutS	2598	500, 1455
lpg1808	porphobilinogen synthase	996	500, 1455
lpg1809	hypothetical protein	393	500, 1455
lpg1810	long chain fatty acid transporter	1479	500, 1455
lpg1812	ATP-dependent DNA helicase (UvrD/Rep helicase)	3231	500, 1455
lpg1813	ATPase (Mrp)	1074	500, 1455
lpg1816	major facilitator family transporter	1290	500, 1455
lpg1824	acyl CoA dehydrogenase	1170	500, 1455
lpg1836	coiled coil domain protein	1419	500, 1455
lpg1839	glycyl tRNA synthetase, beta subunit	2067	500, 1455
lpg1845	lipoprotein VacJ-like	783	500, 1455
lpg1847	glutamate-cysteine ligase	1296	500, 1455
lpg1850	rhodanese domain protein	351	500, 1455
lpg1855	peptidyl prolyl cis-trans isomerase D	1875	500, 1455
lpg1874	general secretion pathway protein L	1140	500, 1455
lpg1882	lactoylglutathione lyase	441	500, 1455
lpg1893	major facilitator family transporter	1281	500, 1455
lpg1906	transporting ATPase	534	500, 1455
lpg1910	D-alanyl-D-alanine carboxypeptidase	1263	500, 1455

lpg1911	glutamate tRNA synthetase catalytic subunit	1485	500, 1455
lpg1917	amino acid antiporter	1422	500, 1455
lpg1918	hypothetical protein	1431	500, 1455
lpg1921	glycoprotease (O-sialoglycoprotein endopeptidase)	672	500, 1455
lpg1943	hypothetical protein	258	500, 1455
lpg1944	hypothetical protein	1047	500, 1455
lpg1945	3',5'-cyclic nucleotide phosphodiesterase	984	500, 1455
lpg1999	pterin 4 alpha carbinolamine dehydratase	342	500, 1455
lpg2002	transmembrane protein YajC, preprotein translocase subunit	336	500, 1455
lpg2014	pyridoxal-5'-phosphate dependent enzyme family	696	500, 1455
lpg2015	pyrroline-5-carboxylate reductase	789	500, 1455
lpg2025	chaperone protein DnaK, heat shock protein Hsp70	1950	500, 1455
lpg2036	Maf-like protein (septum formation)	603	500, 1455
lpg2037	enolase	1269	500, 1455
lpg2040	mevalonate diphosphate decarboxylase	969	500, 1455
lpg2046	ABC transporter, ATP binding protein	738	500, 1455
lpg2049	hypothetical protein	300	500, 1455
lpg2189	drug efflux protein	954	500, 1455
lpg2193	sulfate transporter	1554	500, 1455
lpg2200	hypothetical protein	537	500, 1455
lpg2201	replication factor C subunit (activator I)	1434	500, 1455
lpg2202	hypothetical protein	333	500, 1455
lpg2207	hypothetical protein	1224	500, 1455
lpg2208	zinc binding dehydrogenase	1014	500, 1455
lpg2214	nucleoside-diphosphate sugar epimerase	924	500, 1455
lpg2231	3-oxoacyl reductase	753	500, 1455
lpg2232	3-oxoacyl-(acyl carrier protein) synthase III FabH	1011	500, 1455
lpg2242	hypothetical protein	1326	500, 1455
lpg2243	uracil phosphoribosyltransferase	645	500, 1455
lpg2245	C4-dicarboxylate transport protein	1293	500, 1455
lpg2247	DedA family protein	756	500, 1455
lpg2250	alcohol dehydrogenase, iron containing	1161	500, 1455
lpg2259	periplasmic, osmotically inducible protein Y-like	312	500, 1455
lpg2273	glycerol-3-phosphate binding periplasmic protein	1314	500, 1455
lpg2300	ankyrin repeat domain protein	1404	500, 1455
lpg2303	chorismate synthase AroC	1059	500, 1455
lpg2304	adenine specific methylase	933	500, 1455
lpg2313	hypothetical protein	1293	500, 1455
lpg2316	3-hydroxybutyrate dehydrogenase	783	500, 1455
lpg2320	hypothetical protein	474	500, 1455
lpg2323	type II secretion system protein (twitching	1122	500, 1455

	motility protein)		
lpg2325	hypothetical protein	837	500, 1455
lpg2336	peptide chain release factor 1 (RF-1)	1089	500, 1455
lpg2339	hypothetical protein	834	500, 1455
lpg2346	transcriptional regulator	939	500, 1455
lpg2350	alkylhydroperoxide reductase, AhpC/TSA family	639	500, 1455
lpg2355	amidase (enantiomer selective)	1410	500, 1455
lpg2356	transmembrane protein	834	500, 1455
lpg2358	30S ribosomal protein S21	240	500, 1455
lpg2359	hypothetical protein	444	500, 1455
lpg2393	bacterioferritin (cytochrome b1)	480	500, 1455
lpg2401	putative secreted esterase	1527	500, 1455
lpg2433	hypothetical protein	1761	500, 1455
lpg2439	NADPH-dependent FMN reductase domain protein	552	500, 1455
lpg2454	acetyltransferase, GNAT family, ElaA-like protein	447	500, 1455
lpg2457	two component response regulator	411	500, 1455
lpg2467	cytochrome c3 hydrogenase alpha chain	1293	500, 1455
lpg2468	sulfhydrogenase delta subunit	786	500, 1455
lpg2469	hydrogenase/sulfur reductase gamma subunit	846	500, 1455
lpg2493	small heat shock protein HspC2	495	500, 1455
lpg2506	sensor histidine kinase/response regulator LuxN	1263	500, 1455
lpg2507	hypothetical protein	696	500, 1455
lpg2515	structural toxin protein (hemagglutinin/hemolysin) RtxA	366	500, 1455
lpg2518	hypothetical protein	342	500, 1455
lpg2526	hypothetical protein	1368	500, 1455
lpg2530	3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase	1044	500, 1455
lpg2536	ferredoxin reductase	957	500, 1455
lpg2547	chaperonin CsaA	336	500, 1455
lpg2552	hypothetical protein	1668	500, 1455
lpg2577	hypothetical protein	759	500, 1455
lpg2581	2-oxoisovalerate dehydrogenase, E1 component, alpha and beta fusion	2271	500, 1455
lpg2586	cysteine protease, papain C1 family	1095	500, 1455
lpg2590	chromosome partitioning protein ParB (SpoOJ)	792	500, 1455
lpg2595	peptide deformylase	513	500, 1455
lpg2605	hypothetical protein	417	500, 1455
lpg2608	UDP-3-O-acyl-N-acetylglucosamine deacetylase	915	500, 1455
lpg2616	UDP-N-muramoylalanine-D-glutamate ligase	1344	500, 1455
lpg2625	carbamoyl phosphate synthase, large subunit	3204	500, 1455
lpg2630	permease	1002	500, 1455

lpg2631	aminopeptidase A/I	1485	500, 1455
lpg2634	leucine aminopeptidase	1365	500, 1455
lpg2635	integral membrane protein (putative virulence factor) MviN, possible role in motility	1572	500, 1455
lpg2643	hypothetical protein	660	500, 1455
lpg2650	50S ribosomal protein L27	279	500, 1455
lpg2655	sensory box protein, EAL domain, GGDEF domain, signal transduction protein	1116	500, 1455
lpg2660	transmembrane protein	663	500, 1455
lpg2662	pantoate-beta-alanine ligase	759	500, 1455
lpg2671	zinc protease (peptidase, M16 family)	1326	500, 1455
lpg2679	D-isomer specific 2-hydroxyacid dehydrogenase	945	500, 1455
lpg2680	UDP-N-acetylmuramyl tripeptide synthase	1029	500, 1455
lpg2700	sugar kinase	1482	500, 1455
lpg2702	stringent starvation protein A	621	500, 1455
lpg2703	ubiquinol-cytochrome c reductase, cytochrome c1	741	500, 1455
lpg2708	ferredoxin 2Fe-2S protein	216	500, 1455
lpg2710	phenylalanyl tRNA synthetase, beta subunit	2382	500, 1455
lpg2711	phenylalanyl tRNA synthetase, alpha subunit	1029	500, 1455
lpg2713	translational initiation factor IF-3	390	500, 1455
lpg2725	inner membrane protein	630	500, 1455
lpg2736	uroporphyrinogen III methylase	753	500, 1455
lpg2757	hypothetical protein	861	500, 1455
lpg2760	DNA-binding response regulator	729	500, 1455
lpg2762	hypothetical protein	717	500, 1455
lpg2765	HIT family hydrolase	342	500, 1455
lpg2772	initiation factor IF2-beta (IF-2 gamma, IF-2 alpha)	2607	500, 1455
lpg2778	NADH dehydrogenase I, L subunit	1974	500, 1455
lpg2782	NADH dehydrogenase I, H subunit	1023	500, 1455
lpg2785	NADH dehydrogenase I, E subunit	504	500, 1455
lpg2788	NADH dehydrogenase I, B subunit	477	500, 1455
lpg2794	phosphoglucomutase/phosphomannomutase MrsA	1368	500, 1455
lpg2805	peptide transport protein, POT family	1503	500, 1455
lpg2806	hypothetical protein	1383	500, 1455
lpg2808	shikimate-5-dehydrogenase	798	500, 1455
lpg2814	aminopeptidase	1269	500, 1455
lpg2819	tyrosine phosphatase II superfamily protein	960	500, 1455
lpg2824	DNA repair protein RecN	1668	500, 1455
lpg2848	ribonuclease, T2 family	1014	500, 1455
lpg2859	MoxR protein (ATPase) methanol dehydrogenase regulatory protein	996	500, 1455
lpg2860	hypothetical protein	477	500, 1455
lpg2861	nitrogen regulation protein	990	500, 1455

lpg2865	6-pyruvoyl tetrahydropterin synthase, putative	495	500, 1455
lpg2867	thioesterase	393	500, 1455
lpg2869	prolipoprotein diacylglyceryl transferase	771	500, 1455
lpg2874	hypothetical protein	885	500, 1455
lpg2880	endonuclease III	636	500, 1455
lpg2890	glucose inhibited division protein B	627	500, 1455
lpg2898	cytochrome c	1614	500, 1455
lpg2903	ubiquinone/menaquinone biosynthesis methyltransferase UbiE	753	500, 1455
lpg2904	hypothetical protein	624	500, 1455
lpg2908	peptide methionine sulfoxide reductase	870	500, 1455
lpg2925	outer membrane efflux protein	1629	500, 1455
lpg2927	hypothetical protein	1350	500, 1455
lpg2933	oxidoreductase, 3-octaprenyl-4-hydroxybenzoate carboxy-lyase	1467	500, 1455
lpg2934	transcription termination factor Rho	1272	500, 1455
lpg2935	RSc1188; probable thioredoxin 1	327	500, 1455
lpg2955	integration host factor beta subunit	312	500, 1455
lpg2957	stomatin like transmembrane protein	780	500, 1455
lpg2965	peroxynitrite reductase, AhpC/Tsa family	606	500, 1455
lpg2967	superoxide dismutase	591	500, 1455
lpg2969	hypothetical protein	786	500, 1455
lpg2975	hypothetical protein	2616	500, 1455
lpg2983	ATP synthase gamma chain, ATP synthase F1 gamma chain	867	500, 1455
lpg2987	ATP synthase F0, C subunit	276	500, 1455
lpg2994	hypothetical protein	357	500, 1455
lpg2997	alkane-1-monooxygenase	1167	500, 1455
lpg2999	astacin protease	801	500, 1455
lpg0002	DNA polymerase III beta chain	1104	1455
lpg0004	DNA gyrase subunit B	2421	1455
lpg0005	peptidylarginine deiminase	1047	1455
lpg0021	alpha helix protein	480	1455
lpg0022	hypothetical protein	2124	1455
lpg0023	transmembrane protein	543	1455
lpg0025	Rcp	573	1455
lpg0028	ubiquinone biosynthesis protein COQ7	714	1455
lpg0032	leucine aminopeptidase	1194	1455
lpg0035	hypothetical protein	357	1455
lpg0037	arginine 3rd transport system periplasmic binding protein	744	1455
lpg0040	integral membrane protein	996	1455
lpg0043	hypothetical protein	861	1455
lpg0047	chloramphenicol acetyltransferase	696	1455
lpg0048	acetyltransferase	879	1455
lpg0052	carboxyphosphoenolpyruvate phosphonomutase	894	1455

lpg0075	hypothetical protein	315	1455
lpg0076	hypothetical protein	582	1455
lpg0089	hypothetical protein	447	1455
lpg0091	conserved domain protein	465	1455
lpg0094	ribose-5-phosphate isomerase A	651	1455
lpg0095	cytosolic IMP-GMP specific 5'-nucleotidase	1380	1455
lpg0100	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase	1071	1455
lpg0102	3-oxoacyl-(acyl carrier protein) synthase	1230	1455
lpg0105	cytochrome oxidase-like	597	1455
lpg0106	xanthine/uracil permease	1281	1455
lpg0110	hypothetical protein	1803	1455
lpg0111	squalene and phytoene synthases	1041	1455
lpg0115	hypothetical protein	309	1455
lpg0117	glycine cleavage system H protein	378	1455
lpg0119	hypothetical protein	480	1455
lpg0122	ABC transporter, ATP binding protein	1299	1455
lpg0125	GTP binding protein EngB	603	1455
lpg0129	methylmalonate-semialdehyde dehydrogenase	1527	1455
lpg0130	hypothetical protein	2469	1455
lpg0138	glyceraldehyde 3-phosphate dehydrogenase	1029	1455
lpg0140	hypothetical protein	1335	1455
lpg0153	hypothetical protein	315	1455
lpg0183	amine oxidase, flavin containing	1497	1455
lpg0188	acyl CoA transferase/carnitine dehydratase	1077	1455
lpg0194	catalase/(hydro)peroxidase KatG	2250	1455
lpg0197	hypothetical protein	318	1455
lpg0206	membrane protein	414	1455
lpg0209	hypothetical protein	1968	1455
lpg0213	inner membrane protein, LrgB family protein	708	1455
lpg0217	phosphoribosylaminoimidazole carboxylase, ATPase subunit	1080	1455
lpg0227	hypothetical protein	1107	1455
lpg0229	heme oxygenase	834	1455
lpg0238	glycine betaine aldehyde dehydrogenase	1467	1455
lpg0239	4-aminobutyrate aminotransferase	1353	1455
lpg0243	short chain dehydrogenase	600	1455
lpg0244	pyridine nucleotide-disulfide oxidoreductase	1395	1455
lpg0256	conserved domain protein	1047	1455
lpg0264	hypothetical protein	699	1455
lpg0267	magnesium and cobalt transport protein CorA	1053	1455
lpg0269	hypothetical protein	1563	1455
lpg0271	bifunctional pyrazinamidase/nicotinamidase	642	1455

lpg0276	Ras GEF	1500	1455
lpg0282	hypothetical protein	777	1455
lpg0295	mannose-1-phosphate guanyltransferase	663	1455
lpg0298	peptidyl-prolyl cis-trans isomerase D (SurA)	1341	1455
lpg0299	pyridoxal phosphate biosynthetic protein PdxA	975	1455
lpg0301	hypothetical protein	546	1455
lpg0319	50S ribosomal protein L1	696	1455
lpg0320	50S ribosomal protein L10	534	1455
lpg0321	50S ribosomal protein L7/L12	381	1455
lpg0322	DNA-directed RNA polymerase beta subunit	4107	1455
lpg0323	DNA-directed RNA polymerase beta' subunit	4248	1455
lpg0324	30S ribosomal protein S12	381	1455
lpg0325	30S ribosomal protein S7	528	1455
lpg0330	50S ribosomal protein L4	609	1455
lpg0332	50S ribosomal protein L2	828	1455
lpg0335	30S ribosomal protein S3	657	1455
lpg0336	50S ribosomal protein L16/(L10E)	414	1455
lpg0337	50S ribosomal subunit protein L29	195	1455
lpg0338	30S ribosomal protein S17	255	1455
lpg0339	50S ribosomal protein L14	366	1455
lpg0340	50S ribosomal protein L24	330	1455
lpg0341	50S ribosomal protein L5	561	1455
lpg0343	30S ribosomal protein S8	390	1455
lpg0347	50S ribosomal protein L30/(L7E)	186	1455
lpg0348	50S ribosomal protein L15	435	1455
lpg0349	preprotein translocase SecY	1335	1455
lpg0353	30S ribosomal protein S4	621	1455
lpg0355	50S ribosomal protein L17	384	1455
lpg0356	single strand binding protein	489	1455
lpg0357	major facilitator family transporter	1368	1455
lpg0359	acyl carrier protein	414	1455
lpg0361	3-oxoacyl-(acyl carrier protein) synthase II, C-terminal	1293	1455
lpg0363	lipid A biosynthesis acyltransferase	846	1455
lpg0365	hypothetical protein	2691	1455
lpg0366	diaminopimelate epimerase	834	1455
lpg0369	carboxylesterase/phospholipase	678	1455
lpg0370	oligoketide cyclase/lipid transporter protein	435	1455
lpg0371	hypothetical protein	273	1455
lpg0372	small protein A, tmRNA-binding	345	1455
lpg0374	hypothetical protein	387	1455
lpg0377	hypothetical protein	744	1455
lpg0380	hypothetical protein	720	1455

lpg0382	osmotically inducible protein Y	567	1455
lpg0384	excinuclease ABC A subunit	2856	1455
lpg0387	ABC transporter, permease protein	774	1455
lpg0391	SM20-related protein	540	1455
lpg0392	zinc metalloprotease	708	1455
lpg0393	hypothetical protein	864	1455
lpg0394	methylated DNA protein cysteine S-methyltransferase	456	1455
lpg0395	50S ribosomal protein L19	366	1455
lpg0396	tRNA (guanine N1) methyltransferase	771	1455
lpg0399	30S ribosomal protein S16	261	1455
lpg0400	signal recognition particle protein Ffh	1377	1455
lpg0404	amino acid antiporter	1404	1455
lpg0405	hypothetical protein	591	1455
lpg0406	hypothetical protein	342	1455
lpg0407	hypothetical protein	444	1455
lpg0413	hypothetical, SCO1/SenC family protein	642	1455
lpg0418	6-phosphogluconate dehydratase	1839	1455
lpg0421	D-xylose (galactose, arabinose)-proton symporter	1422	1455
lpg0422	glucoamylase	1350	1455
lpg0423	transcriptional regulator, cro family	237	1455
lpg0424	hypothetical protein	540	1455
lpg0425	ferrochelatase	999	1455
lpg0426	cold shock protein CspD	234	1455
lpg0428	glyoxylase domain hypothetical protein	429	1455
lpg0432	hypothetical protein	900	1455
lpg0433	hypothetical protein	381	1455
lpg0440	hypothetical protein	213	1455
lpg0442	IcmS	345	1455
lpg0443	IcmR	363	1455
lpg0444	IcmQ	600	1455
lpg0445	IcmP (DotM)	1143	1455
lpg0446	IcmO (DotL)	2352	1455
lpg0447	LphA (DotK)	570	1455
lpg0448	IcmM (DotJ)	285	1455
lpg0449	IcmL (DotI)	639	1455
lpg0450	IcmK (DotH)	1086	1455
lpg0452	IcmG (DotF)	810	1455
lpg0454	IcmD (DotP)	282	1455
lpg0455	IcmJ (DotN)	645	1455
lpg0457	TphA (ProP)	1257	1455
lpg0458	IcmF	2922	1455
lpg0459	IcmH (DotU)	786	1455
lpg0460	phosphoribosylamineimidazolecarboxamide formyltransferase	1590	1455
lpg0462	acetyl CoA carboxylase, biotin carboxylase	1350	1455

	subunit		
lpg0468	lipase A	852	1455
lpg0469	endonuclease/exonuclease/phosphatase family protein	774	1455
lpg0471	phenol hydroxylase	747	1455
lpg0473	hypothetical protein	297	1455
lpg0475	sugar transport PTS system phosphocarrier HPr protein	270	1455
lpg0476	sigma-54 modulation protein	300	1455
lpg0478	50S ribosomal protein L33	165	1455
lpg0482	endo-1,4 beta-glucanase	1164	1455
lpg0491	amino acid (glutamine) ABC transporter, periplasmic amino acid binding protein	735	1455
lpg0498	leucine-, isoleucine-, valine-, threonine-, and alanine-binding protein	1179	1455
lpg0499	carboxy-terminal protease	1338	1455
lpg0500	peptidase, M23/M37 family	1170	1455
lpg0506	outer membrane protein	2361	1455
lpg0507	outer membrane protein OmpH	501	1455
lpg0510	(3R)-hydroxymyristoyl-(acyl carrier protein) dehydratase	453	1455
lpg0511	acyl-(acyl carrier protein)-UDP-N-acetylglucosamine acyltransferase	771	1455
lpg0512	CrcB protein, camphor resistance	405	1455
lpg0513	seryl tRNA synthetase	1281	1455
lpg0528	succinate dehydrogenase cytochrome b556 subunit C	393	1455
lpg0530	succinate dehydrogenase flavoprotein subunit A	1770	1455
lpg0534	succinyl CoA synthetase beta chain	1221	1455
lpg0535	succinyl CoA synthetase alpha chain	876	1455
lpg0539	hypothetical protein	390	1455
lpg0542	DNA binding protein Fis	282	1455
lpg0547	outer membrane lipoprotein LolB	600	1455
lpg0548	phosphopantetheine adenylyltransferase	600	1455
lpg0552	suppressor of GroEL (SugE)	321	1455
lpg0556	hypothetical protein	621	1455
lpg0558	stearoyl-CoA-9-desaturase	1188	1455
lpg0560	acetyoacetyl CoA reductase	747	1455
lpg0561	acetyoacetyl CoA reductase	747	1455
lpg0562	hypothetical protein	399	1455
lpg0563	hypothetical protein	357	1455
lpg0566	spore maturation protein B	534	1455
lpg0568	tyrosyl tRNA synthetase	1272	1455
lpg0577	transferase	537	1455
lpg0583	phosphate transporter	1254	1455
lpg0584	hypothetical phosphate transport regulator	672	1455
lpg0585	hypothetical protein	765	1455
lpg0587	YqgF	429	1455

lpg0588	aspartate carbamoyltransferase	894	1455
lpg0591	hypothetical protein	261	1455
lpg0592	nitrogen regulatory P-II transcription regulator	375	1455
lpg0594	hypothetical protein	186	1455
lpg0602	ATP transporter, ABC binding component, ATP-binding protein	753	1455
lpg0604	aminotransferase	1245	1455
lpg0605	nitrogen fixation protein (Fe-S cluster formation) NifU	450	1455
lpg0608	hypothetical SAM-dependent methyltransferase	915	1455
lpg0612	alcohol dehydrogenase (NADP-dependent, zinc-type)	1047	1455
lpg0614	hypothetical protein	405	1455
lpg0616	GTP cyclohydrolase I PLUS perhaps regulatory protein	1251	1455
lpg0624	hypothetical protein	378	1455
lpg0626	DNA uptake/competence protein ComA	2208	1455
lpg0627	type IV pilin	450	1455
lpg0630	type IV fimbrial biogenesis PilW related protein, transmembrane)	1068	1455
lpg0640	heat shock protein, HslVU, proteasome-related peptidase subunit	549	1455
lpg0641	ATP dependent Hsl protease, ATP binding subunit	1344	1455
lpg0643	ribonuclease BN	1239	1455
lpg0651	malate oxidoreductase	1236	1455
lpg0652	major facilitator family transporter	1293	1455
lpg0654	DNA adenine methylase	819	1455
lpg0656	tryptophan/tyrosine permease	1191	1455
lpg0657	outer membrane protein, OmpA family protein	750	1455
lpg0658	HlyD family secretion protein	858	1455
lpg0659	ABC transporter ElsE	1743	1455
lpg0660	ABC transporter permease protein	1122	1455
lpg0663	soluble lytic murein transglycosylase	1821	1455
lpg0665	putative transmembrane protein	465	1455
lpg0672	acetoacetate decarboxylase ADC	765	1455
lpg0673	signal peptide protein	273	1455
lpg0674	adenylate cyclase	1314	1455
lpg0677	hypothetical protein	267	1455
lpg0678	arginine ABC transporter, periplasmic binding protein	765	1455
lpg0679	adenylyl transferase	2742	1455
lpg0680	dipeptidyl aminopeptidase/acylaminoacyl peptidase	1269	1455
lpg0685	Fe-S oxidoreductase	1308	1455
lpg0687	Hsp10, 10 kDa chaperonin GroES	291	1455
lpg0688	Hsp60, 60K heat shock protein HtpB	1653	1455
lpg0692	ABC type dipeptide/oligopeptide/nickel	1821	1455

