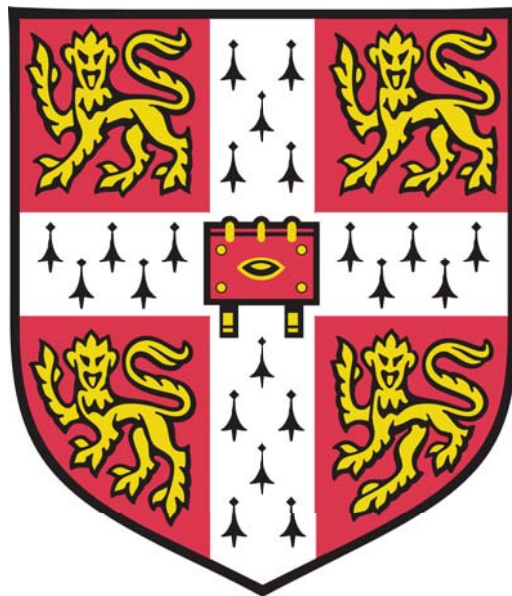


# The diversity of disease-causing and environmental *Legionella pneumophila*

**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

<b>1. INTRODUCTION</b> .....	<b>1</b>
<b>1.1. THE HISTORY AND CLASSIFICATION OF <i>L. PNEUMOPHILA</i></b> .....	<b>1</b>
1.1.1. The first recognised outbreak caused by <i>L. pneumophila</i> .....	1
1.1.2. Earlier isolation of <i>L. pneumophila</i> .....	1
1.1.3. <i>L. pneumophila</i> classification .....	2
1.1.4. Microbiological characteristics of <i>L. pneumophila</i> .....	3
<b>1.2. THE LIFE CYCLE AND PATHOGENESIS OF <i>L. PNEUMOPHILA</i></b> .....	<b>4</b>
1.2.1. Protozoa: the natural hosts of <i>L. pneumophila</i> .....	4
1.2.2. The “accidental” infection of humans.....	5
1.2.3. The intracellular life cycle of <i>L. pneumophila</i> .....	5
1.2.4. The Dot/Icm secretion system.....	7
<b>1.3. DISEASE CAUSED BY <i>L. PNEUMOPHILA</i></b> .....	<b>9</b>
1.3.1. Legionnaires’ disease.....	9
1.3.2. Pontiac fever.....	10
1.3.3. Extra-pulmonary disease.....	11
<b>1.4. MICROBIOLOGICAL IDENTIFICATION AND DETECTION</b> .....	<b>12</b>
1.4.1. Culture methods.....	12
1.4.2. Serologic diagnosis.....	13
1.4.3. Direct fluorescent antibody testing .....	14
1.4.4. Urine antigen detection.....	14
1.4.5. PCR-based detection.....	15
<b>1.5. TYPING METHODS AND OUTBREAK INVESTIGATIONS</b> .....	<b>15</b>
1.5.1. Pulsed field gel electrophoresis .....	15
1.5.2. Amplified fragment length polymorphism.....	16
1.5.3. Monoclonal antibody subgrouping .....	16
1.5.4. Sequence-based typing.....	16
<b>1.6. EPIDEMIOLOGY OF LEGIONNAIRES’ DISEASE</b> .....	<b>17</b>
1.6.1. The incidence of Legionnaires’ disease .....	17
1.6.2. Sporadic cases, clusters and outbreaks .....	19
1.6.3. Common sources of infection .....	20
1.6.4. Transmission.....	20
1.6.5. Host risk factors .....	21

