Developing Metagenomic Methods for *Legionella* Detection

Sharon Marie Carney

National Heart & Lung Institute,
Imperial College London
&
Public Health England

Submitted for the degree of Doctorate of Philosophy

Abstract

Legionella is a Gram-negative bacterium naturally present in freshwater and soil. The bacteria can enter, colonise and multiply in man-made water systems. Infection through inhalation of aerosols containing the bacteria can cause Legionnaires' disease (LD) an atypical, severe pneumonia in individuals with underlying risk factors. *Legionella pneumophila* serogroup 1, the most widely studied species, is reported to account for more than 90 % of all clinical isolates related to LD in England. *L. pneumophila* is isolated routinely at Public Health England however *Legionella* is a slow growing bacterium, typically taking from three to five days to grow. Additionally, it has been reported that *L. pneumophila* is isolated in only 60 % of urinary antigen-positive cases.

Here, the application of metagenomic sequencing was investigated in *Legionella* positive clinical and environmental samples, the hypothesis being that metagenomic sequencing may provide a more time efficient result and may reveal previously undetected heterogeneity in clinical and environmental cases.

The results demonstrate that *L. pneumophila* genomes can be captured and sequenced from patients with LD and from environmental source samples without prior culture using a targeted capture approach. The data generated also demonstrate that *Legionella* diversity within environmental sources as well as a clinical case could be captured. Importantly, the work has demonstrated the first successful application of *in silico* 7-loci sequence-based typing and 50 core gene MLST to *Legionella* data generated by a metagenomic method.

Overall, this thesis demonstrates the proof of concept of targeted metagenomic sequencing of *L. pneumophila* directly from multiple patients and environmental sources as well as the ability to capture a variety of sequence types. Furthermore, the challenges of implementing metagenomic sequencing for routine diagnostic use and future avenues for technical optimisation of the targeted capture approach are outlined.

Acknowledgements

I would like to express my sincere gratitude to my supervisors Prof. Miriam Moffatt, Prof. William Cookson and Dr. Victoria Chalker for their expertise, advice and encouragement. I am very grateful to you for giving me the opportunity to work on this project.

I am especially grateful to my associate supervisor Dr. Michael Cox for his knowledge, helpfulness and time and being inspiring and kind in the difficult moments.

My thanks also go to Prof. Michael Lovett for his technical guidance during the work for Chapter 4.

I would like to thank the members of the Genomic Medicine Section, NHLI, especially Verdiana, Claire and Huw for their friendship and laughs and Leah, Colin, Anca and Kenny for their help with problem-solving and technical advice. I also wish to acknowledge the help and guidance of members of the Respiratory and Vaccine-Preventable Bacteria Unit at Public Health England.

Many thanks to my friends Bill and Sonal and my cousin Lorraine who supported and looked out for me when I moved back to London. Thanks to my partner Bartosz for being understanding, thoughtful and supportive and providing a continuous flow of memes.

Finally, I would especially like to thank my parents Bridget and Fintan and my sisters Elaine and Niamh who constantly support and encourage me.

Declaration of Originality

I, Sharon Carney declare I wrote this thesis and performed the work therein. Work performed by others is referenced appropriately.

Copyright Declaration

Table of Contents

List of Figures

List of Tables

Abbreviations

AFLP	Amplified fragment length polymorphism
asd	aspartate-B-semialdehyde dehydrogenase gene
BAL	Bronchial alveolar lavage fluid
BCYE	Buffered charcoal yeast extract
BRC	Biomedical Research Centre
bp	Base pair
CDC	Centre for Disease Control
CO_2	Carbon dioxide
COG	Cluster of Orthologous Groups of Proteins
Cot	Concentration over time
COPD	Chronic Obstructive Pulmonary Disorder
CTAB	Hexadecytrimethylammonium bromide
dATP	Deoxyadenosine triphosphate
DEPC	Diethyl pyrocarbonate
DFA	Direct fluorescent antibody assays
DH	Department of Health
DNA	Deoxyribonucleic acid
dCTP	Deoxycytidine triphosphate
dGTP	Deoxyguanosine triphosphate
dsDNA	Double-stranded DNA
DSMZ	Deutsche Sammlung von Mikroorganismen und Zellulturen GmbH
dTTP	Deoxythymidine triphosphate
dUTP	Deoxyuridine triphosphate
ECDC	European Centre for Disease Prevention and Control
EDTA	Ethylenediaminetetraacetate
EEA	European Economic Area
EIA	Enzyme Immunoassay
ESGLI	European Study Group for *Legionella* Infection
ESR	Institute of Environmental Science and Research Ltd
EtOH	Ethanol
EU	European Union
FFPE	Formalin-fixed paraffin embedded
flaA	Flagellin gene
FluC	Firefly luciferase
ftp	File transfer protocol
g	Gram
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
H_2O	Water
HPS	Health Protection Scotland
ICT	Immunochromatographic test
IMT	Incident Management Team
IRAS	Integrated Research Application System
ISO	Isolate
kbp	Kilobase pairs
L	Litre
LCV	*Legionella*-containing vacuole
LD	Legionnaires' Disease

LINEs	Long interspersed nuclear element
LLAPs	*Legionella*-like amoebal pathogens
LME	Lysing matrix E tube
Lp1	*Legionella pneumophila* serogroup 1
LPS	Lipopolysaccharide Antigen (LPS)
LTR	Long terminal repeat
mAb	Monoclonal antibody
M	Molar
mip	Macrophage infectivity potentiator gene
ml	Mililitre
MLST	Multi-locus sequence-based typing
mompS	Major outer membrane precursor protein gene
NIH	National Institute of Health
Na_2HPO_4	Dibasic diphosphate
NaCl	Sodium chloride
NaOH	Sodium hydroxide
NEB	New England Biolabs
NETs	Neutrophil extracellular traps
neuA	Lipopolysaccharide biosynthesis gene
neuAh	Lipopolysaccharide biosynthesis homolog gene
ng	Nanogram
nm	Nanometre
nM	Nanomolar
NNDS	National Notifiable Diseases Surveillance System
ONT	Oxford Nanopore Technologies
ORF	Open Reading Frame
OUT	Operational Taxonomic Unit
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
PFGE	Pulsed field gel electrophoresis
PHE	Public Health England
pilE	Type IV pilin gene
pM	Picomolar
PMA	Propium monoazide
proA	Zinc metalloprotease gene
qPCR	Quantitative Polymerase Chain Reaction
REC	Research Ethics Committee
RER	Rough endoplasmic reticulum
RNA	Ribonucleic acid
rpm	Revolutions per minute
rRNA	Ribosomal ribonucleic acid
RSB	Resuspension buffer
RVPBRU	Respiratory and Vaccine-Preventable Bacteria Unit
SBT	Sequence-based typing
SDS	Sodium dodecyl sulphate
Sg	Serogroup
SINEs	Short interspersed nuclear element
SNP	Single nucleotide polymorphism
SSC	Saline sodium citrate