	transport, ATPase component		
lpg0698	hypothetical protein	852	1455
lpg0699	outer membrane protein TolC	1368	1455
lpg0704	enhanced entry protein EnhA	612	1455
lpg0712	endo-1,4-beta-xylanase-like	696	1455
lpg0716	hypothetical protein	1014	1455
lpg0719	valyl tRNA synthase	2766	1455
lpg0720	multidrug resistance protein	3048	1455
lpg0721	RND efflux membrane fusion protein, acriflavin resistance protein E	1263	1455
lpg0722	hypothetical protein	405	1455
lpg0723	hypothetical, His rich	438	1455
lpg0724	hypothetical periplasmic or secreted lipoprotein	312	1455
lpg0725	serine hydroxymethyltransferase	1254	1455
lpg0730	transmembrane permease	1053	1455
lpg0732	hypothetical protein	639	1455
lpg0734	glutamine dependent NAD+ synthetase	1611	1455
lpg0737	hypothetical signal peptide protein	435	1455
lpg0738	replicative DNA helicase	1383	1455
lpg0739	alanine racemase	1074	1455
lpg0740	17kDa common antigen	450	1455
lpg0741	hypothetical protein	525	1455
lpg0742	hypothetical protein	1254	1455
lpg0747	hypothetical protein	819	1455
lpg0749	imidazole glycerol phosphate synthase, cyclase subunit HisF	765	1455
lpg0752	N-acetylneuraminic acid synthetase	1071	1455
lpg0753	polysialic acid biosynthesis	1134	1455
lpg0754	acetyltransferase	609	1455
lpg0755	pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biosynthesis	1503	1455
lpg0759	glucose-6-phosphate isomerase	1503	1455
lpg0781	global regulator (carbon storage regulator)	249	1455
lpg0786	cell cycle protein MesJ	1302	1455
lpg0791	macrophage infectivity potentiator (Mip)	708	1455
lpg0802	sulfate transporter	1704	1455
lpg0804	choloylglycine hydrolase	999	1455
lpg0810	hypothetical protein	318	1455
lpg0811	rod shape determining protein MreB	1044	1455
lpg0815	hypothetical protein	699	1455
lpg0816	isocitrate dehydrogenase, NADP-dependent	1269	1455
lpg0818	ATP binding protease component ClpA	2274	1455
lpg0826	exonuclease VII, large subunit	1332	1455
lpg0829	two component histidine kinase, GGDEF domain protein/EAL domain protein	1920	1455
lpg0833	indole-3-glycerol phosphate synthase	777	1455

lpg0835	anthranilate synthase component II	579	1455
lpg0836	ABC transporter, ATP binding protein	726	1455
lpg0837	hypothetical protein	510	1455
lpg0838	hypothetical protein	570	1455
lpg0842	toluene tolerance protein Ttg2B	783	1455
lpg0843	toluene tolerance protein Ttg2C	477	1455
lpg0845	hypothetical protein	282	1455
lpg0846	hypothetical BolA like protein	246	1455
lpg0847	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	1269	1455
lpg0848	hypothetical TIGR00486	759	1455
lpg0851	membrane fusion protein	1017	1455
lpg0852	hypothetical protein	594	1455
lpg0856	heme exporter protein CcmA	741	1455
lpg0858	heme exporter protein CcmC	792	1455
lpg0859	cytochrome c-type biogenesis protein CcmD	132	1455
lpg0860	cytochrome c-type biogenesis protein CcmE	432	1455
lpg0862	thiol:disulfide interchange protein DsbE	534	1455
lpg0867	ATP-dependent DNA helicase RecQ	1827	1455
lpg0869	3-hydroxyisobutyryl Coenzyme A hydrolase	780	1455
lpg0872	peptide chain release factor 3	1581	1455
lpg0874	NAD(P) transhydrogenase	1422	1455
lpg0875	transmembrane NAD(P) transhydrogenase	297	1455
lpg0877	hypothetical transporter	555	1455
lpg0878	hypothetical protein	300	1455
lpg0882	hypothetical protein	435	1455
lpg0887	N-succinyl-diaminopimelate desuccinylase	1134	1455
lpg0888	2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase DapD	831	1455
lpg0889	1-acyl-sn-glycerol-3-phosphate acyltransferase	894	1455
lpg0892	kynurenine 3-monooxygenase	1350	1455
lpg0896	hypothetical protein	378	1455
lpg0897	Na/Ca antiporter	960	1455
lpg0899	A/G specific adenine glycosylase	1068	1455
lpg0900	hypothetical protein	1569	1455
lpg0902	hypothetical protein	903	1455
lpg0904	hydrolase, isochorismatase family	546	1455
lpg0905	3-oxoacyl-(acyl carrier protein) reductase	744	1455
lpg0906	flagellar biosynthesis/type III secretory pathway chaperone	498	1455
lpg0907	negative regulator of flagellin synthesis	321	1455
lpg0908	flagella basal body P-ring formation protein FlgA	702	1455
lpg0909	cytochrome c5	408	1455
lpg0910	enhanced entry protein EnhA	558	1455
lpg0915	cell division transmembrane protein FtsL	279	1455

lpg0917	UDP-N-acetylmuramyl-tripeptide synthetase MurE	1452	1455
lpg0918	erythronate-4-phosphate dehydrogenase	1053	1455
lpg0920	phosphatidylglycerophosphatase B	648	1455
lpg0925	penicillin binding protein 1A	2385	1455
lpg0926	hypothetical protein	1011	1455
lpg0927	type IV pilus biogenesis protein PilM	1065	1455
lpg0928	type IV pilus biogenesis protein PilN	549	1455
lpg0929	type IV pilus biogenesis protein PilO	636	1455
lpg0932	shikimate kinase	528	1455
lpg0933	3-dehydroquinate synthetase	1110	1455
lpg0934	DamX-related protein	1452	1455
lpg0935	universal stress protein A (UspA)	432	1455
lpg0937	isoleucyl tRNA synthetase	2796	1455
lpg0938	lipoprotein signal peptidase	366	1455
lpg0940	LidA	2190	1455
lpg0941	hypothetical protein	3126	1455
lpg0942	GTP-binding protein Era	936	1455
lpg0943	DNA repair protein RecO	690	1455
lpg0946	pyridoxal phosphate biosynthetic protein PdxJ	807	1455
lpg0949	carrier/transport protein	675	1455
lpg0951	TldD protein	1443	1455
lpg0953	AMP-binding protein	1668	1455
lpg0955	transmembrane protein	1263	1455
lpg0956	hypothetical protein	1179	1455
lpg0960	peptide ABC transporter, permease protein	978	1455
lpg0961	peptide ABC transporter, permease protein	1374	1455
lpg0962	DNA polymerase III, alpha subunit	3447	1455
lpg0966	nucleoside-diphosphate sugar epimerases	1878	1455
lpg0970	amino acid permeases	1464	1455
lpg1117	hypothetical protein	474	1455
lpg1119	major acid phosphatase	1065	1455
lpg1122	membrane bound lytic murein transglycosylase D	1329	1455
lpg1131	cyclopropane fatty acid synthase	1167	1455
lpg1135	bacterial regulatory proteins, TetR family	612	1455
lpg1136	hypothetical protein	903	1455
lpg1137	hypothetical protein	969	1455
lpg1138	spermidine/putrescine-binding periplasmic protein PotD	1023	1455
lpg1144	hypothetical protein	507	1455
lpg1146	thermostable carboxypeptidase 1	1482	1455
lpg1147	hypothetical protein	504	1455
lpg1148	hypothetical protein	1512	1455
lpg1155	pyruvate decarboxylase	1680	1455
lpg1161	phosphoribosyltransferase	663	1455

lpg1164	acetylornithine deacetylase	1155	1455
lpg1165	uridine kinase	798	1455
lpg1166	hypothetical protein	2013	1455
lpg1167	hypothetical protein	522	1455
lpg1171	hypothetical protein	420	1455
lpg1172	TPR repeat protein	1488	1455
lpg1174	two component response regulator PilR	1329	1455
lpg1176	Zn-dependent protease	1449	1455
lpg1178	riboflavin synthase, alpha subunit RibE	615	1455
lpg1186	competence lipoprotein ComL	783	1455
lpg1188	Kup system potassium uptake protein	1896	1455
lpg1189	hypothetical protein	1002	1455
lpg1190	SAM-dependent methyltransferase	1173	1455
lpg1191	glycosyl hydrolase family 3	1188	1455
lpg1195	phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase	720	1455
lpg1196	amidotransferase HisH	600	1455
lpg1197	histidinol phosphatase and imidazoleglycerol-phosphate dehydratase = bifunctional protein HisB	1059	1455
lpg1203	cytochrome D ubiquinol oxidase, subunit II	1137	1455
lpg1205	cold shock domain family protein CspA	240	1455
lpg1206	sigma 54 modulation protein YhbH	573	1455
lpg1207	hypothetical protein	456	1455
lpg1208	transcriptional regulator MarR family	420	1455
lpg1212	IAA acetyltransferase/MarR transcriptional regulatory protein	972	1455
lpg1214	2-acylglycerophosphoethanolamine acyltransferase	1308	1455
lpg1215	oxygen-dependent coproporphyrinogen III oxidase	951	1455
lpg1216	flagellar basal body rod protein FlgB	393	1455
lpg1217	flagellar basal body rod protein FlgC	423	1455
lpg1220	flagellar basal body rod protein FlgF	747	1455
lpg1221	flagellar basal body rod protein FlgG	786	1455
lpg1277	ABC transporter ATP binding protein	1827	1455
lpg1278	hypothetical protein	318	1455
lpg1279	hypothetical protein	372	1455
lpg1281	hypothetical protein	384	1455
lpg1282	stationary phase surival protein SurE	756	1455
lpg1284	stationary phase specific sigma factor RpoS	1080	1455
lpg1286	YebC	744	1455
lpg1288	Holliday junction DNA helicase RuvA	600	1455
lpg1291	two component sensor kinase	1416	1455
lpg1292	DNA-binding response regulator	678	1455
lpg1293	intracellular septation protein A	546	1455
lpg1294	membrane bound lytic murein	1440	1455

	transglycosylase D		
lpg1297	5,10-methylenetetrahydrofolate dehydrogenase	855	1455
lpg1300	integral membrane protein	468	1455
lpg1301	oxidoreductase	1299	1455
lpg1303	phosphoribosyl anthranilate isomerase	624	1455
lpg1305	tryptophan synthetase, alpha chain TrpA	819	1455
lpg1306	glutaminyl-tRNA synthetase	1656	1455
lpg1307	cysteinyl-tRNA synthetase	1371	1455
lpg1320	type II protein secretion LspD	2376	1455
lpg1324	multidrug resistance efflux pump	1215	1455
lpg1331	protease DO	1401	1455
lpg1332	hypothetical protein	726	1455
lpg1333	ribosomal large subunit pseudouridine synthase D, RluD	966	1455
lpg1334	tRNA thiotransferase	1344	1455
lpg1336	enhanced entry protein EnhA	747	1455
lpg1337	flagellar protein FliS	411	1455
lpg1339	hypothetical protein	282	1455
lpg1341	acetyl CoA carboxylase, carboxyltransferase, beta subunit	885	1455
lpg1342	FolC bifunctional protein	1287	1455
lpg1344	colicin V	534	1455
lpg1346	DNA polymerase III, delta subunit	1026	1455
lpg1349	apolipoprotein N-acyltransferase	1536	1455
lpg1350	L-lysine dehydrogenase	1137	1455
lpg1352	3-hydroxyacyl CoA dehydrogenase oxidoreductase protein/	2370	1455
lpg1359	general secretion pathway protein LspJ	618	1455
lpg1363	type II protein secretion LspF	1200	1455
lpg1369	chaperone Hsp90 HtpG	1872	1455
lpg1370	hypothetical protein	297	1455
lpg1372	oxidoreductase	963	1455
lpg1373	ribonuclease HII	576	1455
lpg1375	penicillin binding protein 2	1881	1455
lpg1376	hypothetical protein	471	1455
lpg1377	hypothetical protein	339	1455
lpg1385	hypothetical protein	378	1455
lpg1388	hypothetical protein	393	1455
lpg1391	50S ribosomal protein L32	192	1455
lpg1392	fatty acid/phospholipid synthesis protein PlsX	1029	1455
lpg1397	beta-ketoacyl-acyl carrier protein synthase II	1239	1455
lpg1398	periplasmic solute-binding protein	999	1455
lpg1401	type 4 fimbrial biogenesis protein PilZ	342	1455
lpg1403	hypothetical protein	1101	1455
lpg1404	major facilitator family transporter	1293	1455

lpg1405	multidrug translocase MdfA, chloramphenicol resistance pump Cmr	1281	1455
lpg1406	glycosyltransferase	1176	1455
lpg1408	choline kinase	1152	1455
lpg1409	hypothetical protein	771	1455
lpg1410	transcriptional regulatory protein	597	1455
lpg1411	adenylate kinase	711	1455
lpg1415	citrate synthase	1272	1455
lpg1416	purine nucleoside phosphorylase II	882	1455
lpg1417	DNA gyrase, A subunit	2616	1455
lpg1419	3-phosphoshikimate 1-carboxyvinyltransferase	1302	1455
lpg1421	30S ribosomal protein S1	1740	1455
lpg1429	hypothetical protein	537	1455
lpg1430	4-hydroxybenzoate octaprenyltransferase UbiA	849	1455
lpg1432	FAD linked oxidase	1797	1455
lpg1435	cytidine deaminase	396	1455
lpg1441	phosphate starvation-inducible protein PhoH	951	1455
lpg1444	tryptophanyl tRNA synthetase	1218	1455
lpg1445	hypothetical protein	792	1455
lpg1446	hypothetical protein	591	1455
lpg1447	pseudouridine synthase	747	1455
lpg1451	hypothetical protein	312	1455
lpg1453	hypothetical protein	519	1455
lpg1455	phospholipase C	1257	1455
lpg1456	23S rRNA (uracil-5-)methyltransferase RumA	1335	1455
lpg1459	aspartate aminotransferase	1194	1455
lpg1460	hypothetical protein	810	1455
lpg1461	single stranded DNA specific exonuclease RecJ	1740	1455
lpg1462	zinc binding TIM barrel protein, YjbN family	1005	1455
lpg1463	preprotein translocase; secretion protein SecA	2709	1455
lpg1466	hypothetical protein	747	1455
lpg1472	biotin synthase BioB	948	1455
lpg1473	8-amino-7-oxononanoate synthase	1146	1455
lpg1474	biotin biosynthesis protein BioH	720	1455
lpg1475	dethiobiotin synthetase	639	1455
lpg1476	hypothetical protein	315	1455
lpg1477	transmembrane protein	630	1455
lpg1482	hypothetical protein	837	1455
lpg1483	serine/threonine-protein kinase	1590	1455
lpg1484	hypothetical protein	810	1455
lpg1487	acetyltransferase, GNAT family	531	1455
lpg1502	dihydrolipoamide dehydrogenase	1440	1455

lpg1504	pyruvate dehydrogenase E1 component oxidoreductase protein AceE	2682	1455
lpg1505	hypothetical protein	390	1455
lpg1506	inner membrane protein AmpE	780	1455
lpg1511	lipoate-protein ligase B	600	1455
lpg1513	type I secretion system LssZ	615	1455
lpg1517	HlyD family secretion protein	1137	1455
lpg1519	purine/pyrimidine phosphoribosyltransferase	570	1455
lpg1520	hypothetical protein	327	1455
lpg1524	type 4 (IV) prepilin-like protein leader peptide processing enzyme PilD	870	1455
lpg1527	hypothetical protein	456	1455
lpg1529	2-methylcitrate dehydratase PrpD	1473	1455
lpg1530	2-methylcitrate synthase	1119	1455
lpg1535	rubredoxin (rubredoxin-type Fe(Cys)4 protein)	177	1455
lpg1536	transmembrane protein	423	1455
lpg1539	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase FolK	426	1455
lpg1541	GTP-binding protein EngA	1389	1455
lpg1542	PQQ (pyrrolo quinoline) WD40-like repeat, enzyme repeat domain protein	1164	1455
lpg1545	DNA-binding protein, putative	570	1455
lpg1546	fimbrial biogenesis and twitching motility protein PilF	783	1455
lpg1547	radical SAM enzyme, Cfr family	1161	1455
lpg1549	hypothetical protein	885	1455
lpg1550	tRNA-(ms(2)io(6)a)-hydrolase(tRNA hydroxylase)	636	1455
lpg1553	septum site determining protein MinC	711	1455
lpg1554	long chain fatty acid-CoA ligase	1710	1455
lpg1559	pyruvate dehydrogenase E1 beta subunit	975	1455
lpg1564	integral membrane protein	744	1455
lpg1566	thiamine biosynthesis oxidoreductase ThiO	1071	1455
lpg1567	thiamine (thiazole) biosynthesis protein ThiG	792	1455
lpg1568	phosphomethylpyrimidine kinase ThiD/thiamin-phosphate pyrophosphorylase fused protein ThiE	1488	1455
lpg1573	biopolymer transport protein TolR	456	1455
lpg1575	esterase	396	1455
lpg1577	RNA polymerase sigma E factor RpoE	564	1455
lpg1578	hypothetical protein	450	1455
lpg1579	glycine cleavage T protein	1059	1455
lpg1580	cytochrome b-561 transmembrane protein	549	1455
lpg1582	hypothetical protein	441	1455
lpg1584	(CDP-alcohol) phosphatidyltransferase	768	1455
lpg1585	hypothetical protein	489	1455

lpg1587	hypothetical thiol-disulfide isomerase and thioredoxins family	558	1455
lpg1589	50S ribosomal protein L9	450	1455
lpg1592	30S ribosomal protein S6	339	1455
lpg1593	carbon storage regulator CsrA	198	1455
lpg1595	hypothetical protein	735	1455
lpg1596	enoyl CoA hydratase	2019	1455
lpg1605	hypothetical protein	480	1455
lpg1612	transcriptional regulator SkgA, mercury resistance	753	1455
lpg1618	beta-lactamase AmpS	843	1455
lpg1623	hydrogenase	1032	1455
lpg1624	alpha/beta hydrolase	984	1455
lpg1636	acetyltransferase, GNAT family	849	1455
lpg1638	drug:proton antiporter	1314	1455
lpg1640	transmembrane protein	558	1455
lpg1645	hypothetical protein	576	1455
lpg1650	myo-inositol catabolism protein IolD	1872	1455
lpg1653	D-xylose-proton symporter	1416	1455
lpg1656	hypothetical protein	954	1455
lpg1657	NG,NG-dimethylarginine dimethylaminohydrolase	768	1455
lpg1661	hypothetical protein	1119	1455
lpg1662	putative transport protein	597	1455
lpg1663	hypothetical protein	507	1455
lpg1667	hypothetical protein	1392	1455
lpg1679	hypothetical protein	714	1455
lpg1682	oxidoreductase, short chain dehydrogenase/reductase family	852	1455
lpg1696	proline dehydrogenase/delta-1-pyrroline-5-carboxylate dehydrogenase = bifunctional PutA protein	3165	1455
lpg1697	hypothetical protein	678	1455
lpg1698	ProQ-like, activator of ProP osmoprotectant transporter	372	1455
lpg1699	3-demethylubiquinone-9 3-methyltransferase UbiG	693	1455
lpg1701	kinectin 1 (kinesin receptor)	1683	1455
lpg1705	carboxypeptidase G2	1224	1455
lpg1706	arginine/ornithine succinyltransferase	1044	1455
lpg1707	succinylglutamic-5-semialdehyde dehydrogenase	1491	1455
lpg1710	hypothetical protein	366	1455
lpg1711	ribosome recycling factor	558	1455
lpg1712	uridylate kinase	744	1455
lpg1713	translation elongation factor Ts (EF-Ts)	900	1455
lpg1714	30S ribosomal protein S2	816	1455
lpg1720	protein-PII uridylyltransferase	2586	1455
lpg1723	inosine-5'-monophosphate dehydrogenase	1473	1455

lpg1724	septum site-determining protein MinD	831	1455
lpg1725	similar to cell division inhibitor MinE, putative pseudogene	221	1455
lpg1727	hydrolase	786	1455
lpg1731	sn-glycerol-3-phosphate transmembrane ABC transporter	879	1455
lpg1732	quinone oxidoreductase	1002	1455
lpg1733	chloride channel protein EriC (voltage gated)	1284	1455
lpg1734	anthranilate synthase (glutamine amidotransferase) component I	2157	1455
lpg1743	Fis transcriptional activator	288	1455
lpg1747	RNA methyltransferase	774	1455
lpg1749	signal peptide peptidase	957	1455
lpg1750	ClpB protein	2577	1455
lpg1751	hypothetical protein	1311	1455
lpg1752	hypothetical protein	648	1455
lpg1753	UDP-N-acetylmuramate:L-alanyl-gamma-D-glutamyl-meso-diaminopimelate ligase	1368	1455
lpg1754	hypothetical protein	621	1455
lpg1755	transmembrane protein	3822	1455
lpg1757	nucleotide binding protein FliI	1353	1455
lpg1758	flagellar assembly protein FliH	639	1455
lpg1762	sigma 54-dependent response regulator	1362	1455
lpg1764	ATPase associated with chromosome architecture	1305	1455
lpg1765	outer membrane lipoprotein carrier protein	615	1455
lpg1766	cell division protein FtsK	2385	1455
lpg1767	thioredoxin reductase	1050	1455
lpg1768	leucyl/phenylalanyl-tRNA protein transferase	669	1455
lpg1770	translation initiation factor IF-1	222	1455
lpg1778	peptide chain release factor 2 (RF-2)	1008	1455
lpg1788	flagellar biosynthetic protein FliQ	270	1455
lpg1792	flagellar protein	1068	1455
lpg1793	hypothetical protein	255	1455
lpg1803	hypothetical protein	936	1455
lpg1804	hypothetical 17.2kDa protein, CinA-related competence damage protein	495	1455
lpg1806	outer membrane protein	1689	1455
lpg1807	periplasmic protein	2541	1455
lpg1814	hypothetical protein	555	1455
lpg1815	hydrogen peroxide-inducible genes activator OxyR	891	1455
lpg1821	dihydroorotate oxidase	1167	1455
lpg1823	hypothetical protein	687	1455
lpg1825	acyl CoA C-acetyltransferase	1185	1455
lpg1826	hypothetical protein	321	1455
lpg1830	hydroxymethylglutaryl-CoA lyase	909	1455