1.6.6. Travel-associated Legionnaires' disease .....	21
1.6.7. The distribution of <i>L. pneumophila</i> subtypes in clinical disease.....	23
<b>1.7. THE ENVIRONMENTAL DISTRIBUTION OF <i>L. PNEUMOPHILA</i> .....</b>	<b>24</b>
1.7.1. <i>L. pneumophila</i> in the natural environment.....	24
1.7.2. The colonisation of man-made water systems by <i>L. pneumophila</i> .....	25
1.7.3. The control of <i>L. pneumophila</i> in man-made water systems.....	26
<b>1.8. WHOLE GENOME SEQUENCING TECHNOLOGIES.....</b>	<b>27</b>
1.8.1. The history of sequencing .....	27
1.8.2. Second generation sequencing technologies.....	28
1.8.3. Third generation sequencing technologies .....	30
1.8.4. Bioinformatic advances.....	30
1.8.5. Applications of bacterial WGS.....	32
<b>1.9. APPLICATION OF WGS TO <i>L. PNEUMOPHILA</i> .....</b>	<b>33</b>
1.9.1. The structure and features of the <i>L. pneumophila</i> genome.....	33
1.9.2. The population structure, diversity and evolution of <i>L. pneumophila</i> .....	34
1.9.3. WGS in outbreak investigations.....	36
<b>1.10. THESIS OUTLINE .....</b>	<b>37</b>
<b>2. MATERIALS &amp; METHODS.....</b>	<b>39</b>
<b>2.1. CULTURE AND DNA EXTRACTION .....</b>	<b>39</b>
<b>2.2. WHOLE GENOME SEQUENCING .....</b>	<b>39</b>
<b>2.3. DE NOVO ASSEMBLY OF ILLUMINA SEQUENCE DATA.....</b>	<b>40</b>
<b>2.4. CONTROL FOR SAMPLE MIX-UP THROUGH DETERMINATION OF SEQUENCE TYPE .....</b>	<b>40</b>
<b>2.5. MAPPING OF ILLUMINA SEQUENCE DATA .....</b>	<b>40</b>
<b>2.6. PHYLOGENETIC ANALYSIS .....</b>	<b>41</b>
<b>2.7. STATISTICAL ANALYSES AND FIGURES .....</b>	<b>42</b>
<b>3. RECENT EMERGENCE OF FIVE MAJOR DISEASE-ASSOCIATED STS.....</b>	<b>43</b>
<b>3.1. INTRODUCTION .....</b>	<b>44</b>
<b>3.2. MATERIALS &amp; METHODS.....</b>	<b>46</b>
3.2.1. Bacterial isolates .....	46
3.2.2. Whole genome sequencing .....	47
3.2.3. Mapping of sequence reads and phylogenetic analysis .....	47
3.2.4. Time-dependent phylogenetic reconstruction .....	48
3.2.5. Estimation of the age of the ST1, ST23, ST47 and ST62 lineages .....	48
3.2.6. Gene content analysis.....	49

3.2.7. Searching for evidence of positive selection using CodeML.....	49
3.2.8. Identification of genes with high nucleotide similarity in the five STs.....	50
3.2.9. Identification of recombination donors.....	50
<b>3.3. RESULTS .....</b>	<b>51</b>
3.3.1. Independent emergence of the five STs .....	51
3.3.2. Investigation of the diversity within the five STs.....	52
3.3.3. Dating the emergence of the five lineages.....	56
3.3.4. Analysis of the spread of the disease-associated STs.....	59
3.3.5. Evidence for convergent evolution.....	66
<b>3.4. DISCUSSION .....</b>	<b>76</b>
<b>4. DYNAMICS AND IMPACT OF HOMOLOGOUS RECOMBINATION ON THE EVOLUTION OF <i>LEGIONELLA PNEUMOPHILA</i> .....</b>	<b>82</b>
<b>4.1. INTRODUCTION .....</b>	<b>83</b>
<b>4.2. MATERIALS &amp; METHODS.....</b>	<b>85</b>
4.2.1. Bacterial isolates.....	85
4.2.2. Reference genomes.....	86
4.2.3. Mapping, recombination detection, phylogenetic analysis and BAPS clustering .....	86
4.2.4. Detection of MGEs .....	87
4.2.5. Identification of homologous recombination hotspots.....	88
4.2.6. Inference of recombination donors .....	88
<b>4.3. RESULTS .....</b>	<b>88</b>
4.3.1. Contribution of homologous recombination to <i>L. pneumophila</i> diversity.....	88
4.3.2. Hotspots of homologous recombination in <i>L. pneumophila</i> .....	93
4.3.3. Inference of recombination donors .....	101
<b>4.4. DISCUSSION .....</b>	<b>109</b>
<b>5. EVALUATION OF AN OPTIMAL WGS-BASED TYPING SCHEME FOR <i>LEGIONELLA PNEUMOPHILA</i> .....</b>	<b>115</b>
<b>5.1. INTRODUCTION .....</b>	<b>116</b>
<b>5.2. MATERIALS &amp; METHODS.....</b>	<b>118</b>
5.2.1. Bacterial isolates.....	118
5.2.2. Study design.....	119
5.2.3. <i>De novo</i> assembly .....	119
5.2.4. Mapping/SNP-based analysis .....	120