SSPE	Sodium chloride-sodium phosphate-EDTA
ST	Sequence Type
STEC	Shiga Toxigenic *Escherichia coli*
TB	Tuberculosis
TBE	Tris Borate EDTA
TC	Target Capture
TE	Tris-EDTA
UAT	Urinary Antigen Test
UCL	University College London
UKAS	UK's National Accreditation Body
µl	Microlitre
µM	Micromolar
UTP	Uridine triphosphate
UV	Ultraviolet
V	Volts
VCF	Variant Call Format
WG	Whole genome
WGA	Whole genome amplification
WGS	Whole genome sequencing
WHO	World Health Organisation
x g	Relative centrifugal force
ZMW	Zero-mode waveguide
°C	Degrees Celsius

Chapter 1.

Introduction

1.1 The Emergence and Identification of *Legionella*

1.1.1 Discovery of *Legionella pneumophila*

In July 1976, the Centre for Disease Control (CDC) was alerted to an outbreak of respiratory illness among attendees of the Pennsylvania American Legion convention hosted by the Bellevue-Stratford Hotel in Philadelphia, USA. A total of 182 members of the American Legion developed severe respiratory illness with 29 individuals dying after returning home from the conference (Fraser *et al.,* 1977). Other suspected cases began to emerge including individuals residing within one block of the hotel and nearby pedestrians. An etiologic agent for the illness, which became known as Legionnaires' Disease (LD), was not identified until 6 months later. First evidence of the causative agent was recognised when guinea pigs were inoculated with patient lung tissue and a Gram-negative bacillus was isolated (McDade *et al.,* 1979). Subsequent egg-yolk sac cultures of the bacillus were used as antigen to examine suspected patient serum specimens by indirect immunofluorescence. Increases in antibody titre indicated that the newly isolated bacterium, termed *Legionella pneumophila*, was the causative agent of the outbreak (McDade *et al.,* 1979).

1.1.2 Retrospective Identification of Legionnaires' Disease Cases and Outbreaks

Subsequent analysis of unsolved cases of respiratory illness from prior decades revealed that *L. pneumophila* had in fact been isolated from a pneumonia patient in 1947 with an identification label of a "rickettsia-like agent" (McDade *et al.,* 1979). In a 1944 study, *Legionella* was recovered from guinea pigs infected with clinical material (Tatlock *et al.,* 1944). Furthermore, in a retrospective study of unsolved outbreaks of pneumonia, *L. pneumophila* was found to be the causative agent of a number of the outbreaks; including an outbreak in a meat-packing factory in 1957 (Osterholm *et al.,* 1983), an outbreak among attendees of a convention taking place in a hotel in Philadelphia (with close proximity to the Bellevue-Stratford) in 1974 (Terranova *et al.,* 1978) and a hospital outbreak in 1965 (Thacker *et al.,* 1978). Interestingly, an outbreak of non-pneumonic respiratory illness was observed in Pontiac, Michigan in 1968 with no fatalities (Glick *et al.,* 1978). The causative agent was identified ten years later (in 1978) as *L. pneumophila* despite the milder clinical symptoms and a reduced incubation period of 36 hours. These studies led to the recognition that *L. pneumophila* could cause two clinically distinct syndromes and the milder form was therefore termed “Pontiac Fever” (Glick *et al.,* 1978). Together, clinical syndromes caused by *Legionella* infection are referred to as “legionellosis”.

1.1.3 Identification of *Legionella* in the Environment

The first environmental identification of *L. pneumophila* was from non-outbreak-related freshwater by direct immunofluorescence staining (Fliermans *et al.,* 1979). A larger study of freshwater sources demonstrated the ubiquity of *L. pneumophila* in this habitat (Fliermans *et al.,* 1981). Airborne transmission of the bacterium was first hypothesised based on epidemiological evidence from a retrospectively identified 1965 LD outbreak (Thacker *et al.,* 1978). Furthermore, the Pontiac outbreak (Glick *et al.,* 1978) provided evidence of transmission from water via an air-handling device. In a critical review by Muder *et al.,* 1986, a number of modes of transmission based on water aerosolisation and aspiration were hypothesised based on epidemiological evidence from previous cases and outbreaks. These modes included transmission by cooling towers and evaporative condensers, showers, humidifiers as well as respiratory therapeutic devices. The source of bacterial dispersal during the 1976 Philadelphia outbreak was later identified as the hotel air-conditioning system (Fraser *et al.,* 1977).

Soon after the Philadelphia outbreak, other *Legionella* species began to be recognised. They were isolated in the previous decades during research on rickettsial disease. These species were retrospectively identified as *L. bozemanii* and *L. micdadei* (Hebert *et al.,* 1980).

1.2 Current Taxonomic Classification and Species

After the discovery of *L. pneumophila*, a new genus “Legionella” was proposed (Brenner *et al.,* 1979). The full scientific classification was designated as follows: Domain: Bacteria, Phylum: Proteobacteria, Class: Gammaproteobacteria, Order: Legionellales, Family: Legionellaceae, Genus: *Legionella* (Brenner *et al.,* 1979). In the years since, 65 species of the *Legionella* genus have been identified in the environment and/or clinically. In addition to species, four *L. pneumophila* subspecies have been described: *pneumophila*, *fraseri*, *pasculeii* (Brenner *et al.,* 1988) all of which were recognised by DNA hybridisation experiments and the fourth recently described subspecies *raphaeli* (Kozak-Muiznieks *et al.,* 2018). *L. pneumophila* is divided into 16 serogroups, all of which have a clinical association. *L. pneumophila* serogroup 1 is reported to account for 85.6 % of culture positive cases in Europe (ECDC, 2017). In some countries *L. longbeachae* has an equal culture positivity rate to *L. pneumophila* in clinical cases, with different geographic distributions within those countries. In 2015, 47 % of LD clinical isolates in Australia were *L. longbeachae* (NNDS, 2019), 65 % in New Zealand in 2017 (ESR, 2017) and 50 % of clinical isolates in Thailand (Phares *et al.,* 2007). Of the total 65 *Legionella* species identified to date, 31 have been implicated in clinical disease (**Table 1.1**).