lpg1831	acetoacetyl CoA synthetase	2001	1455
lpg1832	hypothetical protein	417	1455
lpg1833	D-methionine transport ATP binding protein MetN	1101	1455
lpg1834	ATP binding protein, permease protein	648	1455
lpg1835	29 kDa immunogenic protein	792	1455
lpg1837	SAM-dependent methyltransferase	732	1455
lpg1838	histidinol phosphate phosphatase	531	1455
lpg1840	glycyl tRNA synthetase, alpha subunit	924	1455
lpg1841	27 kDa outer membrane protein	786	1455
lpg1842	DNA dependent ATPase I and helicase II	2193	1455
lpg1843	proline iminopeptidase	960	1455
lpg1846	glutathione synthetase	963	1455
lpg1849	hypothetical protein	282	1455
lpg1851	hypothetical protein	663	1455
lpg1854	enoyl reductase	807	1455
lpg1858	HupB DNA binding protein HU-beta	303	1455
lpg1859	ATP-dependent protease La	2451	1455
lpg1860	ATP-dependent Clp protease, ATP binding subunit ClpX	1281	1455
lpg1861	ATP-dependent Clp protease, proteolytic subunit ClpP	645	1455
lpg1870	transmembrane protein	393	1455
lpg1871	signal peptidase I (lepB-1)	783	1455
lpg1873	membrane bound lytic murein transglycosylase	1044	1455
lpg1883	transmembrane protein	528	1455
lpg1887	hypothetical protein	354	1455
lpg1888	hypothetical protein	1332	1455
lpg1889	lipase	966	1455
lpg1891	hypothetical protein HI1736	303	1455
lpg1892	hypothetical protein	384	1455
lpg1894	chloride channel protein (voltage gated)	1314	1455
lpg1895	hypothetical protein	525	1455
lpg1896	hypothetical protein	480	1455
lpg1904	integral membrane protein	903	1455
lpg1905	ectonucleoside triphosphate diphosphohydrolase I	1182	1455
lpg1908	glutathione S-transferase	612	1455
lpg1913	6-phosphofructokinase	1245	1455
lpg1915	Tfp pilus assembly protein, major type IV pilin class A	423	1455
lpg1919	3-deoxy-manno-octulosonate cytidylyltransferase	753	1455
lpg1920	tetraacyldisaccharide-1-P-4'-kinase	180	1455
lpg1924	hypothetical protein	2793	1455
lpg1927	hypothetical protein	270	1455
lpg1942	3-hydroxyacyl CoA dehydrogenase	855	1455

lpg1949	hypothetical protein	1341	1455
lpg1993	polysaccharide deacetylase	879	1455
lpg1994	(outer) membrane bound lytic murein transglycosylase family protein	1194	1455
lpg2000	protein export protein SecF	918	1455
lpg2001	protein export protein SecD	1857	1455
lpg2004	S-adenosylmethionine:tRNA ribosyltransferase-isomerase	1050	1455
lpg2007	aspartyl protease	492	1455
lpg2009	guanosine-3, 5-bis(diphosphate)-3-pyrophosphohydrolase	2148	1455
lpg2010	guanylate kinase	630	1455
lpg2011	stress-induced protein	867	1455
lpg2012	ribonuclease PH	708	1455
lpg2013	twitching motility protein PilT	1035	1455
lpg2017	hypothetical protein	615	1455
lpg2018	hypothetical protein	282	1455
lpg2020	transcriptional regulator OruR, AraC family	1014	1455
lpg2021	adenosylhomocysteinase	1326	1455
lpg2023	carbamoyl phosphate synthase, small subunit	1125	1455
lpg2024	heat shock protein DnaJ, chaperone protein	1140	1455
lpg2027	2-keto-3-deoxy-D-arabino-heptulosonate 7-phosphate synthase	1338	1455
lpg2028	uroporphyrinogen decarboxylase	1062	1455
lpg2029	dihydroneopterin aldolase FolB, putative kinase	339	1455
lpg2031	arginyl tRNA synthetase	1770	1455
lpg2032	transporter, permease	1137	1455
lpg2033	ATP dependent DNA helicase RecG	2073	1455
lpg2034	cation efflux family protein	1164	1455
lpg2038	transmembrane protein	270	1455
lpg2039	putative mevalonate kinase	882	1455
lpg2041	radical activating enzyme	654	1455
lpg2042	outer membrane protein	969	1455
lpg2043	peptidoglycan associated lipoprotein	531	1455
lpg2044	conserved domain protein	603	1455
lpg2045	ABC transport system periplasmic substrate binding protein	924	1455
lpg2047	ABC transporter, permease	1125	1455
lpg2048	hypothetical protein	852	1455
lpg2051	isopentenyl-diphosphate delta-isomerase	1029	1455
lpg2052	hydroxymethylglutaryl CoA reductase	1299	1455
lpg2175	(2-pyrone-4,6-)dicarboxylic acid hydrolase	768	1455
lpg2176	sphingosine-1-phosphate lyase I	1827	1455
lpg2178	probable multidrug-efflux system transmembrane protein	3156	1455
lpg2186	polyketide synthase, type I	11343	1455
lpg2187	hypothetical protein	402	1455

lpg2194	(beta)-carbonic anhydrase	627	1455
lpg2203	alginate O-acetylation protein AlgJ	1560	1455
lpg2204	alginate O-acetylation protein	1422	1455
lpg2206	hypothetical protein	1101	1455
lpg2210	hypothetical protein	1050	1455
lpg2211	hypothetical protein	405	1455
lpg2212	acetylpolyamine aminohydolase	1281	1455
lpg2213	hemin binding protein Hbp	453	1455
lpg2220	hypothetical protein	1500	1455
lpg2222	TPR repeat protein, protein-protein interaction	1128	1455
lpg2225	expressed protein (GH3 homolog)	1530	1455
lpg2228	3-oxoacyl-(acyl carrier protein) synthase III	1062	1455
lpg2233	acyl carrier protein	228	1455
lpg2234	multidrug resistance protein D	1368	1455
lpg2235	sterol desaturase	1203	1455
lpg2238	transmembrane protein	387	1455
lpg2240	dipeptidyl aminopeptidase/acylaminoacyl peptidase	1209	1455
lpg2246	hypothetical protein	510	1455
lpg2248	hypothetical protein	2235	1455
lpg2249	glutamine amidotransferase, class I	699	1455
lpg2255	hypothetical protein	252	1455
lpg2256	metallo-beta-lactamase superfamily protein	1419	1455
lpg2258	hypothetical protein	291	1455
lpg2260	PHA synthase	1851	1455
lpg2261	phosphate acetyl/butyryltransferase family protein) includes: (de)hydratase mit MaoC domain)	1407	1455
lpg2262	acetate kinase	1119	1455
lpg2263	curved DNA binding protein DnaJ	891	1455
lpg2266	hypothetical protein	552	1455
lpg2267	prolidase	1239	1455
lpg2271	hypothetical protein	651	1455
lpg2272	transmembrane protein	492	1455
lpg2274	glycerophosphoryl diester esterase	720	1455
lpg2275	hypothetical protein	708	1455
lpg2276	Glu/Leu/Phe/Val dehydrogenase	1074	1455
lpg2277	O-methyltransferase, SAM-dependent	657	1455
lpg2278	4-hydroxyphenylpyruvate dioxygenase	1086	1455
lpg2279	fumarylacetoacetate hydrolase	999	1455
lpg2280	glutathione S-transferase	639	1455
lpg2281	hypothetical protein	561	1455
lpg2282	asparaginyl tRNA synthetase	1437	1455
lpg2285	lipoprotein ABC transporter	1356	1455
lpg2295	ribosomal large subunit (23S rRNA) pseudouridine synthase C	954	1455

lpg2297	ribonuclease E	2004	1455
lpg2298	inclusion membrane protein A	1278	1455
lpg2302	aspartate semialdehyde dehydrogenase	1023	1455
lpg2306	rhodanese domain protein	420	1455
lpg2307	glutaredoxin 3	255	1455
lpg2310	glutamate racemase	867	1455
lpg2312	hypothetical protein	303	1455
lpg2314	dihydropicolinate synthase	873	1455
lpg2315	hypothetical protein	285	1455
lpg2318	chemotaxis (motility protein A) transmembrane	906	1455
lpg2319	chemotaxis (motility protein B) transmembrane	939	1455
lpg2321	serine transporter	1362	1455
lpg2322	cardiac ankyrin repeat protein	1926	1455
lpg2327	CG18304 gene product	894	1455
lpg2328	hypothetical protein	384	1455
lpg2334	hypothetical protein	279	1455
lpg2335	glutamyl tRNA reductase	1383	1455
lpg2338	DnaK suppressor protein	477	1455
lpg2340	3-deoxy-D-manno-oct-2-ulosonic acid transferase	1266	1455
lpg2343	lysophospholipase A	978	1455
lpg2347	2,4-dienoyl-CoA reductase FadH1	2025	1455
lpg2348	superoxide dismutase (copper-zinc)	489	1455
lpg2352	malate dehydrogenase	993	1455
lpg2353	NUDIX hydrolase	564	1455
lpg2354	(oxygen-independent) coproporphyrinogen III oxidase	1128	1455
lpg2386	hypothetical protein	564	1455
lpg2388	amino acid permease	1731	1455
lpg2389	catalase-peroxidase KatB	2196	1455
lpg2391	SdbC	1305	1455
lpg2396	transcriptional regulator	1071	1455
lpg2404	hypothetical protein	915	1455
lpg2405	mutator MutT protein	402	1455
lpg2411	hypothetical protein	828	1455
lpg2413	hypothetical protein	438	1455
lpg2414	hypothetical protein	294	1455
lpg2434	hypothetical protein	495	1455
lpg2435	hypothetical protein	1044	1455
lpg2436	hypothetical protein	381	1455
lpg2438	florfenicol efflux pump	1191	1455
lpg2440	glutathione S-transferase	1017	1455
lpg2442	PhnB protein	411	1455
lpg2443	hypothetical protein	558	1455
lpg2445	hypothetical protein	486	1455

lpg2453	hypothetical protein	450	1455
lpg2459	guanylate cyclase	561	1455
lpg2460	hypothetical protein	384	1455
lpg2461	hypothetical protein	639	1455
lpg2463	peptide aspartate b-dioxygenase	720	1455
lpg2472	hydrogenase expression/formation protein HypD	1107	1455
lpg2473	hydrogenase expression/formation protein HypC	228	1455
lpg2475	hydrogenase expression/formation protein HypB	759	1455
lpg2476	hydrogenase nickel incorporation protein HypA	342	1455
lpg2483	hypothetical protein	558	1455
lpg2484	ribosomal protein Ham1	585	1455
lpg2485	TPR domain protein	1716	1455
lpg2487	deoxyuridinetriphosphatase	471	1455
lpg2491	hypothetical protein	564	1455
lpg2495	homospermidine synthase	1419	1455
lpg2497	hypothetical protein	654	1455
lpg2500	carbonic anhydrase Mig5	738	1455
lpg2513	RND multidrug efflux membrane fusion protein	1167	1455
lpg2514	outer membrane efflux protein (RND multidrug efflux)	1563	1455
lpg2516	major facilitator family transporter	1266	1455
lpg2517	transcriptional regulator, AsnC family	474	1455
lpg2520	hypothetical protein	369	1455
lpg2531	chorismate mutase/prephenate dehydratase (P-protein)	585	1455
lpg2532	aspartate aminotransferase	1167	1455
lpg2534	hypothetical protein	432	1455
lpg2535	myoglobin-like	408	1455
lpg2538	hypothetical protein	1416	1455
lpg2544	membrane-bound lytic murein transglycosylase A	1374	1455
lpg2549	transcriptional regulator, AraC-family	771	1455
lpg2554	rare lipoprotein A	483	1455
lpg2576	hypothetical, uroporphyrin-III C-methyltransferase	378	1455
lpg2578	hypothetical protein	255	1455
lpg2579	hypothetical protein	414	1455
lpg2580	glutaryl CoA dehydrogenase	1158	1455
lpg2585	D-alanyl-D-alanine dipeptidase	732	1455
lpg2587	probable thermolabile hemolysin	1551	1455
lpg2589	D-alanyl-D-alanine carboxypeptidase, fraction B; penicillin binding protein 4	1794	1455
lpg2592	hypothetical protein	750	1455
lpg2596	signal peptide protein, LysM domain protein	1038	1455

lpg2598	hypothetical protein	417	1455
lpg2601	hypothetical protein	441	1455
lpg2602	conserved domain protein	423	1455
lpg2604	hypothetical protein	804	1455
lpg2606	glutamine amidotransferase	867	1455
lpg2611	cell division protein FtsQ	720	1455
lpg2614	UDP-N-acetylmuramate:L-alanine ligase MurC	1410	1455
lpg2615	cell division protein FtsW	1185	1455
lpg2619	cell division protein ZipA	780	1455
lpg2621	acid phosphatase, class B	681	1455
lpg2622	hypothetical protein	1062	1455
lpg2624	transcription elongation factor GreA	483	1455
lpg2626	hypothetical protein	273	1455
lpg2628	membrane protein	753	1455
lpg2629	permease	1071	1455
lpg2632	DNA polymerase III, chi subunit	435	1455
lpg2636	30S ribosomal protein S20	267	1455
lpg2641	enhanced entry protein EnhA	723	1455
lpg2645	excinuclease ABC subunit	1857	1455
lpg2651	50S ribosomal protein L21	312	1455
lpg2652	50S ribosomal protein L25, ribosomal 5S rRNA E-loop binding protein	660	1455
lpg2653	peptidyl tRNA hydrolase	570	1455
lpg2656	octaprenyl diphosphate synthase IspB	969	1455
lpg2658	ferrous iron transporter A	228	1455
lpg2659	ATPase N2B (nucleotide (GTP) binding protein)	1092	1455
lpg2661	3-methyl-2-oxobutanoate hydroxymethyltransferase	867	1455
lpg2663	hypothetical protein	534	1455
lpg2666	probable hydrolase	882	1455
lpg2667	RNA polymerase sigma-32 factor RpoH	879	1455
lpg2668	cell division ATP transporter FtsX	930	1455
lpg2672	zinc protease (peptidase, M16 family)	1305	1455
lpg2673	N6-adenine specific methylase	546	1455
lpg2677	5'-nucleotidase	1728	1455
lpg2678	hypothetical protein	798	1455
lpg2682	hypothetical with two candidate membrane-spanning segments	708	1455
lpg2684	hypothetical protein	861	1455
lpg2687	IcmV	456	1455
lpg2688	IcmW	456	1455
lpg2690	LphB	1632	1455
lpg2692	hypothetical protein	531	1455
lpg2693	hypothetical SnoK-like protein	801	1455
lpg2694	phytanoyl-CoA dioxygenase	858	1455

lpg2696	tRNA delta(2)-isopentenylpyrophosphate transferase	966	1455
lpg2698	N-acetylmuramoyl-L-alanine amidase	1431	1455
lpg2701	stringent starvation protein B	396	1455
lpg2704	ubiquinol-cytochrome c reductase, cytochrome b	1215	1455
lpg2705	ubiquinol-cytochrome c reductase, iron-sulfur subunit	627	1455
lpg2706	30S ribosomal protein S9	432	1455
lpg2707	50S ribosomal protein L13	459	1455
lpg2709	integration host factor (IHF) alpha subunit	300	1455
lpg2712	50S ribosomal protein L20	360	1455
lpg2714	threonyl tRNA synthase	1941	1455
lpg2716	hypothetical protein	288	1455
lpg2717	hypothetical protein	486	1455
lpg2719	hypothetical protein	1152	1455
lpg2720	cNMP binding domain-containing protein	1032	1455
lpg2722	NADH-dependent flavin oxidoreductase, Oye family	1077	1455
lpg2724	hypothetical protein	345	1455
lpg2726	peptidylprolyl cis-trans isomerase B (cyclophilin-type) Lcy	495	1455
lpg2727	queuine/archaeosine tRNA-ribosyltransferase	1167	1455
lpg2732	(two component) response regulator	1026	1455
lpg2735	porphobilinogen deaminase	966	1455
lpg2737	uroporphyrinogen III methylase	1125	1455
lpg2739	cation efflux system protein	924	1455
lpg2740	hypothetical protein	663	1455
lpg2741	oligoribonuclease	564	1455
lpg2742	tRNA nucleotidyltransferase	1275	1455
lpg2743	EngC GTPase	978	1455
lpg2755	hypothetical protein	339	1455
lpg2756	recombinational DNA repair protein RecR	600	1455
lpg2758	hypothetical protein	1998	1455
lpg2763	Mg2+ and Co2+ transporter CorB, hemolysin	1266	1455
lpg2766	GTP cyclohydrolase I	567	1455
lpg2769	30S ribosomal protein S15 (S15/S13E)	276	1455
lpg2773	N utilization substance protein A	1479	1455
lpg2774	hypothetical protein	444	1455
lpg2777	NADH dehydrogenase I, M subunit	1506	1455
lpg2779	NADH dehydrogenase I, K subunit	306	1455
lpg2780	NADH dehydrogenase I, J subunit	660	1455
lpg2781	NADH dehydrogenase I, I subunit	501	1455
lpg2783	NADH dehydrogenase I, G subunit	2352	1455
lpg2786	NADH dehydrogenase I, D subunit	1269	1455
lpg2787	NADH dehydrogenase I, C subunit	684	1455

lpg2789	NADH dehydrogenase I, A subunit	357	1455
lpg2791	preprotein translocase, SecG subunit	306	1455
lpg2792	triosephosphate isomerase (TIM)	750	1455
lpg2795	7,8-dihydropteroate synthase	876	1455
lpg2796	cell division protein FtsH	1920	1455
lpg2797	ribosomal RNA large subunit methyltransferase J	744	1455
lpg2798	RNA-binding protein containing KH domina, putative pseudogene	251	1455
lpg2799	O-acetyltransferase	1977	1455
lpg2809	aminopeptidase N	2598	1455
lpg2812	sporulation protein	1524	1455
lpg2817	heat shock protein 33, redox regulated chaperonin	864	1455
lpg2818	hypothetical protein	498	1455
lpg2822	virulence regulator BipA	1827	1455
lpg2823	sugar kinase	888	1455
lpg2825	cold shock protein CspE	207	1455
lpg2827	hypothetical protein	978	1455
lpg2833	acyl-CoA thioester hydrolase	381	1455
lpg2835	thiopurine S-methyltransferase	666	1455
lpg2836	glucosamine-fructose-6-phosphate aminotransferase, isomerizing	1815	1455
lpg2837	phospholipase/lecithinase/hemolysin, lysophospholipase A, glycerophospholipid-cholesterol acyltransferase	1302	1455
lpg2838	rhodanese domain protein	765	1455
lpg2842	PhoH protein (phosphate starvation inducible protein)	1407	1455
lpg2847	hypothetical protein	963	1455
lpg2851	protoporphyrinogen oxidase	1509	1455
lpg2853	hypothetical protein, KQDN repeats	1659	1455
lpg2855	TPR (repeat) domain protein	933	1455
lpg2858	hypothetical protein	912	1455
lpg2868	thymidylate synthase (TS)	795	1455
lpg2872	(di)nucleoside polyphosphate hydrolase	528	1455
lpg2873	L-asparaginase I (cytoplasmic)	1011	1455
lpg2875	UDP-N-acetylglucosamine pyrophosphorylase	1386	1455
lpg2879	hypothetical protein	1752	1455
lpg2881	iron-sulfur cluster binding protein	615	1455
lpg2883	3-octaprenyl-4-hydroxybenzoate carboxy-lyase	570	1455
lpg2884	hypothetical protein	738	1455
lpg2885	hypothetical protein	555	1455
lpg2886	ExsB protein	702	1455
lpg2887	phosphomannose isomerase GDP mannose pyrophosphorylase	1494	1455
lpg2891	sporulation initiation inhibitor protein Soj	771	1455
lpg2894	cytochrome c oxidase, subunit III	870	1455

lpg2897	cytochrome c oxidase, subunit II	1206	1455
lpg2899	ferredoxin component, putative pseudogene	350	1455
lpg2900	CapM protein, capsular polysaccharide biosynthesis	1029	1455
lpg2901	transporter, LysE family	606	1455
lpg2905	ubiquinone biosynthesis AarF	1650	1455
lpg2907	hypothetical protein	1263	1455
lpg2916	hypothetical protein	537	1455
lpg2924	lipoprotein	1146	1455
lpg2926	bis(5'-nucleosyl)tetraphosphatase, symmetrical	846	1455
lpg2928	dimethyladenosine transferase	771	1455
lpg2929	aspartate-1-decarboxylase	402	1455
lpg2931	hypothetical protein	324	1455
lpg2937	fumarate hydratase	1395	1455
lpg2951	cystathionine beta synthase	951	1455
lpg2953	hypothetical protein	726	1455
lpg2956	deoxycytidine triphosphate deaminase	567	1455
lpg2960	major outer membrane protein	972	1455
lpg2962	sodium-type flagellar protein	900	1455
lpg2963	dihydroorotase, homodimeric type	1080	1455
lpg2964	ribonuclease T	624	1455
lpg2966	glutaredoxin-related protein	270	1455
lpg2968	N-acetylornithine aminotransferase ArgD	1170	1455
lpg2970	glycerophosphoryl diester phosphodiesterase	789	1455
lpg2971	malate dehydrogenase (NAD-linked), malic enzyme	1671	1455
lpg2972	SUA5/yciO/yrdC family:Sua5/YciO/YrdC/YwlC protein family	969	1455
lpg2974	phosphatidylserine decarboxylase	852	1455
lpg2976	hypothetical protein	1530	1455
lpg2982	H+-transporting two-sector ATPase, ATP synthase F1 subunit beta	1377	1455
lpg2985	ATP synthase F1, delta subunit	558	1455
lpg2986	ATP synthase F0, B subunit	471	1455
lpg2990	hypothetical protein	144	1455
lpg2991	hemolysin, lipoprotein	588	1455
lpg2993	phosphoheptose isomerase	600	1455
lpg2995	lipoprotein	1812	1455
lpg2996	tetrapyrrole (corrin/porphyrin) methylase	852	1455
lpg2998	sulfate transporter	2178	1455
lpg3002	inner membrane protein, 60 kDa	1671	1455

Appendix Table 15. 200 "accessory" genes used in the gene presence/absence scheme. The reference gene sequences are deposited in the ENA under the accession numbers, FJOD01000001-FJOD01000200.

