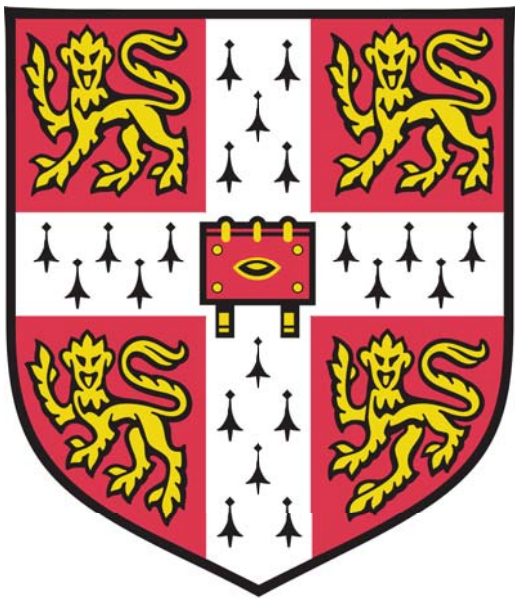
The diversity of disease-causing and environmental *Legionella pneumophila*

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