Table 1.1 *Legionella* Species Associated or Not Associated with Clinical Disease

Associated with Clinical Disease	**No Known Clinical Association**
L. anisa (Gorman *et al.,* 1985, Borstein *et al.,* 1989)	*L. adelaidensis* (Benson *et al.,* 1991)
L. birminghamensis (Wilkinson *et al.,* 1987)	*L. beliardensis* (Lo Presti *et al.,* 2001)
L. bozemanii (Brenner *et al.,* 1980, Mitchell *et al*., 1984, Tang *et al.,* 1984)	*L. brunensis* (Wilkinson *et al.,* 1988) *L. busanensis* (Park *et al.,* 2003)
L. cardiaca (Pearce *et al.,* 2012)	*L. cherrii* (Brenner *et al.,* 1985)
L. cincinnatiensis (Thacker *et al.,* 1988)	*L. drancourtii* (La Scola *et al.,* 2004)
L. clemsonensis (Palmer *et al*., 2016)	*L. dresdenensis* (Lück *et al.,* 2010)
L. donaldsonii (Hookey *et al.,* 1996, Han *et al.,* 2015)	*L. drozanskii* (Adeleke *et al.,* 2001)
L. dumoffii (Brenner *et al.,* 1980, Joly *et al.,* 1986)	*L. fairfieldensis* (Thacker *et al.,* 1991)
L. erythra (Brenner *et al.,* 1985, von Baum *et al.,* 2008)	*L. fallonii* (Adeleke *et al.,* 2001)
L. feeleii (Herwaldt *et al.,* 1984, Thacker *et al.,* 1985, Patluke *et al.,* 1986)	*L. geestiana* (Dennis *et al.,* 1993) *L. gemonospecies* (Benson *et al.,* 1996)
L. gormanii (Morris *et al.,* 1980, Griffith *et al.,* 1988)	*L. gratiana* (Bornstein *et al.,* 1989)
L. hackeliae (Brenner *et al.,* 1985, Wilkinson *et al.,* 1985)	*L. gresilensis* (Lo Presti *et al.,* 2001)
L. jamestowniensis (Brenner *et al.,* 1985, Edelstein *et al.,* 2017)	*L. impletisoli* (Kuroki *et al.,* 2007)
L. jordanis (Cherry *et al.,* 1982, Littrup *et al.,* 1987)	*L. israelensis* (Bercovier *et al.,* 1986)
L. lansingensis (Thacker *et al.,* 1992)	Candidatus *L. jeonii* (Park *et al.,* 2004)
L. londiniensis (Dennis *et al.,* 1993, Stallworth *et al.,* 2012)	*L. massiliensis* (Campocasso *et al.,* 2012)
L. longbeachae (McKinney *et al.,* 1981, Bibb *et al.,* 1981)	*L. moravica* (Wilkinson *et al.,* 1988)
L. lytica (Hookey *et al.,* 1996, Han *et al.,* 2015)	*L. nautarum* (Dennis *et al.,* 1993)
L. maceachernii (Brenner *et al.,* 1985, Wilkinson *et al.,* 1985)	*L. norrlandica* (Rizzardi *et al.,* 2015)
L. micdadei (Myerowitz *et al.,* 1979, Hébert *et al.,* 1980)	*L. quateirensis* (Dennis *et al.,* 1993)
L. nagasakiensis (Yang *et al.,* 2012)	*L. quinlivanii* (Benson *et al.,* 1989)
L. oakridgensis (Orrison *et al.,* 1983, Tang *et al.,* 1985)	*L. rowbothamii* (Adeleke *et al.,* 2001)
L. parisiensis (Brenner *et al.,* 1985, Lo Presti *et al.,* 1997)	*L. santicrucis* (Brenner *et al.,* 1985)
L. pittsburghensis (Pasculle *et al.,* 1980)	*L. saoudiensis* (Bajrai *et al.,* 2016)
L. pneumophila (Brenner *et al.,* 1979)	*L. shakespearei* (Verma *et al.,* 1992)
L rubrilucens (Brenner *et al.,* 1985, Matsui *et al.,* 2010)	*L. spiritensis* (Brenner *et al.,* 1985)
L. sainthelensi (Campbell *et al.,* 1984, Chereshsky *et al*., 1986, Benson *et al.,* 1990)	*L. steigerwaltii* (Brenner *et al.,* 1985) *L. taurinensis* (Lo Presti *et al.,* 1999)
L. steelei (Edelstein *et al.,* 2012)	*L. thermalis* (Ishizaki *et al.,* 2016)
L. tucsonensis (Thacker *et al.,* 1989)	*L. tunisiensis* (Campocasso *et al.,* 2012)
L. wadsworthii (Edelstein *et al.,* 1982)	*L. worsleiensis* (Dennis *et al.,* 1993)
L. waltersii (Konig *et al.,* 2005, Benson *et al.,* 1996)	*L. yabuuchiae* (Kuroki *et al.,* 2007)

1.3 *Legionella* Physiology and Ecology

Legionellae are aerobic, Gram-negative, Gammaproteobacteria. They exist as rod-shaped bacilli or long filamentous cells that do not form spores. *Legionella* species are ubiquitous freshwater inhabitants, present in rivers, lakes, ponds, streams, sub-surface waters and hot springs (Fliersmans *et al.,* 1981, Qin *et al.,* 2013). They are also present in soil, sediments, potting and compost mixes, from which *L. longbeachae* and *L. pneumophila* species are predominantly associated with infection (Steele *et al.,* 1990, Casati *et al.,* 2009, Pravinkumar *et al.,* 2010, Currie *et al.,* 2014, Travis *et al.,* 2014). They survive in water at temperatures of between 20 degrees Celsius (°C) and 45 °C although they have been reported in water above and below this range.