Gene no.	Annotation	Length (bp)	Reference isolate
1	hypothetical protein	2895	EUL 24
2	hypothetical protein	339	EUL 24
3	hypothetical protein	1410	EUL 24
4	Fatty acid hydroxylase superfamily	873	EUL 24
5	hypothetical protein	621	EUL 24
6	hypothetical protein	1440	EUL 24
7	hypothetical protein	231	EUL 24
8	hypothetical protein	1413	EUL 24
9	Serine/threonine-protein kinase HipA	306	EUL 24
10	Heme NO binding	540	EUL 24
11	Transcriptional repressor smtB homolog	294	EUL 24
12	transcriptional repressor DicA	252	EUL 24
13	Carbon storage regulator	198	EUL 24
14	integrating conjugative element protein PilL	411	EUL 24
15	hypothetical protein	585	EUL 24
16	Phage integrase	804	EUL 24
17	Phage integrase	897	EUL 24
18	hypothetical protein	1260	EUL 24
19	hypothetical protein	666	EUL 24
20	hypothetical protein	1065	EUL 24
21	Hsp20/alpha crystallin family	567	EUL 24
22	Opacity protein and related surface antigens	783	EUL 24
23	hypothetical protein	621	EUL 24
24	conjugal transfer protein TrbB	492	EUL 24
25	conjugal pilus assembly protein TraF	780	EUL 24
26	conjugal transfer mating pair stabilization protein TraN	1809	EUL 24
27	conjugal transfer pilus assembly protein TrbC	663	EUL 24
28	hypothetical protein	2547	EUL 24
29	conjugal transfer pilus assembly protein TraB	1464	EUL 24
30	conjugal transfer protein TraK	732	EUL 24
31	Predicted acetyltransferase	1248	EUL 24
32	aminoalkylphosphonic acid N-acetyltransferase	450	EUL 24
33	Predicted acetyltransferase	1035	EUL 24
34	Domain of unknown function (DUF932)	801	EUL 24
35	hypothetical protein	384	EUL 24
36	Predicted acetyltransferase	1008	EUL 24
37	hypothetical protein	1380	EUL 36
38	phosphonate utilization associated putative membrane protein	831	EUL 36

39	Ribonuclease TTHA0252	1359	EUL 36
40	hypothetical protein	450	EUL 36
41	poly(R)-hydroxyalkanoic acid synthase	1677	EUL 36
42	Predicted membrane protein	525	EUL 36
43	Protein of unknown function (DUF2933)	294	EUL 36
44	Putative protein-S-isoprenylcysteine methyltransferase	663	EUL 36
45	hypothetical protein	819	EUL 36
46	hypothetical protein	759	EUL 36
47	3-ketosteroid-9-alpha-hydroxylase reductase subunit	1887	EUL 36
48	ATP synthase subunit alpha	1470	EUL 36
49	F-type ATPase subunit b	741	EUL 36
50	Lipid-binding protein	273	EUL 36
51	F-ATPase subunit 6	690	EUL 36
52	putative F0F1-ATPase subunit	276	EUL 36
53	F0F1 ATP synthase subunit epsilon	408	EUL 36
54	hypothetical protein	429	EUL 36
55	hypothetical protein	255	EUL 36
56	Predicted transcriptional regulator	201	EUL 36
57	Putative prophage CPS-53 integrase	1179	EUL 36
58	hypothetical protein	270	EUL 36
59	hypothetical protein	807	EUL 36
60	hypothetical protein	207	EUL 36
61	Pathogenicity locus	294	EUL 36
62	Thiocyanate hydrolase subunit beta	423	EUL 36
63	Thiocyanate hydrolase subunit gamma	666	EUL 36
64	HupE / UreJ protein	624	EUL 36
65	Opacity protein and related surface antigens	708	EUL 36
66	SNARE domain	291	EUL 36
67	Universal stress protein E homolog	936	EUL 36
68	Ankyrin repeats (3 copies)	1932	EUL 36
69	HTH-type transcriptional regulator gltR	873	EUL 36
70	Aspartate aminotransferase	1356	EUL 36
71	Proline porter II	1275	EUL 36
72	Hypoxic response protein 1	447	EUL 36
73	hypothetical protein	1434	EUL 36
74	hypothetical protein	870	EUL 36
75	type IV secretion system protein VirB3	279	EUL 36
76	Type IV secretion system protein virB4	2478	EUL 36
77	P-type DNA transfer protein VirB5	708	EUL 36
78	TrbL/VirB6 plasmid conjugal transfer protein	1038	EUL 36
79	Type IV secretion system protein virB8	714	EUL 36
80	Type IV secretion system protein virB9 precursor	750	EUL 36
81	Type IV secretion system protein virB10	1089	EUL 36
82	Conjugal transfer protein traG	1899	EUL 36
83	hypothetical protein	204	EUL 36

84	hypothetical protein	456	EUL 36
85	hypothetical protein	351	EUL 36
86	Bacterial regulatory proteins	699	EUL 36
87	Dot/Icm substrate protein	4605	EUL 36
88	hypothetical protein	318	EUL 36
89	hypothetical protein	1917	EUL 36
90	hypothetical protein	2502	EUL 36
91	Transposase	1194	EUL 36
92	Legionella pneumophila major outer membrane protein precursor	975	EUL 36
93	phenylacetate-CoA ligase	1380	EUL 36
94	hypothetical protein	522	EUL 36
95	hypothetical protein	1137	EUL 36
96	Inner membrane protein ybaL	1692	EUL 36
97	hypothetical protein	249	EUL 36
98	hypothetical protein	1206	EUL 48
99	Rieske [2Fe-2S] domain	258	EUL 48
100	hypothetical protein	1392	EUL 48
101	Flavin reductase like domain	633	EUL 48
102	hypothetical protein	1344	EUL 48
103	precorrin 6A synthase	753	EUL 48
104	hypothetical protein	195	EUL 48
105	Major Facilitator Superfamily	1224	EUL 48
106	hypothetical protein	369	EUL 48
107	Dipeptide and tripeptide permease A	1479	EUL 48
108	ATP synthase subunit beta	1422	EUL 48
109	hypothetical protein	372	EUL 48
110	hypothetical protein	165	EUL 48
111	Ribose-phosphate pyrophosphokinase	930	EUL 48
112	Pyrimidine-nucleoside phosphorylase	1515	EUL 48
113	Serine/threonine-protein kinase HipA	1302	EUL 48
114	serine/threonine protein kinase	978	EUL 48
115	hypothetical protein	948	EUL 48
116	hypothetical protein	150	EUL 48
117	Uncharacterized protein conserved in bacteria	2745	EUL 48
118	Superfamily II helicase and inactivated derivatives	1773	EUL 48
119	Regulator of chromosome condensation (RCC1) repeat	1440	EUL 48
120	hypothetical protein	816	EUL 48
121	hypothetical protein	558	EUL 48
122	hypothetical protein	900	EUL 48
123	Transposase and inactivated derivatives	1173	EUL 54
124	Antitoxin HipB	273	EUL 54
125	Phage integrase	801	EUL 54
126	hypothetical protein	381	EUL 54
127	hypothetical protein	2229	EUL 54

128	Integrase core domain	1032	EUL 54
129	GIY-YIG nuclease superfamily protein	291	EUL 55
130	Dot/Icm substrate protein	4599	EUL 55
131	Ribulose-5-phosphate 4-epimerase and related epimerases and aldolases	1752	EUL 55
132	Ribulose-5-phosphate 4-epimerase and related epimerases and aldolases	1740	EUL 55
133	GIY-YIG nuclease superfamily protein	354	EUL 55
134	hypothetical protein	750	EUL 55
135	anaerobic benzoate catabolism transcriptional regulator	261	EUL 55
136	hypothetical protein	420	EUL 55
137	hypothetical protein	900	EUL 55
138	Phage integrase	900	EUL 55
139	hypothetical protein	171	EUL 55
140	Relaxosome protein	351	EUL 123
141	Conjugal transfer protein traG	1887	EUL 123
142	conjugal transfer protein TrbJ	738	EUL 123
143	conjugal transfer protein TrbG	873	EUL 123
144	conjugal transfer protein TrbC	375	EUL 123
145	Pertussis toxin liberation protein H	963	EUL 123
146	Carbon storage regulator	258	EUL 123
147	hypothetical protein	882	EUL 123
148	Pyrimidine-nucleoside phosphorylase	1512	EUL 123
149	Phosphate acetyltransferase	1404	EUL 123
150	Enoyl-[acyl-carrier-protein] reductase [NADH] FabI	753	EUL 123
151	Spermidine N(1)-acetyltransferase	555	EUL 123
152	Tetrapyrrole (Corrin/Porphyrin) Methylases	774	EUL 123
153	hypothetical protein	183	EUL 123
154	Aminoglycoside phosphotransferase	879	EUL 123
155	hypothetical protein	1923	EUL 63
156	Sodium/proton antiporter nhaA	1152	EUL 63
157	Regulator of chromosome condensation (RCC1) repeat	1440	EUL 63
158	Integrase	897	EUL 63
159	Ankyrin repeats (3 copies)	1524	EUL 63
160	hypothetical protein	1083	EUL 63
161	DNA primase TraC	2196	EUL 63
162	Probable cadmium-transporting ATPase	1905	EUL 69
163	hypothetical protein	1365	EUL 71
164	hypothetical protein	1050	H123640643
165	Calcium-transporting ATPase	2706	EUL 88
166	hypothetical protein	366	LC6408
167	hypothetical protein	903	EUL 159
168	hypothetical protein	240	EUL 167
169	hypothetical protein	342	EUL 167
170	conjugal transfer mating pair stabilization protein	2889	H073900557

	TraG		
171	Transposase and inactivated derivatives	1209	H081180019
172	Murein tetrapeptide carboxypeptidase	939	H113660550
173	Thiocyanate hydrolase subunit alpha	309	H064380001
174	hypothetical protein	1134	H064180019
175	F0F1 ATP synthase subunit gamma	945	H073340594
176	hypothetical protein	468	H073340594
177	Transposase and inactivated derivatives	747	EUL 18
178	Cyn operon transcriptional activator	942	EUL 25
179	Cation efflux system protein CzcC	1329	EUL 25
180	Type IV secretion system protein virB11	1056	EUL 25
181	hypothetical protein	990	EUL 25
182	conjugal transfer protein TrbF	741	EUL 140
183	Uncharacterized conserved protein (contains double-stranded beta-helix domain)	858	EUL 140
184	hypothetical protein	393	EUL 140
185	Site-specific DNA methylase	1413	EUL 126
186	Aminoglycoside phosphotransferase	1614	EUL 4
187	hypothetical protein	1299	EUL 103
188	Putative prophage CPS-53 integrase	1242	EUL 111
189	acetyl-CoA acetyltransferase	1059	EUL 144
190	hypothetical protein	1950	EUL 149
191	Domain of unknown function (DUF1768)	1842	EUL 154
192	conjugal transfer protein TrbL	1440	EUL 162
193	Bacteriophytochrome cph2	2634	EUL 162
194	Superfamily II helicase and inactivated derivatives	1752	EUL 163
195	Guanine deaminase	480	HL06041035
196	hypothetical protein	450	H093380153
197	Antirestriction protein	507	H044500045
198	Ran GTPase-activating protein (RanGAP) involved in mRNA processing and transport	993	H091960009
199	Transcriptional regulatory protein RstA	726	H091960009
200	hypothetical protein	516	H071260094

Appendix Table 16. A summary of sequencing statistics for the typing panel isolates and all isolates used in the chapter (excluding the two complete genomes).

Quality criteria	**Mean (and range)**	
	Typing panel only (n=106)	*All isolates except complete genomes (n=333)*
Number of reads	4,445,116 (2,659,918 – 5,908,566)	4,427,999 (1,750,804 – 20,104,220)
Mapping depth	124.6x (71.4x - 164.2x)	122.8x (49.4x– 211.2x)
% of the reference length mapped	96.9 (92.3-100)	97.3 (92.3-100)

Assembly length (bp)	3,471,546 (3,229,839 – 3,682,698)	3,476,413 (3,229,839 – 3,710,927)
Number of contigs	35.1 (15 - 72)	39.9 (12 - 140)
N50 (bp)	250,252 (86,373 - 726,453)	249,102 (81,272 – 2,134,649)

Appendix Table 17. The number of typable loci in each isolate for each extended MLST scheme. The number of loci identified as typable by BIGSdb (i.e. only excluding absent or truncated loci) and after the additional QC steps (excluding any loci containing "Ns", any loci with <20 nucleotides, or any loci not validated by mapping data) are given.

EUL/ isolate number	**Number of alleles called pre- and post-QC**											
	rMLST (53)		**cgMLST (50)**		**cgMLST (100)**		**cgMLST (500)**		**cgMLST (1455)**		**cgMLST (1521)**	
1	53	53	50	50	100	100	500	500	1455	1455	1521	1521
2	53	53	50	50	100	100	500	500	1455	1455	1521	1521
3	53	53	50	50	100	100	500	500	1455	1455	1521	1521
4	53	53	50	50	100	100	500	500	1455	1455	1521	1518
6	53	52	50	50	100	100	500	500	1455	1455	1521	1519
7	53	53	50	50	100	100	500	500	1455	1455	1521	1519
8	53	53	50	50	100	100	500	500	1455	1455	1521	1520
13	53	53	50	50	100	100	500	500	1455	1455	1521	1520
14	53	52	50	50	100	100	500	500	1455	1455	1521	1520
16	53	53	50	50	100	100	500	500	1455	1455	1521	1521
17	53	53	50	50	100	100	500	500	1455	1455	1521	1521
18	53	52	50	50	100	100	500	500	1455	1454	1519	1517
19	53	53	50	50	100	100	500	500	1455	1455	1521	1520
20	53	53	50	50	100	100	500	500	1455	1455	1521	1520
25	53	53	50	50	100	100	500	500	1455	1455	1521	1520
26	53	53	50	50	100	100	500	500	1455	1454	1521	1520
27	53	53	50	50	100	100	500	500	1455	1455	1521	1517
27 (replicate)	53	53	50	50	100	100	500	500	1455	1455	1521	1519
28	53	53	50	50	100	100	500	500	1455	1455	1521	1519
29	53	53	50	50	100	100	500	500	1455	1455	1521	1520
30	53	53	50	50	100	100	500	500	1455	1455	1521	1521
31	53	53	50	50	100	100	500	500	1455	1455	1521	1519
32	53	53	50	50	100	100	500	500	1455	1455	1521	1521
33	53	53	50	50	100	100	500	500	1455	1455	1521	1520
33 (replicate)	53	53	50	50	100	100	500	500	1455	1455	1521	1521
36	53	53	50	50	100	100	500	500	1455	1455	1521	1520
37	53	53	50	50	100	100	500	500	1455	1455	1521	1521
38	53	53	50	50	100	100	500	500	1455	1455	1521	1520
39	53	53	50	50	100	100	500	500	1455	1455	1521	1521
40	53	53	50	50	100	100	500	500	1455	1455	1521	1520
41	53	53	50	50	100	100	500	500	1455	1455	1521	1520

42	53	53	50	50	100	100	500	500	1455	1455	1521	1520
43	53	53	50	50	100	100	500	500	1455	1455	1521	1520
48	53	52	50	50	100	100	500	500	1455	1455	1521	1519
49	53	53	50	50	100	100	500	500	1455	1455	1521	1521
50	53	53	50	48	100	98	500	497	1455	1451	1521	1514
51	53	53	50	50	100	100	500	500	1455	1454	1521	1519
52	53	53	50	50	100	100	500	500	1455	1455	1521	1521
53	53	53	50	50	100	100	500	500	1455	1455	1521	1520
54	53	53	50	50	100	100	500	499	1455	1454	1521	1519
55	53	53	50	50	100	100	500	500	1455	1455	1521	1521
60	53	53	50	50	100	100	500	500	1455	1455	1521	1520
63	53	53	50	50	100	100	500	500	1455	1455	1521	1520
66	53	53	50	50	100	100	500	500	1455	1455	1521	1520
67	53	53	50	50	100	100	500	500	1455	1455	1521	1521
68	53	52	50	50	100	100	500	500	1455	1455	1521	1520
69	53	53	50	50	100	100	500	500	1455	1455	1521	1521
69 (replicate)	53	53	50	50	100	100	500	500	1455	1455	1521	1520
70	53	53	50	50	100	100	500	500	1455	1455	1521	1521
71	53	53	50	50	100	100	500	500	1455	1455	1521	1518
72	53	53	50	50	100	100	500	500	1455	1455	1521	1521
73	53	53	50	50	100	100	500	500	1455	1455	1521	1521
74	53	53	50	50	100	100	500	500	1455	1454	1521	1521
75	53	53	50	50	100	100	500	500	1455	1455	1521	1520
75 (replicate)	53	52	50	50	100	100	500	500	1455	1455	1520	1519
81	53	53	50	50	100	100	500	500	1455	1455	1521	1521
82	53	53	50	50	100	100	500	500	1455	1454	1521	1520
83	53	53	50	50	100	100	500	500	1455	1455	1521	1520
84	53	53	50	50	100	100	500	500	1455	1455	1521	1519
85	53	53	50	50	100	100	500	500	1455	1455	1521	1520
86	53	53	50	50	100	100	500	500	1455	1455	1521	1520
87	53	53	50	50	100	100	500	500	1455	1455	1521	1521
88	53	53	50	50	100	100	500	500	1455	1455	1521	1520
91	53	53	50	50	100	100	500	500	1455	1454	1521	1520
92	53	53	50	50	100	100	500	500	1455	1455	1521	1520
92 (replicate)	53	53	50	50	100	100	500	500	1455	1455	1521	1521
93	53	53	50	50	100	100	500	500	1455	1455	1521	1518
97	53	52	50	50	100	100	500	500	1455	1455	1521	1521
98	53	53	50	50	100	100	500	500	1455	1455	1521	1521
99	53	53	50	50	100	100	500	500	1455	1455	1521	1521
100	53	53	50	50	100	100	500	500	1455	1455	1521	1520
101	53	53	50	50	100	100	500	500	1455	1455	1521	1519
102	53	53	50	50	100	100	500	500	1455	1455	1521	1519
103	53	53	50	50	100	100	500	500	1455	1455	1521	1518
104	53	53	50	50	100	100	500	500	1455	1454	1521	1519
105	53	53	50	50	100	100	500	500	1455	1455	1521	1521
110	53	53	50	50	100	100	500	500	1455	1455	1521	1521
111	53	53	50	50	100	100	500	500	1455	1454	1519	1512
111 (replicate)	53	52	50	50	100	100	500	500	1455	1455	1518	1519

114	53	53	50	50	100	100	500	500	1455	1454	1521	1521
116	53	52	50	50	100	100	500	500	1455	1455	1521	1521
117	53	53	50	50	100	100	500	500	1455	1455	1521	1520
118	53	53	50	50	100	100	500	500	1455	1455	1521	1521
119	53	53	50	50	100	100	500	500	1455	1455	1521	1521
120	53	53	50	50	100	100	500	500	1455	1455	1521	1521
9	53	53	50	50	100	100	500	500	1455	1455	1521	1519
10	53	53	50	50	100	100	500	500	1455	1455	1521	1521
11	53	53	50	50	100	100	500	500	1455	1455	1521	1521
12	53	53	50	50	100	100	500	500	1455	1455	1521	1520
22	53	53	50	50	100	100	500	500	1455	1455	1521	1521
23	53	53	50	50	100	100	500	500	1455	1455	1521	1521
24	53	53	50	50	100	100	500	500	1455	1455	1521	1517
34	53	53	50	50	100	100	500	500	1455	1455	1521	1521
35	53	53	50	50	100	100	500	499	1455	1454	1521	1518
44	53	53	50	50	100	100	500	500	1455	1455	1521	1520
45	53	53	50	50	100	100	500	500	1455	1455	1521	1521
46	53	53	50	50	100	100	500	500	1455	1455	1521	1521
47	53	53	50	50	100	100	500	500	1455	1455	1521	1521
56	53	53	50	50	100	100	500	500	1455	1455	1521	1521
57	53	53	50	50	100	100	500	500	1455	1455	1521	1519
58	53	53	50	50	100	100	500	500	1455	1455	1521	1521
59	53	53	50	50	100	100	500	500	1455	1454	1521	1518
76	53	53	50	50	100	100	500	500	1455	1455	1521	1518
77	53	53	50	50	100	100	500	500	1455	1455	1521	1520
78	53	53	50	50	100	100	500	500	1455	1455	1521	1520
79	53	53	50	50	100	100	500	500	1455	1455	1521	1521
94	53	53	50	50	100	100	500	500	1455	1454	1521	1520
95	53	53	50	50	100	100	500	500	1455	1455	1521	1519
96	53	53	50	50	100	100	500	500	1455	1455	1521	1519
106	53	52	50	50	100	100	500	500	1455	1455	1521	1519
107	53	52	50	50	100	100	500	500	1455	1455	1521	1520
121	53	52	50	50	100	100	500	500	1455	1455	1521	1521
LC 202/ EUL 153	53	53	50	50	100	100	500	500	1455	1455	1521	1519
LC 206/ EUL 158	53	53	50	50	100	100	500	500	1455	1455	1521	1520
LC 569/ EUL 154	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC 606/ EUL 155	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC 384/ EUL 156	53	53	50	50	100	100	500	500	1455	1455	1516	1514
LC 395/ EUL 159	53	53	50	50	100	100	500	500	1455	1455	1516	1514
LC6379-1/EUL 145	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC6376	53	53	50	50	100	100	500	500	1455	1455	1521	1520
LC6382	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC6391	53	53	50	50	100	100	500	500	1455	1455	1521	1520
LC6394	53	53	50	50	100	100	500	500	1455	1455	1521	1520

LC6397	53	53	50	50	100	100	500	500	1455	1455	1521	1519
LC6406	53	53	50	50	100	100	500	500	1455	1455	1521	1519
LC6407	53	53	50	50	100	100	500	500	1455	1455	1521	1520
LC6408	53	53	50	50	100	100	500	500	1455	1455	1521	1520
LC6411	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC6412	53	52	50	50	100	100	500	500	1455	1455	1521	1518
LC6413	53	53	50	50	100	100	500	500	1455	1455	1521	1520
LC6416	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC6418	53	53	50	50	100	100	500	500	1455	1455	1521	1519
LC6385	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC6388	53	53	50	50	100	100	500	500	1455	1455	1521	1520
LC6409	53	53	50	50	100	100	500	500	1455	1455	1521	1519
LC6410	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC0537/ EUL 132	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC0539/ EUL 133	53	52	50	50	100	100	500	500	1455	1454	1521	1519
LC0540/ EUL 134	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC0565	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC0583	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H034680033	53	52	50	46	100	95	500	480	1450	1400	1511	1455
H034680035 /EUL 165	53	52	50	50	100	100	500	500	1455	1455	1521	1520
H034690056 /EUL 166	53	52	50	50	100	100	500	500	1455	1455	1521	1519
H034800427	53	52	50	49	100	98	499	492	1454	1444	1519	1513
H034980467	53	53	50	50	100	100	500	500	1455	1454	1521	1520
Paris (complete genome)	53	NA	50	NA	100	NA	500	NA	1455	NA	1521	NA
H034800423	52	42	50	39	97	79	479	364	1394	1064	1456	1067
OLDA1 (NCTC12008)	53	53	50	50	100	100	500	500	1455	1454	1521	1519
EUL 109	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H064240448	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC0731	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC0732	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC0763	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC0782	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC0795	53	52	50	50	100	100	500	500	1455	1455	1521	1521
LC0801	53	53	50	50	100	100	500	500	1455	1455	1521	1520
LC5694	53	53	50	50	100	100	500	500	1455	1455	1504	1504
LC5722	53	53	50	50	100	100	500	500	1455	1455	1521	1519
LC5738	53	53	50	50	100	100	500	500	1455	1455	1504	1504
LC5755	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC6163	53	53	50	50	100	100	500	500	1455	1455	1521	1520
LC6267	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC6268	53	53	50	50	100	100	500	500	1455	1455	1521	1519
LC6228	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H041380048	53	53	50	50	100	100	500	500	1455	1455	1521	1521