5.2.5. Extended MLST .....	120
5.2.6. Gene presence/absence profiling.....	122
5.2.7. Kmer-based analysis.....	123
<b>5.3. RESULTS .....</b>	<b>123</b>
5.3.1. Typability .....	124
5.3.2. Reproducibility .....	127
5.3.3. Epidemiological concordance .....	129
5.3.4. Discriminatory power .....	139
5.3.5. Stability .....	143
<b>5.4. DISCUSSION .....</b>	<b>144</b>
<b>6. APPLICATION OF WGS TO NOSOCOMIAL INVESTIGATIONS OF LEGIONNAIRES’ DISEASE.....</b>	<b>149</b>
<b>6.1. INTRODUCTION .....</b>	<b>150</b>
<b>6.2. MATERIALS AND METHODS.....</b>	<b>151</b>
6.2.1. Bacterial isolates .....	151
6.2.2. Whole genome sequencing .....	152
6.2.3. Mapping of sequence reads and phylogenetic analysis .....	152
<b>6.3. RESULTS .....</b>	<b>153</b>
6.3.1. Hospital lineages comprise distinct lineages of <i>L. pneumophila</i> ST1.....	153
6.3.2. WGS can be used to support or refute links between Legionnaires’ disease cases and hospital water systems .....	156
6.3.3. Substantial diversity within single hospital populations .....	165
6.3.4. Evidence for local microevolution within hospital populations.....	166
6.3.5. Long-term stability of hospital strains.....	168
6.3.6. Evidence for hospital seeding via local and international spread of ST1 .....	168
<b>6.4. DISCUSSION .....</b>	<b>169</b>
<b>7. CONCLUSIONS AND FUTURE DIRECTIONS.....</b>	<b>173</b>
<b>7.1. A RESTATEMENT OF THE RESEARCH QUESTIONS AND AIMS .....</b>	<b>173</b>
<b>7.2. KEY FINDINGS AND FUTURE DIRECTIONS.....</b>	<b>174</b>
7.2.1. Five major disease-associated STs have emerged recently and spread rapidly .....	174
7.2.2. Homologous recombination is a major driver of <i>L. pneumophila</i> evolution..	175
7.2.3. A 50-gene cgMLST scheme is suggested as the optimal WGS-based method for <i>L. pneumophila</i> typing.....	176

7.2.4. WGS can be used to successfully confirm or refute links between Legionnaires' disease cases and hospitals.....	177
<b>7.3. CLOSING REMARKS.....</b>	<b>178</b>
<b>8. REFERENCES.....</b>	<b>179</b>
<b>9. APPENDIX.....</b>	<b>203</b>
9.1. CHAPTER 3.....	203
9.2. CHAPTER 4.....	215
9.3. CHAPTER 5.....	241
9.4. CHAPTER 6.....	350

## List of Figures

Figure 1.1. Electron microscope image of *L. pneumophila*

Figure 1.2. The *Legionella*-containing vacuole

Figure 1.3. Infection of *Acanthamoeba castellanii* with *L. pneumophila*

Figure 1.4. Dot/Icm machinery

Figure 1.5. Incidence of Legionnaires' disease

Figure 1.6. Seasonal trend of Legionnaires' disease

Figure 1.7. Distribution of Legionnaires' disease cases by age and sex

Figure 1.8. Incidence of travel-associated Legionnaires' disease

Figure 1.9. Distribution of STs amongst clinical and environmental isolates

Figure 1.10. Population structure of *L. pneumophila*

Figure 3.1. Geographical distribution of STs and their prevalence in clinical and environmental samples