They are acid-tolerant, capable of surviving at a pH ranging from 2.7 to 8.3 and reportedly as low as 2.0 for short periods of time (Sheehan *et al.,* 2005). Notably, Legionellae display both an extracellular and intracellular lifestyle. Extracellularly, Legionellae survive in freshwater, soil or sediment habitats as free-living microorganisms or in biofilm communities, where they can acquire nutrients from other microbial members (Stewart *et al.,* 2012). Due to stress, nutrient shortages, temperature, pH fluctuations and oxidative stress in this environment, *Legionella* exist in a motile, virulent form (Heuner *et al.,* 2008). A recent study identified a gene, *lpg1659*, coding for a predicted membrane protein (LasM) that is strongly expressed when *Legionella* is free-living. It is thought that its expression may aid the long-term extracellular survival and culturability of *Legionella* in water (Li and Faucher*,* 2016). Furthermore, a homolog of *lpg1659* was identified in other free-living bacteria in water (Li and Faucher, 2016).

Legionellae can also survive and replicate within host cells. In the environment, *Legionella* host cells are primarily protozoa, including amoebae, ciliates and excavates (Rowbotham, 1986, Abu Kwaik *et al.,* 1998) (**Figure 1.1**). The protozoa phagocytose Legionellae, thereby providing them with a nutrient-rich shelter where the bacterial cells undergo a morphogenetic shift to a metabolically active form for replication and eventual dissemination. Experimental studies, co-culture and microscopy, have confirmed 3 phyla as protozoal hosts: the Amoebozoa (*Acanthamoebae spp., Hartmanella spp*., *Echinamoeba exudans*, *Dictyostelium discoideum* and *Balamuthia mandrillaris*), the Ciliophora (*Tetrahymena spp., Oxytricha bifaria*, *Stylonychia mytilus* and *Paramecium caudatum*) and the Percolozoa (*Naegleria spp., Tetramitus jugosus* and *Willertia magna*) (Boamah *et al.,* 2017). There is good agreement between experimentally confirmed host protozoa and protozoa co-isolated with Legionellae from the environment (Boamah *et al.,* 2017). A number of factors can influence the interaction between *Legionella* and the protozoal host including the species of protozoa, the preferences of the protozoal host for feeding on certain *Legionella* strains or species, the conditions in the external environment, the relative abundance of protozoal host versus *Legionella* and the presence and influence of other microorganisms (Boamah *et al.,* 2017).

The extracellular lifestyle of *Legionella* species may promote its ability to survive in harsh conditions and adapt to new environments (Oliva *et al.,* 2018). On the other hand, the association of *Legionella* with protozoa assists the replication and distribution of the bacterial cells and acts as a mechanism for biocide and chlorine avoidance and protection from unfavourable thermal conditions. It is hypothesised that both the extracellular and intracellular environments provide training grounds for the selection and persistence of *Legionella* phenotypes that can, in turn, influence their ability to infect human cells (Molmeret *et al.,* 2005, Oliva *et al.,* 2018).

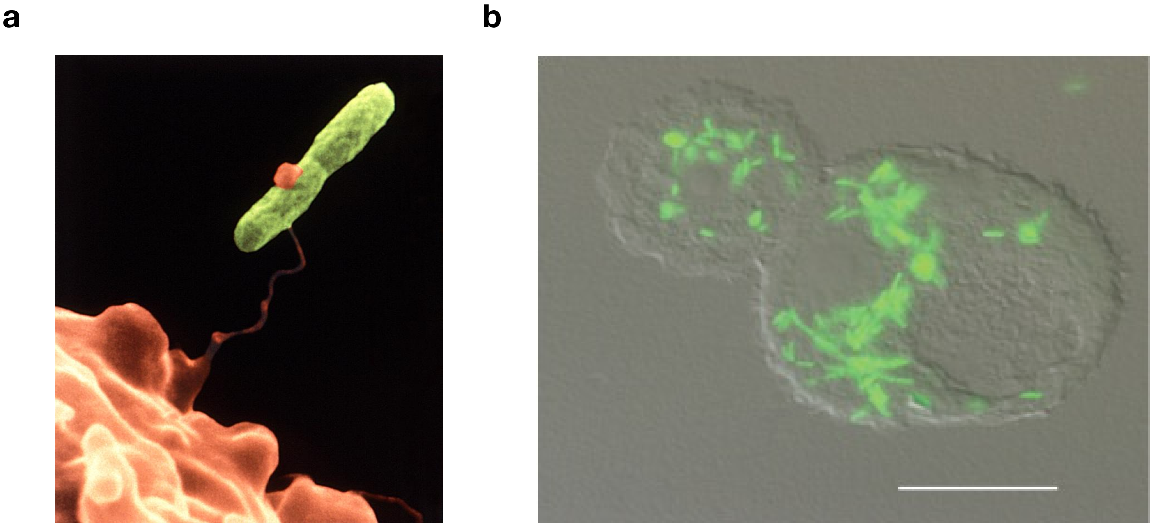


Figure 1.1 In the environment, *Legionella* host cells are primarily protozoa, including amoebae, ciliates and excavates. The protozoa phagocytose Legionellae, providing them with a nutrient-rich shelter for replication. (**a**) Electron micrograph depicting an amoeba as it entraps a *Legionella pneumophila* bacterium (green) with an extended pseudopod. Image is in the public domain. Image credit: CDC/Dr Barry S. Fields, PhD. (**b**) The amoeba *A. castellanii* infected with *L. pneumophila* expressing green fluorescent protein. Scale bar represents 10 μm. Image printed with permission by personal communication with Dr. Mena Abdel-Nour (Citation: Abdel-Nour *et al.*, 2013).

1.4 Sources of Infection, Modes of Transmission and Control

The emergence of legionellosis is considered to be due to human adaption of environments where Legionellae are naturally present. Transmission can occur by inhalation, aspiration and more rarely by direct contact or ingestion (Muder *et al.,* 1986) (**Figure 1.2**). Risk of exposure to *Legionella* exists from water systems, plants or equipment that release an aerosol or mist during operation. In this context, systems of risk include evaporative cooling towers (Fraser *et al.,* 1977, García-Fulgueiras *et al.,* 2003) which can distribute aerosols containing the bacteria greater than 6 km away (Nguyen *et al.,* 2006), potable water systems (Tobin *et al.,* 1980, Castellani Pastoris *et al*., 1999), hot springs (Sommese *et al.,* 1996), thermal spas (Martinelli *et al.,* 2001), domestic plumbing systems (Hayes-Philips *et al.,* 2019), evaporative condensers (Breiman *et al.,* 1990), water systems used in health such as dental equipment (Oppenheim *et al.,* 1987), heated birthing pools (Collins *et al.,* 2016), respiratory devices and nebulisers (Arnow *et al.,* 1982), humidifiers (Moran-Gilad *et al.,* 2012), fountains and water features (O'Loughlin *et al.,* 2007), industrial water systems (Allen *et al.,* 1999), wastewater treatment plants/systems (Kusnetsov *et al.,* 2010), misting devices (Mahoney *et al.,* 1992) and ice machines (Bangsborg *et al.,* 1995).