H041640791	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H042960010	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H061140013	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H071880001	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H073060003	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H080820009	53	53	50	50	100	100	500	500	1455	1454	1521	1520
LC6058	53	53	50	50	100	100	500	500	1455	1455	1521	1520
LC6293	53	52	50	50	100	100	500	500	1455	1455	1521	1519
LC6788	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H062660463	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H073900557	53	53	50	50	100	100	500	500	1455	1455	1521	1520
LC1127	53	53	50	50	100	100	500	500	1455	1454	1521	1520
H084760449	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H085020185	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H090320386	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H044260061	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H093140322	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H093160422	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H092760433	53	52	50	50	100	100	500	500	1455	1454	1521	1518
H100940111	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H101760092	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H101820190	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H102020414	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H101980130	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H103820081	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H120240685	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H104320293	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H113180118	53	53	50	50	100	100	500	500	1455	1455	1504	1504
H113340664	53	53	50	50	100	100	500	500	1455	1455	1504	1504
H113280076	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H113660550	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H114740454	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H115040456	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H111580389	53	52	50	50	100	100	500	500	1455	1455	1521	1520
H113780240	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H083920177	53	53	50	50	100	100	500	500	1455	1455	1519	1517
H084140691	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H081180019	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H103260667	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC464	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC0512	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC0794	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC0798	53	53	50	50	100	100	500	500	1454	1454	1519	1519
LC0536/ EUL 131	53	53	50	50	100	100	500	500	1455	1455	1521	1520
LC230/ EUL 122	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC231/ EUL 123	53	53	50	50	100	100	500	500	1455	1455	1521	1519

LC0462/ EUL 124	53	53	50	50	100	100	500	500	1455	1455	1521	1518
LC0463/ EUL 125	53	53	50	50	100	100	500	500	1455	1455	1521	1520
Lorraine (complete genome)	53	NA	50	NA	100	NA	500	NA	1455	NA	1521	NA
H063920004 /EUL 169	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H064160534 /EULV0410	53	53	50	50	100	100	500	500	1455	1455	1521	1518
H064160538 /EUL 170	53	53	50	50	100	100	500	500	1455	1455	1521	1518
H034700617	53	53	50	50	100	100	500	498	1455	1449	1521	1510
H043580159	53	52	50	50	100	100	500	500	1455	1455	1521	1519
H043580160	53	52	50	50	100	100	500	500	1455	1455	1521	1520
H043660021	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H043680663	53	53	50	50	100	100	500	500	1455	1455	1521	1517
H043700021	53	52	50	50	100	100	500	500	1455	1455	1521	1518
H043790008	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H052920051	53	53	50	50	100	100	500	500	1455	1455	1521	1518
H053540106	53	52	50	50	100	100	500	500	1455	1455	1521	1520
H063660005	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H063660006	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H063760006	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H063660009	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H063680006	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H063680007	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H063740003	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H063740018	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H063780007	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H063780008	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H063860003	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H063960001	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC5759	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H070420013	53	53	50	50	100	100	500	500	1455	1455	1521	1518
LC5822	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H040260015	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H055140095	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H060780053	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H061120064	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H062840608	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H062940111	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H064320006	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H064280005	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H064380002	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H064380001	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H064560527	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H064660638	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H070160015	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H071120010	53	53	50	50	100	100	500	500	1455	1455	1521	1520

H071360036	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H072740002	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H073000045	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H073380007	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H073600182	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H073640185	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H074960018	53	52	50	50	100	100	500	500	1455	1455	1521	1519
H080780059	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H053840008	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H072520002	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H081340222	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H082520613	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H083120262	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H083620580	53	52	50	50	100	100	500	500	1455	1455	1521	1518
H083960064	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H084620118	53	53	50	50	100	100	500	499	1455	1454	1521	1518
H090140214	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H090440226	53	52	50	50	100	100	500	500	1455	1455	1521	1519
H040960441	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H041120007	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H093480403	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H094340202	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H095060125	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H100140151	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H100660110	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H100700025	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H103140121	53	52	50	50	100	100	500	500	1455	1455	1521	1517
H103620160	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H103660126	53	52	50	50	100	100	500	500	1455	1455	1521	1519
H103660121	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H104420240	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H110480273	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H112320437	53	53	50	50	100	100	500	500	1455	1455	1521	1518
H112080616	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H112380374	53	53	50	50	100	100	500	500	1455	1455	1521	1518
H120160499	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H120200371	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H105140391	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H121040204	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H121420445	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H102240357	53	52	50	50	100	100	500	500	1455	1455	1521	1519
H122500497	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H122820408	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H123620597	53	52	50	50	100	100	500	500	1455	1455	1521	1519
H123840629	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H123940534	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H124920387	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H131340777	53	53	50	50	100	100	500	500	1455	1455	1521	1519

H131480353	53	53	50	50	100	100	500	500	1455	1455	1521	1518
H131480354	53	52	50	50	100	100	500	500	1455	1455	1521	1518
H131840211	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H131460248	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H132140863	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H053640534 /EUL 168	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H064180002	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H064180019	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H043540106	53	53	50	50	100	100	500	500	1455	1455	1521	1518
H044120014	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H052780022	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H054280040	53	53	50	50	100	100	500	500	1455	1455	1521	1515
H063680003	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H063840008	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H073660582	53	53	50	50	100	100	500	500	1455	1455	1521	1521
LC5804	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H063760005	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H064240003	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H065040012	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H070140635	53	53	50	50	100	100	500	500	1455	1455	1521	1517
H073020039	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H073320399	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H073440003	53	53	50	50	100	100	500	500	1455	1455	1521	1520
LC6009	53	53	50	50	100	100	500	500	1455	1454	1521	1519
H083140015	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H093400182	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H094760070	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H094800237	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H110480715	53	53	50	50	100	100	500	500	1455	1455	1521	1519
H112840293	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H114100406	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H120240362	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H104640262	53	52	50	50	100	100	500	500	1455	1455	1521	1519
H123140428	53	53	50	50	100	100	500	500	1455	1455	1521	1520
H123460520	53	53	50	50	100	100	500	500	1455	1455	1521	1521
H124360642	53	53	50	50	100	100	500	500	1455	1455	1521	1521
Pontiac-1	53	53	50	50	100	100	499	495	1444	1437	1507	1501

Appendix Table 18. 61 untypable genes in the six extended MLST schemes and the number of affected isolates in the typing panel.

Gene	Scheme(s)	Number of affected isolates in typing panel
lpg0328	cgMLST (1521), rMLST	10
lpg1614	cgMLST (1521)	9

lpg0703	cgMLST (1521)	5
lpg1615	cgMLST (1521)	5
lpg0568	cgMLST (1455)	4
lpg2871	cgMLST (1521)	4
lpg0639	cgMLST (1521)	3
lpg0857	cgMLST (1521)	3
lpg1125	cgMLST (1521)	3
lpg1612	cgMLST (1455), cgMLST (1521)	3
lpg1872	cgMLST (1521)	3
lpg2121	cgMLST (1521)	3
lpg2361	cgMLST (1521)	3
lpg2422	cgMLST (1521)	3
lpg2452	cgMLST (1521)	3
lpg2888	cgMLST (1521)	3
lpg0179	cgMLST (1521)	2
lpg0735	cgMLST (1521)	2
lpg0744	cgMLST (1521)	2
lpg0903	cgMLST (1521)	2
lpg1099	cgMLST (1521)	2
lpg1100	cgMLST (1521)	2
lpg1169	cgMLST (1521)	2
lpg1371	cgMLST (1521)	2
lpg1664	cgMLST (1521)	2
lpg0049	cgMLST (1521)	1
lpg0073	cgMLST (1521)	1
lpg0121	cgMLST (1521)	1
lpg0286	cgMLST (1521)	1
lpg0316	cgMLST (1521)	1
lpg0326	cgMLST (1521)	1
lpg0329	cgMLST (50), cgMLST (100), cgMLST (500), cgMLST (1455), cgMLST (1521), rMLST	1
lpg0342	cgMLST (500), cgMLST (1455), cgMLST (1521), rMLST	1
lpg0409	cgMLST (50), cgMLST (100), cgMLST (500), cgMLST (1455), cgMLST (1521)	1
lpg0549	cgMLST (1521)	1
lpg0707	cgMLST (1521)	1
lpg0952	cgMLST (1521)	1
lpg1181	cgMLST (1521)	1
lpg1187	cgMLST (1521)	1
lpg1199	cgMLST (1521)	1
lpg1209	cgMLST (1521)	1
lpg1335	cgMLST (1521)	1
lpg1564	cgMLST (1455)	1
lpg1567	cgMLST (1455)	1
lpg1581	cgMLST (1521)	1
lpg1665	cgMLST (1521)	1

lpg1751	cgMLST (1455), cgMLST (1521)	1
lpg1868	cgMLST (1521)	1
lpg2016	cgMLST (1521)	1
lpg2044	cgMLST (1455), cgMLST (1521)	1
lpg2146	cgMLST (1521)	1
lpg2196	cgMLST (1521)	1
lpg2208	cgMLST (500), cgMLST (1455), cgMLST (1521)	1
lpg2227	cgMLST (1521)	1
lpg2395	cgMLST (1521)	1
lpg2446	cgMLST (1521)	1
lpg2462	cgMLST (1521)	1
lpg2506	cgMLST (500), cgMLST (1455), cgMLST (1521)	1
lpg2639	cgMLST (1521)	1
lpg2856	cgMLST (1521)	1
lpg2984	cgMLST (1521)	1

Appendix Table 19. The mean and range of mapping coverage, number of contigs and N50 values of isolates that produce complete or incomplete profiles in the extended MLST schemes.

	Mean (and range) of mapping coverage/depth	**Mean (and range) of contigs**	**Mean (and range) of N50 values (bp)**
Isolates with a full profile in all extended MLST schemes	125.4x (77.1x-162.3x)	35.2 (17-72)	246,381 (101,986-657,238)
Isolates with an incomplete profile in 1 or more extended MLST schemes	121.8x (71.4x-164.2x)	35.0 (15-65)	264,222 (86,373-726,453)
Significant difference via unpaired t-test?	No significant difference	No significant difference	No significant difference

Appendix Table 20. The number of accessory genes scored as present, absent or untypable using the gene presence/absence typing method.

EUL/isolate number	**No. genes present**	**No. genes absent**	**No. untypable genes**
1	97	103	0
2	50	148	2
3	97	103	0
4	116	82	2

6	85	112	3
7	117	82	1
8	87	110	3
13	104	93	3
14	104	93	3
16	106	93	1
17	92	108	0
18	37	163	0
19	66	131	3
20	59	139	2
25	55	145	0
26	59	141	0
27	102	96	2
27 (replicate)	102	96	2
28	100	97	3
29	99	101	0
30	62	136	2
31	101	99	0
32	90	109	1
33	62	137	1
33 (replicate)	62	136	2
36	87	113	0
37	92	108	0
38	97	103	0
39	61	138	1
40	30	168	2
41	88	110	2
42	97	103	0
43	96	103	1
48	108	92	0
49	96	103	1
50	86	112	2
51	41	156	3
52	54	143	3
53	96	103	1
54	104	96	0
55	97	103	0
60	96	102	2
63	80	118	2
66	80	118	2
67	96	103	1
68	84	115	1
69	70	130	0
69 (replicate)	69	129	2

70	102	98	0
71	115	85	0
72	47	151	2
73	69	129	2
74	61	137	2
75	87	112	1
75 (replicate)	87	112	1
81	66	132	2
82	101	99	0
83	62	136	2
84	96	104	0
85	101	99	0
86	108	92	0
87	73	124	3
88	101	99	0
91	28	169	3
92	65	132	3
92 (replicate)	66	133	1
93	101	99	0
97	52	144	4
98	52	143	5
99	52	146	2
100	81	119	0
101	85	112	3
102	75	123	2
103	107	92	1
104	107	93	0
105	87	113	0
110	106	94	0
111	59	141	0
111 (replicate)	59	140	1
114	91	109	0
116	58	141	1
117	92	107	1
118	68	132	0
119	97	103	0
120	87	112	1
9	97	103	0
10	97	103	0
11	88	110	2
12	88	110	2
22	53	146	1
23	53	145	2
24	53	146	1

34	62	136	2
35	62	135	3
44	95	105	0
45	97	103	0
46	97	103	0
47	30	168	2
56	108	92	0
57	103	95	2
58	97	103	0
59	42	158	0
76	115	85	0
77	115	85	0
78	69	129	2
79	69	129	2
94	101	99	0
95	101	99	0
96	66	132	2
106	53	144	3
107	52	143	5
121	87	112	1
LC 202/EUL 153	66	132	2
LC 206/EUL 158	66	132	2
LC 569/EUL 154	69	131	0
LC 606/EUL 155	69	131	0
LC 384/EUL 156	66	132	2
LC 395/EUL 159	66	132	2
LC6379-1/EUL 145	74	126	0
LC6376	73	125	2
LC6382	74	126	0
LC6391	74	126	0
LC6394	73	126	1
LC6397	74	126	0
LC6406	74	126	0
LC6407	74	126	0
LC6408	74	126	0
LC6411	74	126	0
LC6412	73	126	1
LC6413	73	126	1
LC6416	74	126	0
LC6418	74	126	0
LC6385	74	126	0
LC6388	74	126	0
LC6409	74	126	0
LC6410	73	125	2

LC0537/EUL 132	69	129	2
LC0539/EUL 133	69	129	2
LC0540/EUL 134	69	129	2
LC0565	69	129	2
LC0583	69	129	2
H034680033	71	124	5
H034680035/EUL 165	73	125	2
H034690056/EUL 166	73	126	1
H034800427	73	125	2
H034980467	72	125	3
Paris	92	107	1
H034800423	95	102	3
OLDA1 (NCTC12008)	97	101	2
EUL 109	107	93	0
H064240448	69	129	2
LC0731	62	136	2
LC0732	62	136	2
LC0763	62	136	2
LC0782	62	136	2
LC0795	62	136	2
LC0801	62	136	2
LC5694	67	132	1
LC5722	55	143	2
LC5738	67	133	0
LC5755	69	129	2
LC6163	70	130	0
LC6267	69	129	2
LC6268	69	129	2
LC6228	73	125	2
H041380048	69	129	2
H041640791	69	128	3
H042960010	69	129	2
H061140013	69	129	2
H071880001	69	129	2
H073060003	69	129	2
H080820009	84	114	2
LC6058	69	129	2
LC6293	69	129	2
LC6788	69	129	2
H062660463	69	129	2
H073900557	79	119	2
LC1127	62	136	2
H084760449	73	125	2
H085020185	73	125	2

H090320386	69	129	2
H044260061	72	125	3
H093140322	73	125	2
H093160422	73	125	2
H092760433	69	129	2
H100940111	69	129	2
H101760092	69	129	2
H101820190	69	129	2
H102020414	72	126	2
H101980130	69	129	2
H103820081	68	129	3
H120240685	69	129	2
H104320293	69	130	1
H113180118	67	132	1
H113340664	67	132	1
H113280076	78	120	2
H113660550	70	128	2
H114740454	69	129	2
H115040456	69	129	2
H111580389	73	125	2
H113780240	69	129	2
H083920177	73	125	2
H084140691	69	129	2
H081180019	69	128	3
H103260667	78	122	0
LC464	69	129	2
LC0512	69	130	1
LC0794	69	129	2
LC0798	69	129	2
LC0536/EUL 131	69	125	6
LC230/EUL 122	89	110	1
LC231/EUL 123	89	111	0
LC0462/EUL 124	87	112	1
LC0463/EUL 125	87	112	1
Lorraine	87	113	0
EUL 169	101	97	2
H064160534	102	98	0
H064160538/EUL 170	101	98	1
H034700617	100	97	3
H043580159	102	98	0
H043580160	102	98	0
H043660021	102	98	0
H043680663	102	98	0
H043700021	111	89	0

H043790008	102	97	1
H052920051	101	99	0
H053540106	102	98	0
H063660005	101	97	2
H063660006	101	97	2
H063760006	100	97	3
H063660009	102	97	1
H063680006	107	89	4
H063680007	101	96	3
H063740003	101	97	2
H063740018	102	97	1
H063780007	102	98	0
H063780008	102	98	0
H063860003	98	96	6
H063960001	100	96	4
LC5759	102	98	0
H070420013	102	98	0
LC5822	102	98	0
H040260015	100	97	3
H055140095	101	97	2
H060780053	102	97	1
H061120064	101	97	2
H062840608	100	98	2
H062940111	102	98	0
H064320006	99	98	3
H064280005	102	97	1
H064380002	102	98	0
H064380001	102	97	1
H064560527	101	97	2
H064660638	102	98	0
H070160015	102	98	0
H071120010	101	97	2
H071360036	101	97	2
H072740002	102	97	1
H073000045	101	97	2
H073380007	102	98	0
H073600182	102	98	0
H073640185	102	97	1
H074960018	101	97	2
H080780059	101	97	2
H053840008	102	98	0
H072520002	102	97	1
H081340222	101	97	2
H082520613	101	97	2

H083120262	102	98	0
H083620580	102	97	1
H083960064	102	97	1
H084620118	102	98	0
H090140214	102	98	0
H090440226	102	97	1
H040960441	101	97	2
H041120007	102	98	0
H093480403	102	98	0
H094340202	102	98	0
H095060125	102	98	0
H100140151	102	98	0
H100660110	102	98	0
H100700025	102	98	0
H103140121	102	98	0
H103620160	101	97	2
H103660126	102	97	1
H103660121	102	98	0
H104420240	102	98	0
H110480273	102	98	0
H112320437	102	98	0
H112080616	102	98	0
H112380374	102	98	0
H120160499	102	98	0
H120200371	100	97	3
H105140391	102	98	0
H121040204	102	98	0
H121420445	102	98	0
H102240357	102	98	0
H122500497	101	98	1
H122820408	101	99	0
H123620597	88	112	0
H123840629	102	98	0
H123940534	102	98	0
H124920387	102	98	0
H131340777	102	98	0
H131480353	102	98	0
H131480354	103	97	0
H131840211	102	98	0
H131460248	102	98	0
H132140863	100	98	2
H053640534/EUL 168	102	98	0
H064180002	77	122	1
H064180019	77	122	1

H043540106	104	96	0
H044120014	112	88	0
H052780022	114	86	0
H054280040	115	85	0
H063680003	124	75	1
H063840008	115	85	0
H073660582	124	76	0
LC5804	126	74	0
H063760005	103	97	0
H064240003	103	95	2
H065040012	104	96	0
H070140635	115	85	0
H073020039	115	85	0
H073320399	114	85	1
H073440003	113	86	1
LC6009	104	96	0
H083140015	104	96	0
H093400182	113	87	0
H094760070	124	76	0
H094800237	124	76	0
H110480715	114	86	0
H112840293	124	76	0
H114100406	104	96	0
H120240362	124	76	0
H104640262	82	118	0
H123140428	114	86	0
H123460520	121	79	0
H124360642	124	76	0
Pontiac-1	113	84	3

Appendix Table 21. Genes scored as untypable in one or more typing panel isolates using the gene presence/absence typing method.

Gene number	Number of affected isolates in typing panel
130	38
87	36
187	10
91	4
81	4
47	4
39	4

133	3
31	3
49	2
46	2
45	2
163	2
171	2
108	2
155	1
172	1
18	1
28	1
123	1
128	1

Appendix Table 22. The number of differences between isolates belonging to an additional 14 epidemiologically "related" sets, as analysed by each of the WGS-based methods. Sets in which one or more isolates could not be fully typed by a particular scheme are marked with an asterisk.

Isolate names	**Mean (and range) of pairwise differences**								
	SNP-based	rMLST (53)	cgMLST (50)	cgMLST (100)	cgMLST (500)	cgMLST (1455)	cgMLST (1521)	Gene pres./ abs.	Kmer-based
LC0731/ LC0732/ LC0763/ LC0782/ LC0795/ LC0801	0.33 (0-1)	0 (0-0)*	0 (0-0)	0 (0-0)	0 (0-0)	0.33 (0-1)	0 (0-0)*	0 (0-0)*	0.064 (0.063-0.065)
H041380048/ H041640791	4	0	0	0	1	1	1	0*	0.066
H093140322/ H093160422	4	1	0	0	0	1	3	0*	0.060
H113180118/ H113340664	1	0	0	0	0	0	1*	0*	0.064
H083920177/ H084140691	5	0	0	0	0	4*	6*	4*	0.075
LC0794/ LC0798	1	0	0	0	0	1*	1*	0*	0.065
EUL 122/ EUL 123	0	0	0	0	0	0	0*	0*	0.062
EUL 124/ EUL 125	2	0	0	0	0	1*	0*	0*	0.064
EUL 169/ H064160534/ EUL 170	1.33 (1-2)	0 (0-0)	0	0	0	0	0.67 (0-1)*	0 (0-0)*	0.064 (0.062-0.064)
H063660005/ H063660006/ H063760006	3.33 (2-5)	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-0)	1.33 (1-2)	0.67 (0-1)*	0 (0-0)*	0.064 (0.061-0.065)
H063680006/ H063680007	11	0	0	0	1	5	7*	10*	0.076

H063780007/ H063780008	2	0	0	0	0	2	1	0	0.064
H131340777/ H131480353/ H131480354/ H131840211	2.17 (0-4)	0 (0-0)*	0 (0-0)	0.5 (0-1)	1.5 (0-3)	2.17 (0-4)	1 (0-2)*	0.5 (0-1)	0.073 (0.061-0.080)
H064180002/ H064180019	1	0	0	0	0	0	1*	0*	0.065

Appendix Table 23. The indices of discrimination (*D*) for 53 ribosomal genes, calculated using 79 epidemiologically "unrelated" isolates from the typing panel.

Gene name	***D* value**
lpg0342/rpsN	0.728
lpg0343/rpsH	0.846
lpg0344/rplF	0.853
lpg0345/rplR	0.586
lpg0346/rpsE	0.771
lpg0347/rpmD	0.611
lpg0348/rplO	0.844
lpg0350/rpmJ	0.025
lpg0351/rpsM	0.677
lpg0352/rpsK	0.758
lpg0353/rpsD	0.873
lpg0355/rplQ	0.848
lpg0395/rplS	0.801
lpg0399/rpsP	0.542
lpg0478/rpmG	0.607
lpg0479/rpmB	0.802
lpg0650/rpmE	0.520
lpg1391/rpmF	0.625
lpg1421/rpsA	0.866
lpg1589/rplI	0.836
lpg1591/rpsR	0.730
lpg1592/rpsF	0.808
lpg1714/rpsB	0.858
lpg2358/rpsU	0.249
lpg2636/rpsT	0.561
lpg2650/rpmA	0.075
lpg2651/rplU	0.678
lpg2706/rpsI	0.802
lpg2707/rplM	0.677
lpg2712/rplT	0.525
lpg2769/rpsO	0.824

lpg3005/rpmH	0.516
Unannotated/*rpmI*	0.075

Appendix Table 24. The indices of discrimination (*D*) for 200 accessory genes, calculated using 79 epidemiologically "unrelated" isolates from the typing panel.