Figure 3.2. Population structure of *L. pneumophila* highlighting five major disease-associated STs of interest

Figure 3.3. Distribution of SNPs in isolates belonging to STs 47 (A), 1 (B), 23 (C), 37 (D) and 62 (E)

Figure 3.4. Linear regression analyses of root-to-tip distances against sampling date in each of the five STs

Figure 3.5. Time-dependent phylogenetic reconstruction of the ST37 lineage inferred using a Bayesian coalescent model in BEAST

Figure 3.6. Maximum likelihood trees of the ST1 (A), ST23 (B), ST37 (C), ST62 (D) and ST47 (E) lineages.

Figure 3.7. Nucleotide diversity of the five STs across the genome

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

Figure 3.9. Tanglegrams comprising maximum likelihood trees of 32 STs of *L. pneumophila* that are representative of the known species diversity

Figure 4.1. Generation of diversity in the six major disease-associated STs

Figure 4.2. Relative frequency of homologous recombination events and vertically inherited mutations

Figure 4.3. Types of change introduced by vertically inherited mutations and homologous recombination

Figure 4.4. Size of detected homologous recombination regions in the six STs

Figure 4.5. Homologous recombination events detected in the ST1 lineage

Figure 4.6. Hotspot 6 in the ST1 lineage

Figure 4.7. The LPS locus comprising hotspot 3 in the ST1 lineage

Figure 4.8. Maximum likelihood tree of 536 *L. pneumophila* isolates that are coloured by BAPS cluster

Figure 4.9. Similarity of the recombined regions to the predicted donors

Figure 4.10. Predicted recombination donor clusters

Figure 4.11. Diversity of recombination donors across the genome in the ST1 lineage

Figure 4.12. Sequence similarity between donors and recipients

Figure 4.13. Maximum likelihood tree of 81 ST1 isolates with predicted recombination events mapped onto the branches

Figure 5.1. Pairwise differences between typing panel isolates using different WGS-based methods (A-H)

Figure 5.2. Index of discrimination ( $D$ ) and epidemiological concordance ( $E$ ) of the current and WGS-based methods

Figure 5.3. Neighbour-net tree of the typing panel isolates constructed using the 100-gene cgMLST scheme

Figure 5.4. Pairwise SNP differences between epidemiologically “unrelated” and “related” isolates belonging to some of the major disease-associated STs (A-E)

Figure 5.5. Maximum likelihood tree of 74 ST37 isolates with isolates coloured by their epidemiological relatedness

Figure 6.1. Maximum likelihood tree of 229 ST1 and “ST1-derived” isolates including those from or associated with hospitals

Figure 6.2. Time frame of legionellosis incidents and collection of environmental isolates at Hospital A

Figure 6.3. Phylogeny of isolates from Hospital A and the surrounding area

Figure 6.4. Zoomed-in sections of the maximum likelihood tree presented in Figure 6.1

Figure 6.5. A plan of Hospital A

## List of Tables

Table 1.1. Classification of *L. pneumophila*

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

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

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

Table 3.3. Homoplastic SNPs on three or four of the branches leading to STs 1, 23, 37, 47 and 62

Table 3.4. Highly similar genes in the five STs

Table 3.5. Recombination events that occurred on the branches leading to STs 47 and 37 and their predicted origin

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

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

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

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

Table 5.1. Typability of the WGS-based methods

Table 5.2. Number of differences identified between sequencing replicates using each of the WGS-based methods

Table 5.3. Index of discrimination ( $D$ ) and epidemiological concordance ( $E$ ) of the current and tested WGS-based typing methods

Table 5.4. Number of differences identified between isolates from epidemiologically “related” sets using each of the WGS-based methods

Table 5.5. Differentiation between isolates from major disease-associated STs

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

## 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