A number of conditions in man-made water systems increase the risk of microbial growth to levels that may cause legionellosis. These factors include the maintenance of a water temperature that encourages *Legionella* growth (or lack of ability to control water temperature), low or no water flow or water stagnation (potentially due to dead-legs or blind-ends), inadequate backflow protection, systems built with inappropriate materials which promote microbial growth and biofilm formation, lack of a continuous, good quality potable water supply entering the system, lack of an appropriate disinfection routine and lack of regular monitoring to ensure risk parameters are being met (European Technical Guidelines, 2017). Monitoring efforts should include testing biocide levels, pH, temperature, water turbidity, presence of dissolved solids and a microbiological and chemical assessment.

Efforts to control *Legionella* in man-made water systems should include a regular schedule of treatment with biocides, corrosion inhibitors and scale inhibitors (European Technical Guidelines, 2017). The presence of protozoa in water systems is deemed a risk factor for legionellosis. In one study, the quantity of *L. pneumophila* in biofilms correlated directly with the biomass of protozoa (Liu *et al.,* 2012). Other identified sources causing legionellosis include soils, potting mixtures and compost (Hughes *et al.,* 1994, Koide *et al.,* 1999, Cramp *et al.,* 2010). The mechanisms of transmission are not fully understood but are hypothesised to be through inhalation or ingestion of dust particles, not washing hands after gardening and being close to dripping hanging baskets (O' Connor *et al.,* 2007). The first case of probable human-to-human transmission of LD was reported in 2016 (Correia *et al.,* 2016).

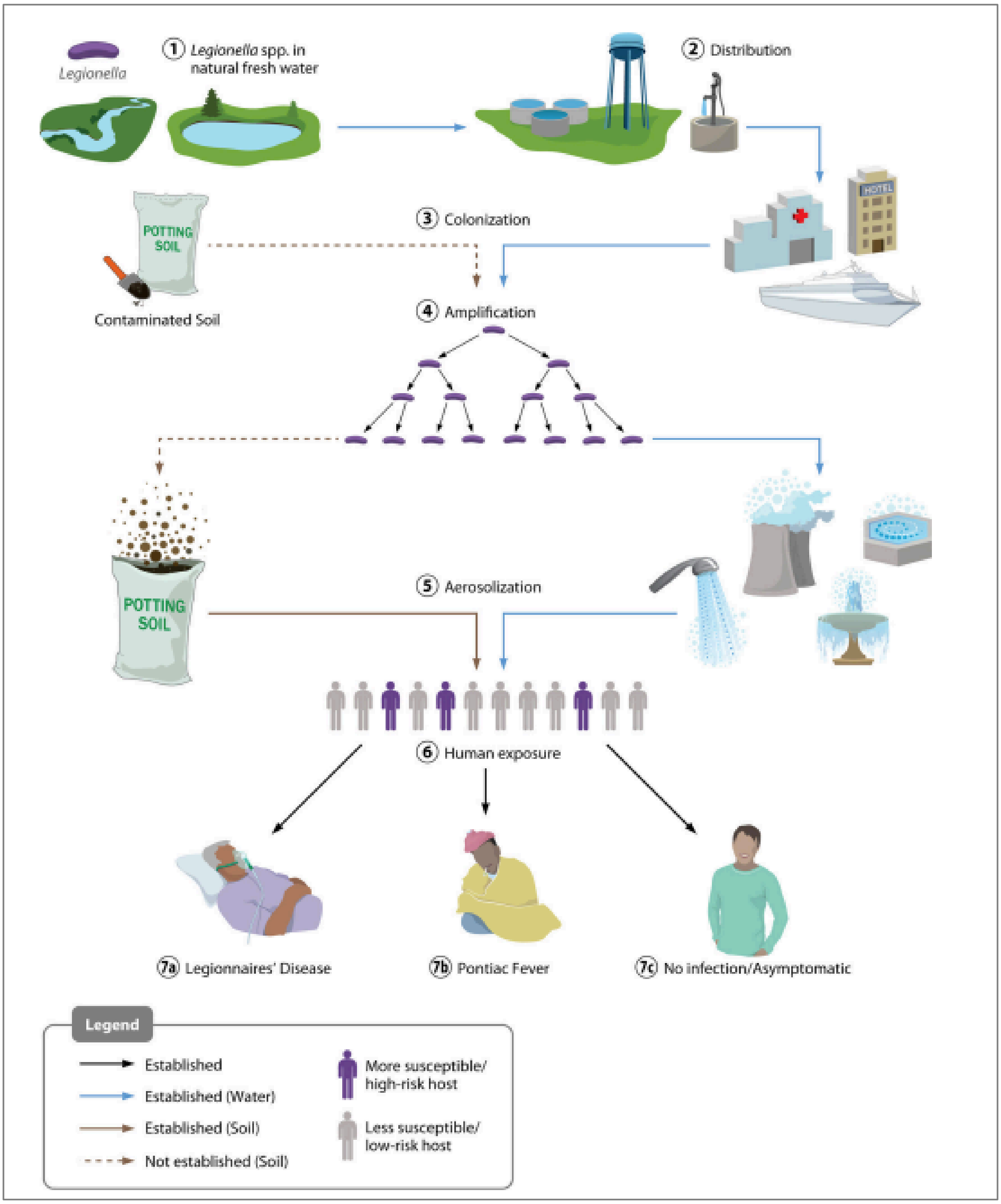


Figure 1.2 Route of *Legionella* dissemination from (1) the natural environment, (2) distribution to the man-made environment, (3) colonisation of water systems/soil, (4) amplification, (5) aerosolization, (6) human exposure and (7) infection/no infection/asymptomatic. Image reprinted with permission from the American Society for Microbiology Copyright © 2015 (Citation: Mercante *et al.,* 2015).