Gene number	***D* value**
1	0.500
2	0.142
3	0.418
4	0.504
5	0.356
6	0.483
7	0.483
8	0.493
9	0.506
10	0.488
11	0.483
12	0.477
13	0.477
14	0.477
15	0.477
16	0.342
17	0.483
18	0.504
19	0.504
20	0.488
21	0.500
22	0.342
23	0.327
24	0.327
25	0.327
26	0.327
27	0.327
28	0.327
29	0.327
30	0.327
31	0.471
32	0.356
33	0.356
34	0.184
35	0.506
36	0.477

37	0.506
38	0.500
39	0.383
40	0.327
41	0.327
42	0.395
43	0.502
44	0.327
45	0.327
46	0.327
47	0.327
48	0.327
49	0.327
50	0.295
51	0.295
52	0.295
53	0.327
54	0.243
55	0.502
56	0.500
57	0.500
58	0.312
59	0.295
60	0.383
61	0.506
62	0.506
63	0.506
64	0.506
65	0.463
66	0.493
67	0.488
68	0.383
69	0.356
70	0.356
71	0.356
72	0.493
73	0.383
74	0.483
75	0.488
76	0.493
77	0.295
78	0.497
79	0.483
80	0.506
81	0.488

82	0.497
83	0.204
84	0.477
85	0.477
86	0.477
87	0.500
88	0.261
89	0.261
90	0.261
91	0.463
92	0.502
93	0.502
94	0.356
95	0.483
96	0.456
97	0.500
98	0.497
99	0.497
100	0.497
101	0.497
102	0.497
103	0.497
104	0.497
105	0.497
106	0.477
107	0.025
108	0.312
109	0.477
110	0.504
111	0.504
112	0.327
113	0.493
114	0.477
115	0.502
116	0.477
117	0.438
118	0.074
119	0.463
120	0.506
121	0.164
122	0.506
123	0.312
124	0.493
125	0.506
126	0.504

127	0.243
128	0.471
129	0.506
130	0.493
131	0.504
132	0.504
133	0.502
134	0.504
135	0.506
136	0.370
137	0.463
138	0.456
139	0.502
140	0.278
141	0.428
142	0.447
143	0.447
144	0.395
145	0.418
146	0.418
147	0.447
148	0.395
149	0.261
150	0.261
151	0.261
152	0.261
153	0.356
154	0.383
155	0.488
156	0.488
157	0.477
158	0.395
159	0.502
160	0.493
161	0.074
162	0.224
163	0.164
164	0.261
165	0.327
166	0.204
167	0.477
168	0.164
169	0.383
170	0.327
171	0.295

172	0.025
173	0.506
174	0.000
175	0.050
176	0.025
177	0.502
178	0.504
179	0.025
180	0.050
181	0.500
182	0.164
183	0.356
184	0.502
185	0.456
186	0.383
187	0.483
188	0.074
189	0.502
190	0.164
191	0.504
192	0.295
193	0.483
194	0.025
195	0.395
196	0.471
197	0.502
198	0.395
199	0.488
200	0.164

Appendix Table 25. The indices of discrimination (*D*) for 1455 core genes, calculated using 79 epidemiologically "unrelated" isolates from the typing panel.

Gene name	***D* value**		**Gene name**	***D* value**		**Gene name**	***D* value**
lpg0001	0.862		*lpg0930*	0.733		*lpg2013*	0.873
lpg0002	0.866		*lpg0932*	0.850		*lpg2014*	0.871
lpg0004	0.866		*lpg0933*	0.837		*lpg2015*	0.869
lpg0005	0.881		*lpg0934*	0.846		*lpg2017*	0.859
lpg0009	0.278		*lpg0935*	0.769		*lpg2018*	0.868
lpg0010	0.880		*lpg0936*	0.829		*lpg2020*	0.858
lpg0011	0.879		*lpg0937*	0.887		*lpg2021*	0.881
lpg0014	0.880		*lpg0938*	0.838		*lpg2023*	0.879

lpg0018	0.878		*lpg0940*	0.874		*lpg2024*	0.879
lpg0021	0.835		*lpg0941*	0.864		*lpg2025*	0.868
lpg0022	0.901		*lpg0942*	0.849		*lpg2027*	0.880
lpg0023	0.852		*lpg0943*	0.840		*lpg2028*	0.870
lpg0024	0.793		*lpg0946*	0.809		*lpg2029*	0.784
lpg0025	0.883		*lpg0949*	0.840		*lpg2031*	0.873
lpg0027	0.894		*lpg0950*	0.849		*lpg2032*	0.870
lpg0028	0.896		*lpg0951*	0.861		*lpg2033*	0.873
lpg0032	0.885		*lpg0953*	0.845		*lpg2034*	0.880
lpg0033	0.862		*lpg0954*	0.870		*lpg2036*	0.879
lpg0035	0.731		*lpg0955*	0.861		*lpg2037*	0.874
lpg0037	0.870		*lpg0956*	0.846		*lpg2038*	0.805
lpg0040	0.866		*lpg0957*	0.872		*lpg2039*	0.896
lpg0043	0.862		*lpg0958*	0.864		*lpg2040*	0.889
lpg0047	0.875		*lpg0960*	0.829		*lpg2041*	0.856
lpg0048	0.882		*lpg0961*	0.847		*lpg2042*	0.878
lpg0052	0.877		*lpg0962*	0.858		*lpg2043*	0.806
lpg0059	0.867		*lpg0963*	0.853		*lpg2044*	0.857
lpg0075	0.843		*lpg0966*	0.857		*lpg2045*	0.875
lpg0076	0.857		*lpg0970*	0.848		*lpg2046*	0.872
lpg0078	0.874		*lpg0971*	0.858		*lpg2047*	0.871
lpg0079	0.873		*lpg1117*	0.862		*lpg2048*	0.873
lpg0083	0.864		*lpg1119*	0.874		*lpg2049*	0.782
lpg0084	0.873		*lpg1121*	0.861		*lpg2051*	0.876
lpg0085	0.851		*lpg1122*	0.864		*lpg2052*	0.882
lpg0089	0.841		*lpg1131*	0.875		*lpg2053*	0.884
lpg0091	0.851		*lpg1135*	0.870		*lpg2175*	0.863
lpg0094	0.856		*lpg1136*	0.868		*lpg2176*	0.871
lpg0095	0.860		*lpg1137*	0.857		*lpg2178*	0.878
lpg0098	0.853		*lpg1138*	0.874		*lpg2186*	0.903
lpg0099	0.851		*lpg1139*	0.853		*lpg2187*	0.849
lpg0100	0.859		*lpg1140*	0.853		*lpg2189*	0.855
lpg0101	0.858		*lpg1141*	0.857		*lpg2191*	0.859
lpg0102	0.876		*lpg1143*	0.849		*lpg2193*	0.842
lpg0103	0.854		*lpg1144*	0.871		*lpg2194*	0.861
lpg0104	0.840		*lpg1146*	0.869		*lpg2200*	0.848
lpg0105	0.816		*lpg1147*	0.856		*lpg2201*	0.870
lpg0106	0.866		*lpg1148*	0.865		*lpg2202*	0.848
lpg0110	0.878		*lpg1154*	0.863		*lpg2203*	0.867
lpg0111	0.868		*lpg1155*	0.858		*lpg2204*	0.864
lpg0115	0.845		*lpg1157*	0.855		*lpg2206*	0.864
lpg0116	0.885		*lpg1159*	0.859		*lpg2207*	0.863
lpg0117	0.848		*lpg1161*	0.819		*lpg2208*	0.824
lpg0118	0.778		*lpg1162*	0.852		*lpg2209*	0.854

lpg0119	0.854		*lpg1164*	0.864		*lpg2210*	0.864
lpg0120	0.856		*lpg1165*	0.853		*lpg2211*	0.851
lpg0122	0.882		*lpg1166*	0.858		*lpg2212*	0.862
lpg0125	0.856		*lpg1167*	0.825		*lpg2213*	0.806
lpg0127	0.865		*lpg1171*	0.851		*lpg2214*	0.869
lpg0128	0.864		*lpg1172*	0.875		*lpg2220*	0.871
lpg0129	0.865		*lpg1174*	0.874		*lpg2222*	0.842
lpg0130	0.870		*lpg1176*	0.880		*lpg2225*	0.865
lpg0131	0.858		*lpg1178*	0.885		*lpg2228*	0.888
lpg0136	0.872		*lpg1179*	0.871		*lpg2229*	0.885
lpg0137	0.869		*lpg1180*	0.867		*lpg2231*	0.889
lpg0138	0.863		*lpg1186*	0.865		*lpg2232*	0.853
lpg0140	0.871		*lpg1188*	0.877		*lpg2233*	0.757
lpg0153	0.868		*lpg1189*	0.867		*lpg2234*	0.886
lpg0165	0.854		*lpg1190*	0.883		*lpg2235*	0.886
lpg0175	0.796		*lpg1191*	0.824		*lpg2238*	0.867
lpg0183	0.875		*lpg1195*	0.866		*lpg2240*	0.891
lpg0188	0.875		*lpg1196*	0.868		*lpg2242*	0.904
lpg0189	0.861		*lpg1197*	0.873		*lpg2243*	0.840
lpg0194	0.872		*lpg1198*	0.869		*lpg2245*	0.908
lpg0197	0.833		*lpg1202*	0.856		*lpg2246*	0.833
lpg0206	0.858		*lpg1203*	0.845		*lpg2247*	0.863
lpg0209	0.869		*lpg1205*	0.631		*lpg2248*	0.873
lpg0212	0.878		*lpg1206*	0.883		*lpg2249*	0.850
lpg0213	0.870		*lpg1207*	0.852		*lpg2250*	0.870
lpg0217	0.867		*lpg1208*	0.848		*lpg2255*	0.741
lpg0218	0.857		*lpg1212*	0.853		*lpg2256*	0.870
lpg0227	0.864		*lpg1214*	0.872		*lpg2258*	0.833
lpg0229	0.866		*lpg1215*	0.865		*lpg2259*	0.826
lpg0232	0.777		*lpg1216*	0.839		*lpg2260*	0.852
lpg0238	0.868		*lpg1217*	0.843		*lpg2261*	0.866
lpg0239	0.866		*lpg1218*	0.838		*lpg2262*	0.895
lpg0241	0.871		*lpg1219*	0.853		*lpg2263*	0.881
lpg0243	0.860		*lpg1220*	0.854		*lpg2264*	0.728
lpg0244	0.870		*lpg1221*	0.845		*lpg2266*	0.884
lpg0245	0.882		*lpg1225*	0.858		*lpg2267*	0.889
lpg0248	0.861		*lpg1226*	0.864		*lpg2271*	0.858
lpg0252	0.856		*lpg1276*	0.858		*lpg2272*	0.850
lpg0256	0.862		*lpg1277*	0.908		*lpg2273*	0.866
lpg0257	0.867		*lpg1278*	0.844		*lpg2274*	0.868
lpg0260	0.854		*lpg1279*	0.848		*lpg2275*	0.860
lpg0264	0.858		*lpg1280*	0.859		*lpg2276*	0.856
lpg0267	0.860		*lpg1281*	0.775		*lpg2277*	0.850
lpg0268	0.852		*lpg1282*	0.838		*lpg2278*	0.861

lpg0269	0.864		*lpg1283*	0.862		*lpg2279*	0.856
lpg0271	0.844		*lpg1284*	0.864		*lpg2280*	0.870
lpg0276	0.874		*lpg1285*	0.858		*lpg2281*	0.857
lpg0282	0.893		*lpg1286*	0.851		*lpg2282*	0.853
lpg0287	0.883		*lpg1287*	0.852		*lpg2285*	0.853
lpg0288	0.893		*lpg1288*	0.771		*lpg2295*	0.855
lpg0289	0.896		*lpg1291*	0.853		*lpg2297*	0.868
lpg0290	0.893		*lpg1292*	0.848		*lpg2298*	0.857
lpg0291	0.876		*lpg1293*	0.818		*lpg2299*	0.872
lpg0293	0.902		*lpg1294*	0.858		*lpg2300*	0.854
lpg0294	0.896		*lpg1296*	0.858		*lpg2302*	0.834
lpg0295	0.881		*lpg1297*	0.837		*lpg2303*	0.820
lpg0296	0.889		*lpg1298*	0.765		*lpg2304*	0.883
lpg0298	0.877		*lpg1300*	0.852		*lpg2306*	0.848
lpg0299	0.885		*lpg1301*	0.857		*lpg2307*	0.828
lpg0301	0.876		*lpg1302*	0.852		*lpg2310*	0.889
lpg0317	0.857		*lpg1303*	0.843		*lpg2312*	0.763
lpg0318	0.850		*lpg1304*	0.859		*lpg2313*	0.855
lpg0319	0.876		*lpg1305*	0.853		*lpg2314*	0.850
lpg0320	0.704		*lpg1306*	0.858		*lpg2315*	0.604
lpg0321	0.716		*lpg1307*	0.858		*lpg2316*	0.858
lpg0322	0.900		*lpg1319*	0.818		*lpg2317*	0.867
lpg0323	0.879		*lpg1320*	0.885		*lpg2318*	0.861
lpg0324	0.294		*lpg1323*	0.849		*lpg2319*	0.858
lpg0325	0.688		*lpg1324*	0.861		*lpg2320*	0.863
lpg0329	0.851		*lpg1331*	0.864		*lpg2321*	0.883
lpg0330	0.855		*lpg1332*	0.833		*lpg2322*	0.870
lpg0331	0.566		*lpg1333*	0.873		*lpg2323*	0.870
lpg0332	0.776		*lpg1334*	0.849		*lpg2325*	0.871
lpg0335	0.767		*lpg1336*	0.847		*lpg2327*	0.877
lpg0336	0.844		*lpg1337*	0.848		*lpg2328*	0.834
lpg0337	0.227		*lpg1338*	0.853		*lpg2331*	0.862
lpg0338	0.288		*lpg1339*	0.837		*lpg2333*	0.870
lpg0339	0.830		*lpg1340*	0.866		*lpg2334*	0.796
lpg0340	0.844		*lpg1341*	0.841		*lpg2335*	0.866
lpg0341	0.780		*lpg1342*	0.857		*lpg2336*	0.867
lpg0342	0.724		*lpg1344*	0.847		*lpg2337*	0.854
lpg0343	0.846		*lpg1346*	0.845		*lpg2338*	0.846
lpg0346	0.771		*lpg1347*	0.674		*lpg2339*	0.873
lpg0347	0.611		*lpg1348*	0.860		*lpg2340*	0.874
lpg0348	0.844		*lpg1349*	0.858		*lpg2343*	0.868
lpg0349	0.865		*lpg1350*	0.853		*lpg2345*	0.867
lpg0352	0.758		*lpg1351*	0.852		*lpg2346*	0.866
lpg0353	0.873		*lpg1352*	0.897		*lpg2347*	0.863

lpg0354	0.887		*lpg1358*	0.865		*lpg2348*	0.863
lpg0355	0.848		*lpg1359*	0.863		*lpg2349*	0.850
lpg0356	0.900		*lpg1360*	0.835		*lpg2350*	0.858
lpg0357	0.911		*lpg1363*	0.866		*lpg2352*	0.868
lpg0359	0.873		*lpg1364*	0.849		*lpg2353*	0.765
lpg0361	0.916		*lpg1365*	0.752		*lpg2354*	0.858
lpg0362	0.905		*lpg1366*	0.857		*lpg2355*	0.859
lpg0363	0.850		*lpg1367*	0.859		*lpg2356*	0.874
lpg0365	0.876		*lpg1369*	0.858		*lpg2358*	0.249
lpg0366	0.860		*lpg1370*	0.659		*lpg2359*	0.849
lpg0369	0.877		*lpg1372*	0.854		*lpg2386*	0.857
lpg0370	0.822		*lpg1373*	0.854		*lpg2387*	0.873
lpg0371	0.854		*lpg1374*	0.851		*lpg2388*	0.886
lpg0372	0.851		*lpg1375*	0.856		*lpg2389*	0.888
lpg0374	0.823		*lpg1376*	0.828		*lpg2391*	0.891
lpg0376	0.884		*lpg1377*	0.681		*lpg2393*	0.850
lpg0377	0.857		*lpg1385*	0.843		*lpg2396*	0.882
lpg0380	0.806		*lpg1386*	0.846		*lpg2401*	0.883
lpg0382	0.858		*lpg1388*	0.848		*lpg2404*	0.877
lpg0383	0.851		*lpg1391*	0.625		*lpg2405*	0.697
lpg0384	0.867		*lpg1392*	0.845		*lpg2411*	0.834
lpg0385	0.847		*lpg1394*	0.862		*lpg2413*	0.844
lpg0386	0.855		*lpg1395*	0.830		*lpg2414*	0.761
lpg0387	0.820		*lpg1396*	0.610		*lpg2433*	0.878
lpg0388	0.863		*lpg1397*	0.804		*lpg2434*	0.873
lpg0391	0.855		*lpg1398*	0.865		*lpg2435*	0.866
lpg0392	0.868		*lpg1399*	0.847		*lpg2436*	0.876
lpg0393	0.860		*lpg1400*	0.771		*lpg2438*	0.876
lpg0394	0.867		*lpg1401*	0.735		*lpg2439*	0.877
lpg0395	0.801		*lpg1402*	0.868		*lpg2440*	0.888
lpg0396	0.860		*lpg1403*	0.876		*lpg2442*	0.863
lpg0399	0.542		*lpg1404*	0.870		*lpg2443*	0.867
lpg0400	0.863		*lpg1405*	0.863		*lpg2445*	0.864
lpg0404	0.862		*lpg1406*	0.854		*lpg2453*	0.844
lpg0405	0.850		*lpg1408*	0.858		*lpg2454*	0.851
lpg0406	0.755		*lpg1409*	0.863		*lpg2457*	0.714
lpg0407	0.841		*lpg1410*	0.858		*lpg2459*	0.863
lpg0408	0.845		*lpg1411*	0.850		*lpg2460*	0.868
lpg0409	0.854		*lpg1414*	0.862		*lpg2461*	0.869
lpg0410	0.841		*lpg1415*	0.737		*lpg2463*	0.811
lpg0411	0.857		*lpg1416*	0.811		*lpg2467*	0.884
lpg0413	0.849		*lpg1417*	0.852		*lpg2468*	0.863
lpg0414	0.849		*lpg1419*	0.859		*lpg2469*	0.864
lpg0415	0.771		*lpg1420*	0.838		*lpg2472*	0.880

lpg0418	0.867		*lpg1421*	0.866		*lpg2473*	0.862
lpg0419	0.857		*lpg1422*	0.766		*lpg2475*	0.889
lpg0421	0.858		*lpg1424*	0.850		*lpg2476*	0.748
lpg0422	0.866		*lpg1425*	0.863		*lpg2481*	0.896
lpg0423	0.834		*lpg1429*	0.826		*lpg2483*	0.825
lpg0424	0.849		*lpg1430*	0.845		*lpg2484*	0.867
lpg0425	0.865		*lpg1431*	0.769		*lpg2485*	0.893
lpg0426	0.625		*lpg1432*	0.848		*lpg2487*	0.830
lpg0428	0.858		*lpg1434*	0.830		*lpg2491*	0.841
lpg0432	0.863		*lpg1435*	0.831		*lpg2493*	0.855
lpg0433	0.850		*lpg1441*	0.840		*lpg2494*	0.870
lpg0439	0.859		*lpg1444*	0.841		*lpg2495*	0.881
lpg0440	0.706		*lpg1445*	0.846		*lpg2497*	0.875
lpg0442	0.821		*lpg1446*	0.851		*lpg2500*	0.888
lpg0443	0.847		*lpg1447*	0.847		*lpg2506*	0.908
lpg0444	0.843		*lpg1451*	0.849		*lpg2507*	0.898
lpg0445	0.854		*lpg1452*	0.858		*lpg2513*	0.885
lpg0446	0.870		*lpg1453*	0.847		*lpg2514*	0.876
lpg0447	0.844		*lpg1455*	0.763		*lpg2515*	0.859
lpg0448	0.847		*lpg1456*	0.844		*lpg2516*	0.878
lpg0449	0.758		*lpg1457*	0.876		*lpg2517*	0.695
lpg0450	0.865		*lpg1459*	0.846		*lpg2518*	0.864
lpg0452	0.822		*lpg1460*	0.847		*lpg2520*	0.827
lpg0453	0.827		*lpg1461*	0.855		*lpg2526*	0.863
lpg0454	0.771		*lpg1462*	0.837		*lpg2528*	0.877
lpg0455	0.812		*lpg1463*	0.858		*lpg2530*	0.862
lpg0456	0.813		*lpg1464*	0.845		*lpg2531*	0.855
lpg0457	0.817		*lpg1466*	0.846		*lpg2532*	0.861
lpg0458	0.832		*lpg1469*	0.849		*lpg2534*	0.736
lpg0459	0.797		*lpg1472*	0.867		*lpg2535*	0.856
lpg0460	0.879		*lpg1473*	0.877		*lpg2536*	0.856
lpg0461	0.875		*lpg1474*	0.875		*lpg2538*	0.890
lpg0462	0.860		*lpg1475*	0.853		*lpg2544*	0.890
lpg0463	0.788		*lpg1476*	0.786		*lpg2547*	0.850
lpg0464	0.737		*lpg1477*	0.872		*lpg2549*	0.836
lpg0468	0.850		*lpg1482*	0.842		*lpg2552*	0.849
lpg0469	0.823		*lpg1483*	0.860		*lpg2554*	0.818
lpg0471	0.735		*lpg1484*	0.852		*lpg2576*	0.873
lpg0473	0.651		*lpg1485*	0.842		*lpg2577*	0.890
lpg0474	0.728		*lpg1486*	0.853		*lpg2578*	0.753
lpg0475	0.581		*lpg1487*	0.829		*lpg2579*	0.862
lpg0476	0.654		*lpg1502*	0.856		*lpg2580*	0.866
lpg0477	0.862		*lpg1503*	0.857		*lpg2581*	0.870
lpg0478	0.607		*lpg1504*	0.854		*lpg2585*	0.863

lpg0479	0.802		*lpg1505*	0.596		*lpg2586*	0.863
lpg0481	0.833		*lpg1506*	0.841		*lpg2587*	0.878
lpg0482	0.859		*lpg1507*	0.843		*lpg2589*	0.865
lpg0483	0.866		*lpg1508*	0.853		*lpg2590*	0.859
lpg0485	0.747		*lpg1509*	0.848		*lpg2592*	0.840
lpg0491	0.865		*lpg1511*	0.851		*lpg2594*	0.835
lpg0493	0.845		*lpg1512*	0.864		*lpg2595*	0.846
lpg0497	0.856		*lpg1513*	0.844		*lpg2596*	0.866
lpg0498	0.888		*lpg1514*	0.840		*lpg2597*	0.885
lpg0499	0.858		*lpg1517*	0.846		*lpg2598*	0.834
lpg0500	0.885		*lpg1519*	0.879		*lpg2601*	0.829
lpg0506	0.900		*lpg1520*	0.865		*lpg2602*	0.861
lpg0507	0.873		*lpg1524*	0.863		*lpg2604*	0.867
lpg0510	0.859		*lpg1526*	0.835		*lpg2605*	0.862
lpg0511	0.809		*lpg1527*	0.879		*lpg2606*	0.837
lpg0512	0.669		*lpg1529*	0.888		*lpg2608*	0.835
lpg0513	0.901		*lpg1530*	0.882		*lpg2611*	0.850
lpg0525	0.865		*lpg1531*	0.881		*lpg2614*	0.755
lpg0528	0.836		*lpg1534*	0.882		*lpg2615*	0.841
lpg0529	0.827		*lpg1535*	0.714		*lpg2616*	0.858
lpg0530	0.862		*lpg1536*	0.867		*lpg2619*	0.858
lpg0531	0.828		*lpg1537*	0.867		*lpg2620*	0.908
lpg0532	0.884		*lpg1539*	0.850		*lpg2621*	0.832
lpg0533	0.864		*lpg1540*	0.843		*lpg2622*	0.882
lpg0534	0.860		*lpg1541*	0.887		*lpg2623*	0.840
lpg0535	0.856		*lpg1542*	0.883		*lpg2624*	0.832
lpg0536	0.859		*lpg1543*	0.880		*lpg2625*	0.899
lpg0539	0.838		*lpg1545*	0.849		*lpg2626*	0.831
lpg0540	0.883		*lpg1546*	0.844		*lpg2627*	0.846
lpg0541	0.838		*lpg1547*	0.849		*lpg2628*	0.769
lpg0542	0.483		*lpg1548*	0.837		*lpg2629*	0.790
lpg0547	0.849		*lpg1549*	0.861		*lpg2630*	0.870
lpg0548	0.740		*lpg1550*	0.842		*lpg2631*	0.871
lpg0551	0.846		*lpg1553*	0.845		*lpg2632*	0.764
lpg0552	0.727		*lpg1554*	0.873		*lpg2633*	0.842
lpg0556	0.753		*lpg1558*	0.841		*lpg2634*	0.811
lpg0557	0.827		*lpg1559*	0.844		*lpg2635*	0.829
lpg0558	0.780		*lpg1562*	0.857		*lpg2636*	0.561
lpg0559	0.843		*lpg1564*	0.852		*lpg2641*	0.863
lpg0560	0.834		*lpg1565*	0.850		*lpg2643*	0.869
lpg0561	0.859		*lpg1566*	0.849		*lpg2645*	0.874
lpg0562	0.754		*lpg1567*	0.853		*lpg2650*	0.075
lpg0563	0.672		*lpg1568*	0.864		*lpg2651*	0.678
lpg0564	0.871		*lpg1573*	0.837		*lpg2652*	0.761