1.5 *Legionella* Pathogenesis

The mode of pathogenesis of human phagocytic cells by *Legionella* bacteria is similar to that of protozoa. *Legionella* are considered an “accidental” human pathogen. The pathogenic pathway of the *L. pneumophila* species is the most well-studied. After inhalation, small aerosols containing the bacteria can enter the lower respiratory tract. The bacteria can then bind to alveolar macrophages or monocytes via their pili (Stone *et al.,* 1998) or surface receptors on the host cells (Cianciotto *et al.,* 1992, Mintz *et al.,* 1992, Mintz *et al.,* 1995, Harb *et al.,* 1998, Declerck *et al.,* 2005, Declerck *et al.,* 2007). The bacteria are engulfed by the host cell through one or more reported mechanisms of phagocytosis (Horowitz *et al.,* 1984, Rittig *et al.,* 1992, Bozue *et al.,* 1996, Kwaik *et al.,* 1998, Khelef *et al.,* 2001, Tachado *et al.,* 2018). The *L. pneumophila* Dot/Icm Type IV secretion system (T4SS) (Berger *et al.,* 1993, Isberg *et al.,* 2009) secretes approximately 300 effector/virulence proteins into the host cell. T4SS is present in all studied *Legionella* species (Sánchez-Busó *et al.,* 2014, Burnstein *et al.,* 2016, Joseph *et al.,* 2016) and a recent genome-wide study reported the existence of 18,000 effector proteins across the *Legionella* genus (Gomez-Valero *et al.,* 2019), with some of the effector proteins being eukaryotic-like.

Effector proteins enable the “hijacking” of the host phagocyte-lysosome pathway. This acts as a mechanism for *L. pneumophila* to control membrane transport in the host cell, leading to the creation of the “*Legionella* Containing Vacuole” (LCV). As the LCV does not fuse with lysosomes (Horwitz *et al.,* 1983, Horwitz *et al.,* 1984) it protects the bacteria from enzymatic digestion. The efficiency of LCV formation is believed to be dependent on the infecting *Legionella* species (Newton *et al.,* 2010). The LCV intercepts the endoplasmic reticulum-Golgi vesicle traffic route and recruits small smooth endoplasmic reticulum vesicles and mitochondria to its membrane (Horwitz *et al.,* 1983, Swanson *et al.,* 1995). The membrane begins to resemble that of the rough endoplasmic reticulum (RER) in thickness and protein composition and becomes embedded with ribosomes (Tilney *et al.,* 2001). *L. pneumophila* replicate very efficiently within the LCV and eventually lyse the host cell thereby releasing the new replicants. **Figure 1.3** gives a graphical overview of the infection cycle.

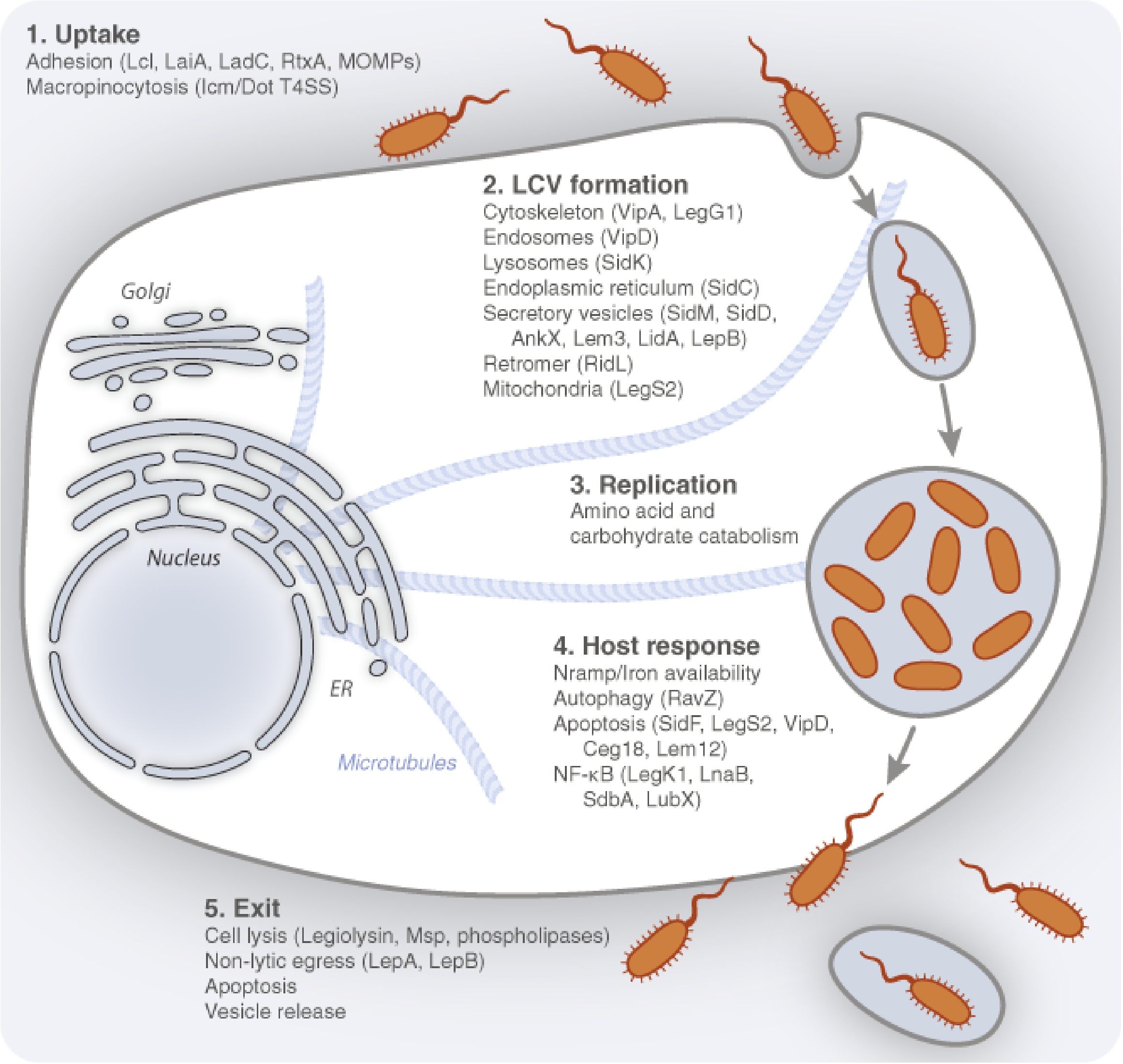

1. Uptake
Adhesion (Lcl, LaiA, LadC, RtxA, MOMPs)
Macropinocytosis (Icm/Dot T4SS)
2. LCV formation
Cytoskeleton (VipA, LegG1)
Endosomes (VipD)
Lysosomes (SidK)
Endoplasmic reticulum (SidC)
Secretory vesicles (SidM, SidD, AnkX, Lem3, LidA, LepB)
Retromer (RidL)
Mitochondria (LegS2)
Golgi
Nucleus
ER
Microtubules
3. Replication
Amino acid and carbohydrate catabolism
4. Host response
Nramp/Iron availability
Autophagy (RavZ)
Apoptosis (SidF, LegS2, VipD, Ceg18, Lem12)
NF-κB (LegK1, LnaB, SdbA, LubX)
5. Exit
Cell lysis (Legiolysin, Msp, phospholipases)
Non-lytic egress (LepA, LepB)
Apoptosis
Vesicle release

Figure 1.3 Overview of *L. pneumophila* infection cycle and associated virulence factors. Five main stages occur in the infection cycle: (1) uptake whereby the bacterium adheres to the surface of alveolar macrophages or monocytes via their pili or surface receptors on the host cells and can become engulfed by a number of phagocytic methods. The bacterial cell harnesses it's Dot/Icm Type IV secretion system to secrete effector proteins into the host cell. The effector proteins enable the "hijacking" of the host phagocyte-lysosome pathway leading to (2) LCV formation which protects the bacteria from enzymatic digestion. The bacteria can now efficiently (3) replicate avoiding a (4) host response and eventually lyse the host cell to (5) exit. Alongside host cell processes, bacterial factors and components implicated in those processes are displayed. Image reprinted with permission from John Wiley and Sons, Cellular Microbiology (License No. 4631370737170) (Citation: Hoffmann *et al.*, 2014).

1.6 Epidemiology

1.6.1 Clinical Features of Infection

1.6.1.1 Legionnaires' Disease

LD is considered a rare and sporadic respiratory infection and is classified as an atypical pneumonia. In many cases the symptoms of infection are non-specific. The incubation period is typically between 2 and 10 days but can sometimes be up to 14 days (Cunha *et al.,* 2016), particularly in immunocompromised individuals. Patients may first experience a prodromal illness (a period between the appearance of initial symptoms and the full development of symptoms). This is followed by clinical features of infection such as fever > 38.8 ^{o}C (and less often > 40 ^{o}C), cough, chills, muscle pain, dyspnoea, chest pain, headache, confusion, delirium, obtundation, seizures or focal problems, myalgia or arthralgia, diarrhoea, nausea, vomiting, abdominal pain (Cunha *et al.,* 2016). A dry cough is reported to occur in approximately 65 % of LD patients (von Baum *et al*., 2008). In those with a productive cough, sputum may be purulent and sometimes blood-streaked (Cunha *et al.,* 2016). Pulmonary infiltrates are visible in chest radiographs from approximately the third day of disease onset. At hospital admission, pleural effusion occurs in 15 to 50 % of patients (Viasus *et al.,* 2013). In immunocompromised patients, round nodular opacities or abscess can appear (Yu *et al.,* 1997). Overall recovery from LD can be slow and post-infection complications such as neurological or neuromuscular disorders, even post-traumatic stress disorder, can occur (Lettinga *et al.,* 2002).

The attack rate of LD is estimated to be 0.1 % to 5 % of individuals in the community and 0.4 % to 14 % of hospitalised individuals (World Health Organisation [WHO], 2007). Antimicrobial therapy is based on the oral or intravenous administration of macrolides, fluoroquinolones, cyclin families or a combination of these. Azithromycin and levofloxacin are typically recommended (Phin *et al.,* 2014). Crucially, beta-lactam antibiotics, usually used to treat bacterial community-acquired pneumonia cases, are ineffective in cases of LD.

1.6.1.2 Pontiac Fever

Pontiac fever is a febrile, non-pneumonic illness. Due to its obscurity, there is no defined definition of the illness. The incubation period is usually from 12 to 48 hours (WHO, 2007). Symptoms can include fever, shivers, headache, muscle aches, tiredness and dry cough. Antimicrobial treatment is usually not required.

1.6.1.3 Extra-Pulmonary Syndromes

Extrapulmonary manifestations of *Legionella* infection are rare and most commonly occur in immunocompromised patients. A previous study showed that *L. pneumophila* could spread from the respiratory system to other parts of the body (Lowry *et al.,* 1993). Infection can also occur without pneumonia. *Legionella* have been implicated in cases of soft tissue, wound and surgical site infections (Brabender *et al.,* 1983, Lowry *et al.,* 1991, Qin *et al.,* 2002, Chee *et al.,* 2007, Mentula *et al.,* 2014), abscesses (Arnow *et al.,* 1983, Anderson *et al.,* 1987, La Scola *et al.,* 1999, Gubler *et al.,* 2001, Charles *et al.,* 2013), cellulitis (Kilborn *et al.,* 1992, Waldor *et al.,* 1993, Loridant *et al.,* 2011), arthritis (Flendrie *et al.,* 2011), osteomyletis (McClelland *et al.,* 2004, Sanchez *et al.,* 2013), ocular (Heriot *et al.,* 2014) and heart manifestations such as myocarditis, pericarditis, post-cardiomyotomy syndrome and endocarditis (Guyot *et al.,* 2007, Tanabe *et al.,* 2009, Samuel *et al.,* 2011, Ishimuro *et al.,* 2012,, Compain *et al.,* 2015). In very rare cases, *Legionella* may spread to the nervous system (Shelburne *et al.,* 2004).

1.6.2 Risk Factors, Seasonality and Incidence

Legionnaires disease is more likely to occur in males greater than 50 years of age (**Figure 1.4**). Occurrence is also more likely in individuals with a history of smoking, heavy alcohol use and underlying medical conditions such as immunosuppression, lung disease, diabetes, heart disease and renal disease (Marston *et al.,* 1994). Recent travel is an additional recognised risk factor. While LD is uncommon in children and neonates, cases have been reported to occur in association with hospital water systems, birthing pools and respiratory equipment (Shachor-Meyouhas *et al*., 2010, Yiallouros *et al.,* 2013. Phin *et al*., 2014).