lpg0565	0.815		*lpg1575*	0.847		*lpg2653*	0.757
lpg0566	0.758		*lpg1576*	0.862		*lpg2654*	0.877
lpg0568	0.868		*lpg1577*	0.840		*lpg2655*	0.866
lpg0577	0.854		*lpg1578*	0.664		*lpg2656*	0.864
lpg0580	0.853		*lpg1579*	0.847		*lpg2657*	0.877
lpg0581	0.833		*lpg1580*	0.878		*lpg2658*	0.772
lpg0583	0.843		*lpg1582*	0.846		*lpg2659*	0.861
lpg0584	0.820		*lpg1584*	0.854		*lpg2660*	0.858
lpg0585	0.798		*lpg1585*	0.790		*lpg2661*	0.856
lpg0586	0.756		*lpg1586*	0.864		*lpg2662*	0.786
lpg0587	0.734		*lpg1587*	0.751		*lpg2663*	0.782
lpg0588	0.840		*lpg1589*	0.836		*lpg2666*	0.865
lpg0591	0.560		*lpg1592*	0.808		*lpg2667*	0.834
lpg0592	0.789		*lpg1593*	0.858		*lpg2668*	0.864
lpg0593	0.854		*lpg1595*	0.859		*lpg2671*	0.861
lpg0594	0.616		*lpg1596*	0.865		*lpg2672*	0.862
lpg0595	0.860		*lpg1597*	0.867		*lpg2673*	0.859
lpg0596	0.862		*lpg1604*	0.847		*lpg2674*	0.761
lpg0598	0.866		*lpg1605*	0.808		*lpg2677*	0.868
lpg0599	0.876		*lpg1612*	0.853		*lpg2678*	0.855
lpg0600	0.864		*lpg1618*	0.863		*lpg2679*	0.846
lpg0601	0.882		*lpg1620*	0.860		*lpg2680*	0.870
lpg0602	0.816		*lpg1623*	0.866		*lpg2682*	0.828
lpg0603	0.885		*lpg1624*	0.854		*lpg2684*	0.872
lpg0604	0.858		*lpg1636*	0.887		*lpg2687*	0.833
lpg0605	0.844		*lpg1638*	0.877		*lpg2688*	0.838
lpg0606	0.793		*lpg1639*	0.874		*lpg2690*	0.866
lpg0607	0.856		*lpg1640*	0.830		*lpg2691*	0.866
lpg0608	0.851		*lpg1641*	0.887		*lpg2692*	0.846
lpg0611	0.858		*lpg1644*	0.800		*lpg2693*	0.846
lpg0612	0.864		*lpg1645*	0.857		*lpg2694*	0.854
lpg0614	0.722		*lpg1646*	0.849		*lpg2696*	0.818
lpg0616	0.880		*lpg1650*	0.881		*lpg2698*	0.895
lpg0618	0.838		*lpg1653*	0.846		*lpg2699*	0.751
lpg0622	0.846		*lpg1656*	0.860		*lpg2700*	0.864
lpg0623	0.834		*lpg1657*	0.845		*lpg2701*	0.843
lpg0624	0.814		*lpg1659*	0.856		*lpg2702*	0.836
lpg0626	0.851		*lpg1661*	0.854		*lpg2703*	0.859
lpg0627	0.822		*lpg1662*	0.861		*lpg2704*	0.871
lpg0629	0.842		*lpg1663*	0.746		*lpg2705*	0.816
lpg0630	0.843		*lpg1666*	0.870		*lpg2706*	0.802
lpg0631	0.814		*lpg1667*	0.855		*lpg2707*	0.677
lpg0633	0.816		*lpg1669*	0.870		*lpg2708*	0.785
lpg0634	0.855		*lpg1672*	0.794		*lpg2709*	0.830

lpg0640	0.839		*lpg1674*	0.837		*lpg2710*	0.863
lpg0641	0.872		*lpg1679*	0.797		*lpg2711*	0.820
lpg0643	0.875		*lpg1680*	0.837		*lpg2712*	0.525
lpg0650	0.520		*lpg1682*	0.846		*lpg2713*	0.803
lpg0651	0.855		*lpg1690*	0.863		*lpg2714*	0.861
lpg0652	0.865		*lpg1696*	0.880		*lpg2716*	0.794
lpg0654	0.824		*lpg1697*	0.727		*lpg2717*	0.776
lpg0656	0.878		*lpg1698*	0.680		*lpg2719*	0.859
lpg0657	0.862		*lpg1699*	0.804		*lpg2720*	0.859
lpg0658	0.889		*lpg1700*	0.723		*lpg2722*	0.851
lpg0659	0.883		*lpg1701*	0.863		*lpg2724*	0.840
lpg0660	0.879		*lpg1705*	0.861		*lpg2725*	0.766
lpg0662	0.887		*lpg1706*	0.852		*lpg2726*	0.845
lpg0663	0.877		*lpg1707*	0.858		*lpg2727*	0.858
lpg0664	0.857		*lpg1710*	0.782		*lpg2732*	0.807
lpg0665	0.816		*lpg1711*	0.794		*lpg2735*	0.831
lpg0667	0.874		*lpg1712*	0.851		*lpg2736*	0.858
lpg0670	0.818		*lpg1713*	0.851		*lpg2737*	0.869
lpg0672	0.847		*lpg1714*	0.858		*lpg2739*	0.871
lpg0673	0.506		*lpg1720*	0.861		*lpg2740*	0.747
lpg0674	0.878		*lpg1721*	0.852		*lpg2741*	0.666
lpg0677	0.616		*lpg1722*	0.862		*lpg2742*	0.804
lpg0678	0.848		*lpg1723*	0.862		*lpg2743*	0.861
lpg0679	0.879		*lpg1724*	0.838		*lpg2755*	0.843
lpg0680	0.871		*lpg1725*	0.838		*lpg2756*	0.766
lpg0685	0.852		*lpg1727*	0.858		*lpg2757*	0.854
lpg0686	0.871		*lpg1730*	0.852		*lpg2758*	0.864
lpg0687	0.393		*lpg1731*	0.850		*lpg2760*	0.839
lpg0688	0.869		*lpg1732*	0.859		*lpg2762*	0.856
lpg0689	0.815		*lpg1733*	0.854		*lpg2763*	0.838
lpg0692	0.881		*lpg1734*	0.857		*lpg2764*	0.830
lpg0697	0.896		*lpg1735*	0.708		*lpg2765*	0.810
lpg0698	0.864		*lpg1736*	0.858		*lpg2766*	0.858
lpg0699	0.813		*lpg1737*	0.862		*lpg2769*	0.824
lpg0700	0.794		*lpg1743*	0.698		*lpg2772*	0.870
lpg0701	0.861		*lpg1744*	0.842		*lpg2773*	0.869
lpg0704	0.820		*lpg1746*	0.854		*lpg2774*	0.840
lpg0712	0.881		*lpg1747*	0.843		*lpg2777*	0.868
lpg0716	0.831		*lpg1748*	0.821		*lpg2778*	0.873
lpg0719	0.866		*lpg1749*	0.827		*lpg2779*	0.522
lpg0720	0.872		*lpg1750*	0.876		*lpg2780*	0.838
lpg0721	0.865		*lpg1751*	0.851		*lpg2781*	0.844
lpg0722	0.441		*lpg1752*	0.747		*lpg2782*	0.860
lpg0723	0.849		*lpg1753*	0.869		*lpg2783*	0.860

lpg0724	0.670		*lpg1754*	0.840		*lpg2785*	0.851
lpg0725	0.863		*lpg1755*	0.853		*lpg2786*	0.868
lpg0726	0.803		*lpg1756*	0.732		*lpg2787*	0.809
lpg0729	0.802		*lpg1757*	0.817		*lpg2788*	0.832
lpg0730	0.809		*lpg1758*	0.787		*lpg2789*	0.803
lpg0732	0.753		*lpg1759*	0.818		*lpg2791*	0.831
lpg0734	0.852		*lpg1761*	0.765		*lpg2792*	0.855
lpg0737	0.842		*lpg1762*	0.820		*lpg2794*	0.841
lpg0738	0.854		*lpg1763*	0.813		*lpg2795*	0.858
lpg0739	0.846		*lpg1764*	0.802		*lpg2796*	0.867
lpg0740	0.809		*lpg1765*	0.789		*lpg2797*	0.839
lpg0741	0.800		*lpg1766*	0.814		*lpg2798*	0.858
lpg0742	0.841		*lpg1767*	0.790		*lpg2799*	0.864
lpg0745	0.833		*lpg1768*	0.709		*lpg2805*	0.861
lpg0747	0.842		*lpg1770*	0.472		*lpg2806*	0.860
lpg0748	0.844		*lpg1771*	0.786		*lpg2808*	0.861
lpg0749	0.728		*lpg1772*	0.770		*lpg2809*	0.868
lpg0752	0.878		*lpg1778*	0.805		*lpg2812*	0.861
lpg0753	0.845		*lpg1779*	0.551		*lpg2814*	0.863
lpg0754	0.849		*lpg1782*	0.827		*lpg2817*	0.855
lpg0755	0.420		*lpg1785*	0.822		*lpg2818*	0.846
lpg0759	0.871		*lpg1788*	0.766		*lpg2819*	0.853
lpg0760	0.792		*lpg1789*	0.828		*lpg2822*	0.858
lpg0781	0.414		*lpg1791*	0.822		*lpg2823*	0.850
lpg0785	0.857		*lpg1792*	0.821		*lpg2824*	0.863
lpg0786	0.908		*lpg1793*	0.777		*lpg2825*	0.000
lpg0791	0.784		*lpg1798*	0.885		*lpg2827*	0.870
lpg0800	0.786		*lpg1800*	0.449		*lpg2833*	0.756
lpg0801	0.782		*lpg1803*	0.771		*lpg2835*	0.863
lpg0802	0.786		*lpg1804*	0.855		*lpg2836*	0.876
lpg0803	0.800		*lpg1805*	0.906		*lpg2837*	0.851
lpg0804	0.736		*lpg1806*	0.825		*lpg2838*	0.846
lpg0805	0.757		*lpg1807*	0.893		*lpg2842*	0.874
lpg0808	0.741		*lpg1808*	0.792		*lpg2843*	0.872
lpg0810	0.772		*lpg1809*	0.531		*lpg2847*	0.851
lpg0811	0.807		*lpg1810*	0.682		*lpg2848*	0.856
lpg0812	0.799		*lpg1811*	0.685		*lpg2851*	0.869
lpg0815	0.701		*lpg1812*	0.758		*lpg2853*	0.855
lpg0816	0.794		*lpg1813*	0.850		*lpg2855*	0.867
lpg0817	0.759		*lpg1814*	0.660		*lpg2858*	0.855
lpg0818	0.832		*lpg1815*	0.881		*lpg2859*	0.853
lpg0821	0.809		*lpg1816*	0.893		*lpg2860*	0.849
lpg0822	0.832		*lpg1821*	0.880		*lpg2861*	0.863
lpg0823	0.649		*lpg1823*	0.870		*lpg2864*	0.872

lpg0824	0.792		*lpg1824*	0.869		*lpg2865*	0.827
lpg0825	0.786		*lpg1825*	0.895		*lpg2867*	0.862
lpg0826	0.818		*lpg1826*	0.557		*lpg2868*	0.864
lpg0829	0.801		*lpg1830*	0.909		*lpg2869*	0.882
lpg0833	0.788		*lpg1831*	0.889		*lpg2872*	0.724
lpg0834	0.807		*lpg1832*	0.857		*lpg2873*	0.858
lpg0835	0.783		*lpg1833*	0.847		*lpg2874*	0.856
lpg0836	0.774		*lpg1834*	0.861		*lpg2875*	0.863
lpg0837	0.768		*lpg1835*	0.840		*lpg2878*	0.855
lpg0838	0.744		*lpg1836*	0.865		*lpg2879*	0.864
lpg0839	0.781		*lpg1837*	0.879		*lpg2880*	0.859
lpg0840	0.799		*lpg1838*	0.861		*lpg2881*	0.848
lpg0841	0.786		*lpg1839*	0.884		*lpg2882*	0.884
lpg0842	0.820		*lpg1840*	0.840		*lpg2883*	0.854
lpg0843	0.452		*lpg1841*	0.867		*lpg2884*	0.854
lpg0845	0.792		*lpg1842*	0.872		*lpg2885*	0.866
lpg0846	0.739		*lpg1843*	0.778		*lpg2886*	0.865
lpg0847	0.809		*lpg1844*	0.782		*lpg2887*	0.862
lpg0848	0.813		*lpg1845*	0.804		*lpg2890*	0.856
lpg0849	0.765		*lpg1846*	0.849		*lpg2891*	0.867
lpg0851	0.817		*lpg1847*	0.856		*lpg2894*	0.868
lpg0852	0.772		*lpg1849*	0.817		*lpg2897*	0.877
lpg0853	0.785		*lpg1850*	0.836		*lpg2898*	0.868
lpg0854	0.699		*lpg1851*	0.854		*lpg2899*	0.836
lpg0856	0.830		*lpg1854*	0.840		*lpg2900*	0.870
lpg0858	0.794		*lpg1855*	0.918		*lpg2901*	0.841
lpg0859	0.099		*lpg1858*	0.233		*lpg2902*	0.863
lpg0860	0.731		*lpg1859*	0.876		*lpg2903*	0.851
lpg0862	0.819		*lpg1860*	0.882		*lpg2904*	0.870
lpg0865	0.799		*lpg1861*	0.872		*lpg2905*	0.879
lpg0866	0.821		*lpg1869*	0.918		*lpg2907*	0.846
lpg0867	0.841		*lpg1870*	0.876		*lpg2908*	0.883
lpg0869	0.834		*lpg1871*	0.921		*lpg2916*	0.869
lpg0870	0.844		*lpg1873*	0.879		*lpg2924*	0.868
lpg0871	0.725		*lpg1874*	0.856		*lpg2925*	0.871
lpg0872	0.819		*lpg1882*	0.907		*lpg2926*	0.779
lpg0873	0.801		*lpg1883*	0.825		*lpg2927*	0.869
lpg0874	0.824		*lpg1887*	0.862		*lpg2928*	0.860
lpg0875	0.502		*lpg1888*	0.880		*lpg2929*	0.827
lpg0877	0.829		*lpg1889*	0.882		*lpg2930*	0.864
lpg0878	0.600		*lpg1891*	0.780		*lpg2931*	0.815
lpg0879	0.827		*lpg1892*	0.862		*lpg2933*	0.877
lpg0880	0.822		*lpg1893*	0.879		*lpg2934*	0.867
lpg0882	0.804		*lpg1894*	0.869		*lpg2935*	0.814

lpg0885	0.813		*lpg1895*	0.793		*lpg2937*	0.867
lpg0886	0.824		*lpg1896*	0.859		*lpg2951*	0.868
lpg0887	0.803		*lpg1904*	0.853		*lpg2953*	0.861
lpg0888	0.806		*lpg1905*	0.852		*lpg2955*	0.735
lpg0889	0.674		*lpg1906*	0.878		*lpg2956*	0.846
lpg0890	0.796		*lpg1908*	0.876		*lpg2957*	0.840
lpg0891	0.867		*lpg1909*	0.882		*lpg2960*	0.850
lpg0892	0.826		*lpg1910*	0.883		*lpg2962*	0.864
lpg0895	0.716		*lpg1911*	0.884		*lpg2963*	0.846
lpg0896	0.811		*lpg1913*	0.887		*lpg2964*	0.863
lpg0897	0.792		*lpg1915*	0.882		*lpg2965*	0.850
lpg0899	0.821		*lpg1916*	0.870		*lpg2966*	0.689
lpg0900	0.832		*lpg1917*	0.911		*lpg2967*	0.843
lpg0901	0.805		*lpg1918*	0.882		*lpg2968*	0.866
lpg0902	0.817		*lpg1919*	0.877		*lpg2969*	0.852
lpg0904	0.801		*lpg1920*	0.785		*lpg2970*	0.855
lpg0905	0.819		*lpg1921*	0.878		*lpg2971*	0.868
lpg0906	0.783		*lpg1924*	0.892		*lpg2972*	0.870
lpg0907	0.807		*lpg1927*	0.796		*lpg2974*	0.867
lpg0908	0.785		*lpg1942*	0.860		*lpg2975*	0.869
lpg0909	0.798		*lpg1943*	0.562		*lpg2976*	0.860
lpg0910	0.795		*lpg1944*	0.869		*lpg2982*	0.850
lpg0911	0.786		*lpg1945*	0.864		*lpg2983*	0.842
lpg0915	0.714		*lpg1949*	0.864		*lpg2985*	0.730
lpg0917	0.807		*lpg1993*	0.849		*lpg2986*	0.708
lpg0918	0.816		*lpg1994*	0.867		*lpg2987*	0.050
lpg0919	0.818		*lpg1999*	0.765		*lpg2990*	0.399
lpg0920	0.771		*lpg2000*	0.866		*lpg2991*	0.829
lpg0921	0.860		*lpg2001*	0.868		*lpg2993*	0.847
lpg0922	0.840		*lpg2002*	0.663		*lpg2994*	0.741
lpg0923	0.842		*lpg2004*	0.869		*lpg2995*	0.810
lpg0924	0.841		*lpg2007*	0.845		*lpg2996*	0.852
lpg0925	0.853		*lpg2008*	0.842		*lpg2997*	0.856
lpg0926	0.849		*lpg2009*	0.857		*lpg2998*	0.868
lpg0927	0.837		*lpg2010*	0.843		*lpg2999*	0.869
lpg0928	0.837		*lpg2011*	0.864		*lpg3002*	0.859
lpg0929	0.827		*lpg2012*	0.807		*lpg3005*	0.516

9.4 Chapter 6

Appendix Table 26. 229 ST1 and ST1-derived isolates used in the study. References are provided for previously published genomes and run accession numbers are provided for genomes that were newly sequenced for this thesis. ST – sequence type; mAb subgroup – monoclonal antibody subgroup; Phil. – Philadelphia; All./France – Allentown/France; Camp – Camperdown; NA – not applicable; U/k – unknown.

Hospital	Isolate	Source	Known exposures during incubation period (up to ~18 days)	Hospital ward (if known)	Date of isolation	Town/ Region	Country	ST/mAb subgroup	Accession number/ Reference
Environmental isolates from hospitals or clinical isolates with confirmed/suspected links to hospitals (n=141)									
A	H072360604 (case 1)	Clinical (pleural fluid)	Hospital A (11-18 days), home	B	24/05/2007	Essex	UK	1/Phil.	ERR1399547
	H072360603 (case 2)	Clinical (sputum)	Hospital A (~12 days)	A	27/05/2007	Essex	UK	1/Phil.	ERR1399550
	H100120270 (case 3)	Clinical (sputum)	Hospital A (~4 days), home and local area	F & G	29/12/2009	Essex	UK	1/Phil.	ERR1399506
	H100120260 (case 4)	Clinical (sputum)	Hospital A (~7 days), home and local area	E	29/12/2009	Essex	UK	1/Phil.	ERR1399540
	H104720329 (case 5)	Clinical (sputum)	Hospital A (~7 days), home and local area	A	19/11/2010	Essex	UK	1/Phil.	ERR1399526
	H113580549 (case 6)	Clinical (post-mortem sample from left lung)	Hospital A (at least 10 days)	H	23/08/2011	Essex	UK	1/Phil.	ERR1399560

	H113580550 (case 6)	Clinical (post-mortem sample of right lung of same patient as above)		H	23/08/2011	Essex	UK	1/Phil.	ERR1399535
	H114820438 (case 7)	Clinical	Hospital A (at least 10 days)	G	24/11/2011	Essex	UK	1/All./ France	ERR1399537
	H072560534	Environmental (carpet cleaner reservoir)	NA	C	09/01/2007	Essex	UK	1/Camp.	ERR1399501
	H072300480	Environmental (hot sink, heat-treated sample)	NA	A	30/05/2007	Essex	UK	1/Phil.	ERR1399554
	H072300481	Environmental (hot sink, untreated sample)	NA	A	30/05/2007	Essex	UK	1/Phil.	ERR1399565
	H072680210	Environmental (day room; hot thermostatic mixing valve)	NA	A	07/06/2007	Essex	UK	1/Phil.	ERR1399562
	H072680211	Environmental (staff room; chilled cold water)	NA	A	07/06/2007	Essex	UK	1/Phil.	ERR1399556
	H072680212	Environmental (steam cleaner)	NA	Multiple	07/06/2007	Essex	UK	1/Camp.	ERR1399551
	H072680213	Environmental (steam cleaner)	NA	Multiple	07/06/2007	Essex	UK	1/Phil.	ERR1399559
	H111920394	Environmental	NA	D	07/06/2007	Essex	UK	1/Phil.	ERR1399545
	H111920398	Environmental	NA	D	07/06/2007	Essex	UK	1/Phil.	ERR1399499
	H111920400	Environmental (sink)	NA	B	07/06/2007	Essex	UK	1/Phil.	ERR1399512
	H111920402	Environmental (same as	NA	B	07/06/2007	Essex	UK	1/Phil.	ERR1399558

		above; acid-treated sample)							
	H111920404	Environmental (wash hand basin)	NA	B	07/06/2007	Essex	UK	1/Phil.	ERR1399544
	H100180614	Environmental (sink, hot tap)	NA	G	31/12/2009	Essex	UK	1/Phil.	ERR1399523
	H100180615	Environmental (sink, hot tap)	NA	G	31/12/2009	Essex	UK	1/Phil.	ERR1399508
	H100180616	Environmental (shower)	NA	E	31/12/2009	Essex	UK	1/Phil.	ERR1399511
	H100180617	Environmental (shower)	NA	E	31/12/2009	Essex	UK	1/Phil.	ERR1399549
	H100280679	Environmental (sink, hot tap)	NA	G	07/01/2010	Essex	UK	1/Phil.	ERR1399539
	H100280682	Environmental (shower)	NA	G	07/01/2010	Essex	UK	1/Phil.	ERR1399516
	H100280683	Environmental (shower)	NA	G	07/01/2010	Essex	UK	1/Phil.	ERR1399505
	H100280685	Environmental (toilet basin)	NA	G	07/01/2010	Essex	UK	1/Phil.	ERR1399503
	H100560548	Environmental (sink, cold tap)	NA	G	14/01/2010	Essex	UK	1/Phil.	ERR1399520
	H100560549	Environmental (shower)	NA	G	14/01/2010	Essex	UK	1/Phil.	ERR1399561
	H112000588	Environmental (from patient's room – although no clinical isolate from patient)	NA	D	12/03/2010	Essex	UK	1/Phil.	ERR1399518
	H104780626	Environmental (sink)	NA	A	19/11/2010	Essex	UK	1/Phil.	ERR1399525
	H104780627	Environmental (sink)	NA	A	19/11/2010	Essex	UK	1/Phil.	ERR1399566
	H104780628	Environmental	NA	A	19/11/2010	Essex	UK	1/Phil.	ERR1399572

		(sink in toilet opposite patient's bed)							
	H113440612	Environmental (basin next to bed 9)	NA	H	24/08/2011	Essex	UK	1/Phil.	ERR1399555
	H113440613	Environmental (shower in room 14)	NA	H	24/08/2011	Essex	UK	1/Phil.	ERR1399533
	H113440614	Environmental (bath in room 13)	NA	H	24/08/2011	Essex	UK	1/Phil.	ERR1399570
	H113440615	Environmental (basin in side room 6)	NA	H	24/08/2011	Essex	UK	1/Phil.	ERR1399536
	H113440616	Environmental (basin in side room 6)	NA	H	24/08/2011	Essex	UK	1/Phil.	ERR1399530
	H114840676	Environmental (toilet, cold water)	NA	G	25/11/2011	Essex	UK	1/All./ France	ERR1399542
	H114840677	Environmental (toilet, hot water)	NA	G	25/11/2011	Essex	UK	1/All./ France	ERR1399553
	H114840678	Environmental (side room 13, cold water)	NA	G	25/11/2011	Essex	UK	1/All./ France	ERR1399498
	H114840679	Environmental (side room 13, hot water)	NA	G	25/11/2011	Essex	UK	1/All./ France	ERR1399567
	H114840680	Environmental (toilet)	NA	G	25/11/2011	Essex	UK	1/Phil.	ERR1399522
	H114840681	Environmental (toilet)	NA	G	25/11/2011	Essex	UK	1/Phil.	ERR1399546
	H120680630	Environmental	NA	I	02/02/2012	Essex	UK	1/Phil.	ERR1399569
B/The Wesley	LP01	Environmental	NA	East Wing/ Cardiac	29/05/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016

Hospital	LP02	Environmental	NA	East Wing/ Cardiac	29/05/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP03	Environmental	NA	East Wing/ Cardiac	29/05/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP04	Environmental	NA	East Wing/ Cardiac	29/05/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP05	Environmental	NA	East Wing/ Cardiac	29/05/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP06	Environmental	NA	Main block/ Hematology HDU	05/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP07	Environmental	NA	Main block/ Hematology HDU	05/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP08	Environmental	NA	Main block/ Palliative care	05/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP09	Environmental	NA	Main block/ Medical centre 1	06/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP10	Environmental	NA	Main block/ Medical centre 1	06/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP11	Environmental	NA	Main block/ Rehabilitation	06/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP12	Environmental	NA	Main block/ Rehabilitation	08/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP14	Environmental	NA	East Wing/ Obstretic	10/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP15	Environmental	NA	Hyperbaric Unit	12/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP16	Environmental	NA	Hemato-Oncology Day Facility	18/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP17	Environmental	NA	Main block/ Medical centre 1	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP18	Environmental	NA	Main block/	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.*

				Medical centre 1					2016
	LP19	Environmental	NA	Main block/ Internal medicine	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP20	Environmental	NA	Main block/ Hematology	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP21	Environmental	NA	Main block/ Hematology HDU	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP23	Environmental	NA	Main block/ Cardiac Catheter Suite	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP24	Environmental	NA	Main block/ Echocardiogr-aphy Laboratory	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP25	Environmental	NA	Main block/ Pediatric	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP26	Environmental	NA	Main block/ Pediatric	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP27	Environmental	NA	Main block/ Pediatric	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP28	Environmental	NA	Main block/ Pediatric	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP29	Environmental	NA	Main block/ Rehabilitation	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP30	Environmental	NA	Main block/ Rehabilitation	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP31	Environmental	NA	Main block/ Rehabilitation	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP32	Environmental	NA	Main block/ Dialysis	21/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP33	Environmental	NA	Main block/ Radiology	25/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP34	Environmental	NA	Main block/ Rehabilitation	28/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016

	LP35	Environmental	NA	Main block/ Rehabilitation	28/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP36	Environmental	NA	Main block/ Rehabilitation	28/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP37	Environmental	NA	Main block/ Rehabilitation	05/07/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP38	Environmental	NA	Main block/ Rehabilitation	05/07/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP39	Environmental	NA	Main block/ Rehabilitation	05/07/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP40	Environmental	NA	East Wing/ Obstretic	10/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP41	Environmental	NA	East Wing/ Breast and Endocrine Surgery	10/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP44 (case 8)	Clinical	Hospital B	Main Block/ Hematology HDU	14/10/2011	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP45 (case 9)	Clinical	Hospital B only	East Wing/ Cardiac	27/05/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP46 (case 9)	Clinical (from same patient as LP45)			31/05/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP47 (case 10)	Clinical	Hospital B only	Main block/ Hematology HDU	07/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
	LP48 (case 10)	Clinical (from same patient as LP47)			17/06/2013	Brisbane	Australia	1/U/k	Bartley *et al.* 2016
C	Paris (case 11)	Clinical	Hospital C (7 days) & another hospital near to Paris (4 days)	Intensive Care Unit (cardiac surgery) and cardiac surgery unit (Hospital C)	04/04/2002	Paris	France	1/Phil.	Cazalet *et al.* 2004

	HL 0051 1015 (case 12)	Clinical	Hospital C only	Nephrology	12/12/2000	Paris	France	1/Phil.	ERR1399510
	HL 0051 4008 (case 13)	Clinical	Hospital C (~17 days)	Intensive Care Unit (cardiac surgery)	18/12/2000	Paris	France	1/Oxford/ OLDA	ERR1399509
	HL 0101 3003 (case 14)	Clinical	Hospital C (~12 days)	Intensive Care Unit (cardiac surgery)	27/12/2000	Paris	France	1/Phil.	ERR1399564
	HL 0102 3034 (case 15)	Clinical	Hospital C (~4 days), home	Intensive Care Unit (cardiac surgery)	21/12/2000	Paris	France	1/Phil.	ERR1399514
	HL 0102 3035 (case 16)	Clinical	Hospital C (~4 days), home	Intensive Care Unit (cardiac surgery)	24/12/2000	Paris	France	1/Phil.	ERR1399517
	LG 0713 5006 (case 17)	Clinical	Hospital C only	Nephrology	22/03/2007	Paris	France	1/Phil.	ERR1399504
	HL 0131 3038	Environmental (shower room 4622)	NA	Oto-Rhino-Laryngology	01/08/2001	Paris	France	1/U/k	ERR1399500
	HL 0131 3039	Environmental (tap room 4622)	NA	Oto-Rhino-Laryngology	01/08/2001	Paris	France	1/U/k	ERR1399502
	Paris 2001 I n2	Environmental (room 7411, HWS, shower)	NA	Nephrology	18/12/2000	Paris	France	1/U/k	ERR1399552
	LG 0713 5007	Environmental (sink 1)	NA	Dialysis room	25/03/2007	Paris	France	1/Oxford/ OLDA	ERR1399563
	LG 0713 5008	Environmental (sink 3)	NA	Dialysis room	26/03/2007	Paris	France	1/Oxford/ OLDA	ERR1399521
D	LG 0918 2002 (case 18)	Clinical	Hospital D (~4 days), home	Internal medicine unit, room 618	08/04/2009	Near Marseille	France	1/Phil.	ERR1399543

	LG 1416 4007 (case 19)	Clinical	Hospital D (~3 days), home (~3 days)	Internal medicine unit, room 610	10/04/2014	Near Marseille	France	1/Phil.	ERR1399515
	LG 1416 4008 (case 20)	Clinical	Hospital D (~5 days)	Internal medicine unit, room 610	07/04/2014	Near Marseille	France	1/Phil.	ERR1399571
	LG 1427 4009	Environmental	NA	Internal medicine unit, room 610	17/06/2014	Near Marseille	France	1/Phil.	ERR1399497
	LG 1427 4010	Environmental	NA	Internal medicine unit, room 610	17/06/2014	Near Marseille	France	1/Phil.	ERR1399507
	LG 0918 2005	Environmental	NA	Internal medicine unit, room 618	16/04/2009	Near Marseille	France	1/Phil.	ERR1399529
E	H103120165 (case 21)	Clinical	Hospital E (at least 10 days)	U/k	08/06/2010	London	UK	1/Phil.	ERR1399524
	H124240908 (case 22)	Clinical	Hospital E (less than 10 days)	U/k	22/10/2012	London	UK	1/Phil.	ERR1441930
	H103340763	Environmental (sink)	NA	Gastroentero-logy and liver specialist ward	10/08/2010	London	UK	1/Phil.	ERR1399528
	H124600775	Environmental	NA	U/k	02/11/2012	London	UK	1/Phil.	ERR1441929
F	H115180236 (case A)	Clinical	Hospital F (3 days)	U/k	26/12/2011	London	UK	1/All./ France	ERR1441936
G	H101460286 (case 23)	Clinical	Hospital G (less than 10 days)	U/k	02/04/2010	Cambridge-shire	UK	1/Oxford/ OLDA	ERR1399527
	H101740836	Environmental	NA	U/k	20/04/2010	Cambridge-shire	UK	1/Oxford/ OLDA	ERR1399532
H	H092520167 (case 24)	Clinical	Hospital H (at least 10 days)	U/k	19/06/2009	London	UK	1/Oxford/ OLDA	ERR1441933
	H092620872 (pick 1: H092620872 24)	Environmental	NA	U/k	19/06/2009	London	UK	1/Oxford/ OLDA	ERR1441934

I	H134660746	Environmental	NA	U/k	Approx. 14/11/2013	Essex	UK	1/Oxford/ OLDA	ERR1441927
J	H102860194	Clinical	Hospital J (less than 10 days)	U/k	20/07/2010	Near London	UK	1/Oxford/ OLDA	ERR1441931
K	H074360702	Environmental	NA	U/k	01/10/2007	Kent	UK	152/ Oxford/ OLDA	Underwood *et al*. 2013
L	EUL 55 (case 25)	Clinical	Hospital L	U/k	01/04/1994	Cáceres province	Spain	1/Oxford/ OLDA	ERR332141
	EUL 58	Environmental	NA	U/k	01/01/1994	Cáceres province	Spain	1/Oxford/ OLDA	ERR376683
M	EUL 93 (case 26)	Clinical	Hospital M only	U/k	19/10/1992	Copenhagen	Denmark	1/Oxford/ OLDA	ERR332179
	EUL 94 (case 27)	Clinical	Hospital M only	U/k	08/12/1992	Copenhagen	Denmark	1/Oxford/ OLDA	ERR376738
	EUL 95	Environmental	NA	U/k	21/01/1993	Copenhagen	Denmark	1/Oxford/ OLDA	ERR376739
N	LG 1019 1002 (case 28)	Clinical	Hospital N only	U/k	28/04/2010	Near Marseille	France	1/Phil.	ERR922491
	LG 1020 3012	Environmental	NA	U/k	27/04/2010	Near Marseille	France	1/Phil.	ERR922492
O	HL 0311 1005	Environmental	NA	Room 1010	07/03/2003	Nice	France	1/Oxford/ OLDA	ERR922485
P	EUL 82	Clinical	Hospital P and home	U/k	29/08/1994	Near Copenhagen	Denmark	1/Oxford/ OLDA	ERR376733
	EUL 85	Clinical	Hospital P	U/k	01/05/1995	Near Copenhagen	Denmark	1/Oxford/ OLDA	ERR376710
Q	EUL 88	Clinical	Hospital Q only	U/k	11/10/1995	Near Copenhagen	Denmark	1/Oxford/ OLDA	ERR332174
R	LG 1139 1124	Environmental	NA	U/k	14/09/2011	Near Lyon	France	1/Oxford/ OLDA	ERR922496
S/ Bundaberg Hospital	LP43	Clinical	Hospital S	U/k	01/03/2001	Bundaberg	Australia	1/U/k	Bartley *et al.* 2016
T	L00-549	Clinical	Hospital T	U/k	2000	Dresden	Germany	1/U/k	ERR923394
U	EUL 157	Environmental	NA	U/k	01/07/2004	Blackpool	UK	8/	ERR376779

								Heysham	
V	HL 0230 4015	Clinical	Hospital V (at least 10 days)	U/k	23/07/2002	Near Paris	France	1/Phil.	ERR922484
W	HL 0416 3014	Clinical	Hospital W (12 days)	U/k	25/03/2004	Brittany	France	1/Oxford/ OLDA	ERR922487
X	LG 1101 1012	Environmental	NA	U/k	09/12/2010	Haute-Marne region	France	1/Phil.	ERR922493
Y	EUL 16	Clinical	Hospital Y	U/k	06/06/1984	Glasgow	UK	5/ Benidorm	ERR376641
Z	ID_6885	Environmental	NA	U/k	29/04/2011	U/k	Spain	1/U/k	Sanchez-Buso *et al.* 2014
α	NIIB80	Clinical	Hospital α	U/k	1981	Nagasaki	Japan	1/U/k	ERR923392
Isolates from or associated with community sources (i.e. with no links to hospitals) (n=47)									
	EUL 84	Clinical			03/04/1995	U/k	Denmark	1/Oxford/ OLDA	ERR376735
	HL 0036 4001	Clinical			22/08/2000	Paris	France	1/Phil.	ERR922483
	HL 0337 3012	Environmental			09/09/2003	Poitiers	France	1/U/k	ERR922486
	LG 0725 3019	Environmental			04/06/2007	Poitiers	France	1/Phil.	ERR1399568
	LG 0725 3022	Environmental			04/06/2007	Poitiers	France	1/U/k	ERR1399531
	LG 1014 3009	Clinical			30/03/2010	U/k	France	1/Oxford/ OLDA	ERR922504
	LG 0940 4015	Clinical			24/09/2009	Lyon	France	1/Phil.	ERR922490
	Lp-032	Environmental			U/k	U/k	Israel	1/U/k	Moran-Gilad *et al.* 2015
	Lp-119	Environmental			23/04/2013	U/k	Israel	1/U/k	Moran-Gilad *et al.* 2015
	Lp-120	Environmental			23/04/2013	U/k	Israel	1/U/k	Moran-Gilad *et al.* 2015
	Lp-121	Environmental			23/04/2013	U/k	Israel	1/U/k	Moran-Gilad *et al.* 2015

	Lp-122	Environmental			23/04/2013	U/k	Israel	1/U/k	Moran-Gilad *et al.* 2015
	Lp-2002694p8	Environmental			U/k	U/k	Israel	1/Oxford/ OLDA	Moran-Gilad *et al.* 2015
	Lp-282-1	Environmental			01/08/2013	U/k	Israel	1/U/k	Moran-Gilad *et al.* 2015
	Lp-283	Environmental			01/08/2013	U/k	Israel	1/U/k	Moran-Gilad *et al.* 2015
	Lp-284	Environmental			01/08/2013	U/k	Israel	1/U/k	Moran-Gilad *et al.* 2015
	Lp-285	Environmental			01/08/2013	U/k	Israel	1/U/k	Moran-Gilad *et al.* 2015
	Lp-286-1	Environmental			01/08/2013	U/k	Israel	1/U/k	Moran-Gilad *et al.* 2015
	Lp-56207	Clinical			U/k	U/k	Israel	1/Oxford/ OLDA	Moran-Gilad *et al.* 2015
	EUL 53	Clinical			01/05/1995	U/k	Spain	1/Oxford/ OLDA	ERR376725
	ID_1688	Environmental			23/06/2004	U/k	Spain	1/U/k	Sanchez-Buso *et al.* 2014
	ID_1690	Environmental			23/06/2004	U/k	Spain	1/U/k	Sanchez-Buso *et al.* 2014
	ID_1828	Environmental			20/09/2004	U/k	Spain	1/U/k	Sanchez-Buso *et al.* 2014
	ID_2041	Environmental			15/06/2005	U/k	Spain	1/Oxford/ OLDA	Sanchez-Buso *et al.* 2014
	ID_2947	Environmental			13/06/2000	U/k	Spain	1/Oxford/ OLDA	Sanchez-Buso *et al.* 2014
	ID_2948	Environmental			13/06/2000	U/k	Spain	1/Oxford/ OLDA	Sanchez-Buso *et al.* 2014
	ID_598	Environmental			07/02/2002	U/k	Spain	1/Pontiac /Knoxville	Sanchez-Buso *et al.* 2014
	ID_747970	Environmental			21/08/2009	U/k	Spain	1/U/k	Sanchez-Buso *et al.* 2014
	ID_891	Environmental			03/09/2002	U/k	Spain	1/Oxford/ OLDA	Sanchez-Buso *et al.* 2014
	EUL 104	Clinical			01/01/1992	U/k	Sweden	1/Oxford/	ERR376745

								OLDA	
	EUL 108	Clinical			01/01/1992	U/k	Sweden	1/All./ France	ERR376748
	EUL 1	Clinical			01/02/1998	Ticino	Switzer-land	1/Phil.	ERR376626
	EUL 3	Clinical			01/10/1989	Ticino	Switzer-land	1/Phil.	ERR376628
	EUL 9	Environmental			01/10/1989	S. Gallen	Switzer-land	1/Phil.	ERR376634
	EUL 10	Environmental			01/10/1989	S. Gallen	Switzer-land	1/Phil.	ERR376635
	H034800423	Environmental			01/11/2003	Hereford	UK	1/Oxford/ OLDA	Reuter *et al.* 2013
	H072740379	Environmental (domestic header tank)			28/06/2007	Woking-ham, Berkshire	UK	1/Phil.	ERR1399548
	H084800579	Clinical			27/11/2008	East of England	UK	1/Oxford/ OLDA	ERR1441923
	H085060063	Environmental (from home of patient from which above isolate, H084800579, was obtained)			About 11/12/2008	Chelmsford, Essex	UK	1/Oxford/ OLDA	ERR1441924
	H091640624 (case B)	Clinical			20/04/2009	Chelmsford, Essex	UK	1/Oxford/ OLDA	ERR1441928
	H091720529	Environmental (from home of patient from which above isolate, H091640624, was obtained)			17/04/2009	East of England	UK	1/Oxford/ OLDA	ERR1441926
	H100200319	Environmental (home of case 3)			30/12/2009	Chadwell, Essex	UK	1/Oxford/ OLDA	ERR1399534

	H100200320	Environmental (home of case 3)			30/12/2009	Chadwell, Essex	UK	1/Oxford/ OLDA	ERR1399538
	H100200321	Environmental (home of case 3)			30/12/2009	Chadwell, Essex	UK	1/NA (mAb all negative)	ERR1399541
	H115260949	Environmental (home of case A who also spent part of their incubation period in Hospital F)			26/12/2011	London	UK	1/Phil.	ERR1441935
	H152640286 (case C)	Clinical			22/06/2015	East of England	UK	1/Oxford/ OLDA	ERR1441925
	H152780272	Environmental (from home of patient from which above isolate, H152780272, was obtained)			01/07/2015	East of England	UK	1/Oxford/ OLDA	ERR1441932
Isolates from a cruise ship (n=3)									
	H073300077	Environmental			Approx. 8/8/2007	NA	NA	1/Oxford/ OLDA	ERR1399519
	H073300079	Environmental			Approx. 8/8/2007	NA	NA	1/Oxford/ OLDA	ERR1399496
	H073360657	Environmental			Approx. 5/8/2007	NA	NA	1/Oxford/ OLDA	ERR1399557
Isolates with an unknown sampling context (n=38)									
	L 3386/03	Environmental			2003	U/k	Austria	1/U/k	ERR922502
	L 3415/03	Environmental			2003	U/k	Austria	1/U/k	ERR922503
	LT 40/04	Clinical			2004	U/k	Austria	1/U/k	ERR922499
	Wien 47-14	Environmental			1996	U/k	Austria	1/U/k	ERR923397
	EUL 90	Clinical			U/k	U/k	Denmark	1/Oxford/	ERR376736

								OLDA	
	E21203	Clinical			2004	U/k	France	1/U/k	ERR923395
	HL 0701 3004	Environmental			03/01/2007	Rueil Malmaison	France	1/Oxford/ OLDA	ERR922488
	LG 0919 2006	Clinical			23/04/2009	Saint Nazaire	France	1/Phil.	ERR922489
	LG 1105 4025	Environmental			19/01/2011	U/k	France	1/Phil.	ERR922494
	EUL 110	Clinical			01/01/1993	Luebeck	Germany	10/Oxford /OLDA	ERR376674
	EUL 113	Environmental			27/02/1995	Hannover	Germany	7/Oxford/ OLDA	ERR363968
	EUL 114	Environmental			27/02/1995	Hannover	Germany	7/Oxford/ OLDA	ERR363969
	EUL 117	Clinical			01/06/2005	U/k	Germany	6/ Benidorm	ERR376755
	EUL 119	Clinical			01/06/2005	U/k	Germany	1/Oxford/ OLDA	ERR376757
	EUL 60	Clinical			01/01/1992	U/k	Greece	1/Phil.	ERR376685
	EUL 62	Environmental			01/01/1989	U/k	Greece	1/Oxford/ OLDA	ERR376687
	EUL 67	Clinical			01/01/1995	U/k	Greece	1/Oxford/ OLDA	ERR376692
	EUL 37	Clinical			01/01/1999	U/k	Italy	1/Phil.	ERR376723
	EUL 42	Clinical			01/01/1999	U/k	Italy	1/Phil.	ERR376667
	EUL 43	Clinical			01/01/1999	U/k	Italy	1/Phil.	ERR376668
	EUL 44	Environmental			01/01/1999	U/k	Italy	1/Phil.	ERR376669
	EUL 45	Clinical			01/01/1999	U/k	Italy	72/Phil.	ERR376670
	EUL 46	Environmental			01/01/1999	U/k	Italy	1/Oxford/ OLDA	ERR376671
	NIIB223	Environmental			1986	U/k	Japan	1/U/k	ERR922500
	NIIB225	Environmental			1986	U/k	Japan	1/U/k	ERR922501
	LG 1118 1044	Environmental			11/07/2009	U/k	Morocco	1/Oxford/ OLDA	ERR922495

	ATCC 35289	Environmental			1988	U/k	Nether-lands	390/NA (sg9)	ERR923391
	EUL 109	Environmental			01/01/1992	U/k	Sweden	1/Oxford/ OLDA	ERR376662
	LP21_Sweden	Clinical			1996-1999	U/k	Sweden	1/U/k	ERR922497
	LP22_Sweden	Clinical			1996-1999	U/k	Sweden	1/U/k	ERR923393
	LP23_Sweden	Clinical			1996-2000	U/k	Sweden	1/U/k	ERR922498
	EUL 13	Clinical			01/01/1994	U/k	UK	1/ Benidorm	ERR376646
	EUL 14	Clinical			06/06/1984	Glasgow	UK	5/ Benidorm	ERR376639
	EUL 17	Clinical			01/01/1993	Ayrshire	UK	7/Phil.	ERR376642
	EUL 21	Environmental			01/01/1999	Glasgow	UK	1/Phil.	ERR376638
	H103620682	Environmental			20/07/2010	Near London	UK	1/Oxford/ OLDA	ERR1441922
	2735	Environmental			2002	U/k	USA	1/U/k	ERR923396
	OLDA1 (NCTC12008)	Clinical			1947	Washington	USA	1/Oxford/ OLDA	ERR434061