LD cases exhibit a seasonality, with cases primarily occurring from June to October in Europe (European Centre for Disease Prevention and Control [ECDC], 2017) (**Figure 1.5**). During 2017, the notification rate for the European Union/European Economic Area (EU/EEA) was 1.8 per 100,000 inhabitants overall (ECDC, 2019) (**Figure 1.6**). In the United States (US) from 2000 to 2015, the notification rate increased from 0.42 to 1.89 per 100,000 inhabitants (Centre for Disease Control [CDC], 2018). From 30 EU/EEA countries, 9,238 cases of LD were reported. A total of 8 % of cases with a known outcome were reported as fatalities. Overall, ECDC reported a 30 % rise in cases during 2017 when compared to 2016 (ECDC, 2019). This increase in incidence could be due to improved diagnostics and surveillance efforts. Additionally, other studies indicate that rising temperatures, higher than average atmospheric changes and increases in humidity may be responsible for the proliferation of LD incidence (Fisman *et al.,* 2005, Hicks *et al.,* 2007).

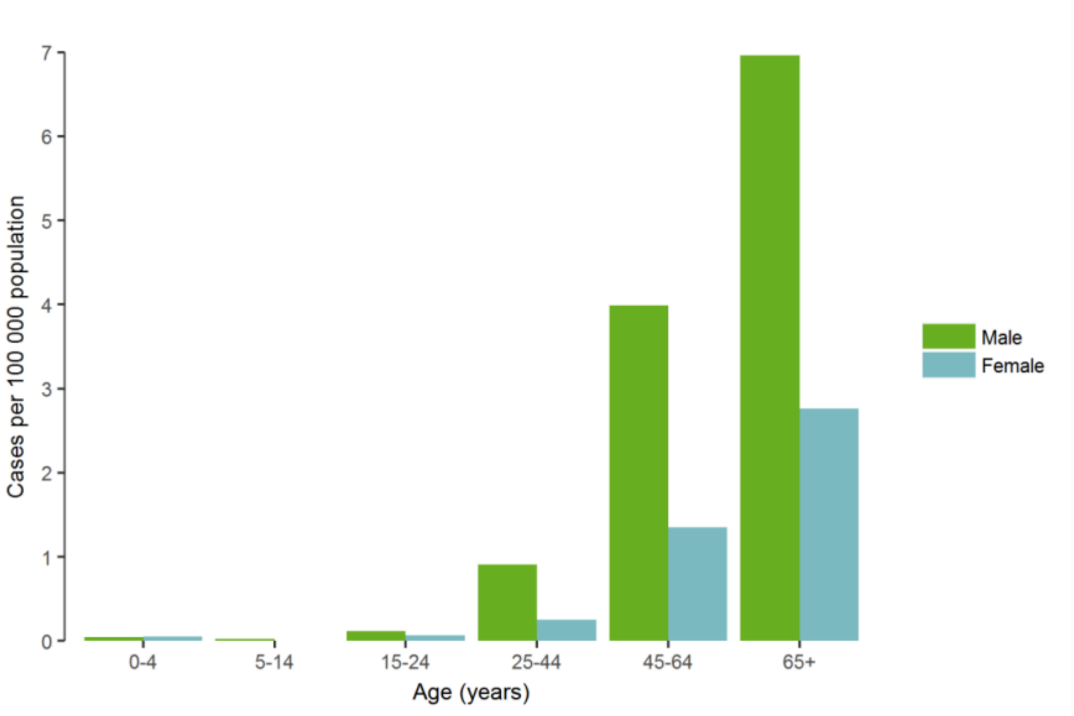
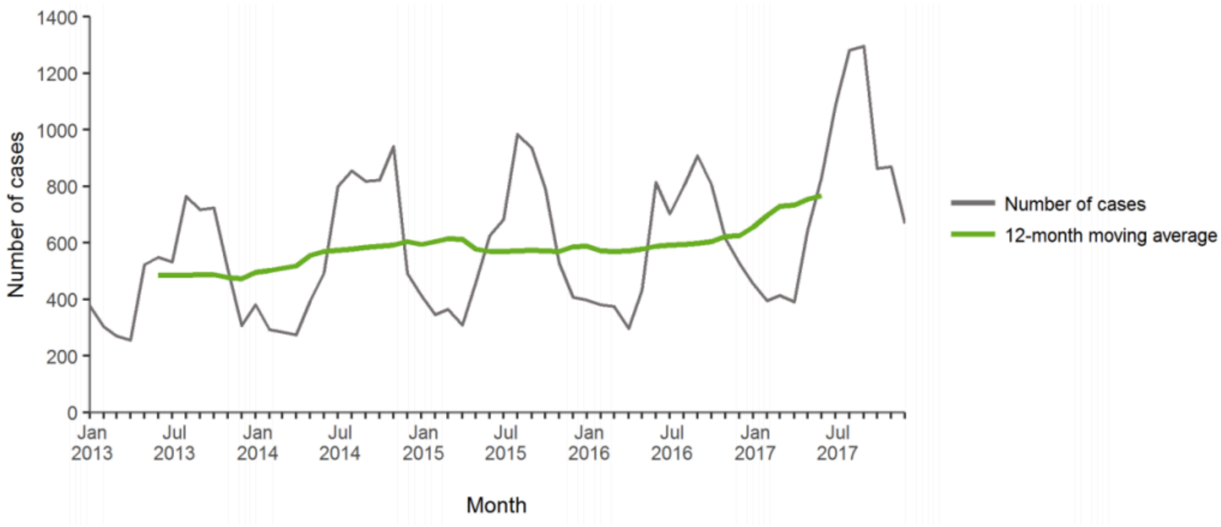


Figure 1.4 Distribution of Legionnaires' disease cases per 100,000 inhabitants (EU/EEA) by age and gender during 2017. Image reprinted with permission from ECDC. (Citation: European Centre for Disease Prevention and Control. Legionnaires' disease. In: ECDC. Annual epidemiological report for 2017. Stockholm: ECDC; 2019.)



Source: Country reports from Austria, Belgium, Bulgaria, the Czech Republic, Cyprus, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden and the United Kingdom.

Figure 1.5 Distribution of Legionnaires' disease cases (EU/EEA) by month from 2013 to 2017. Image reprinted with permission from ECDC. (Citation: European Centre for Disease Prevention and Control. Legionnaires' disease. In: ECDC. Annual epidemiological report for 2017. Stockholm: ECDC; 2019.)

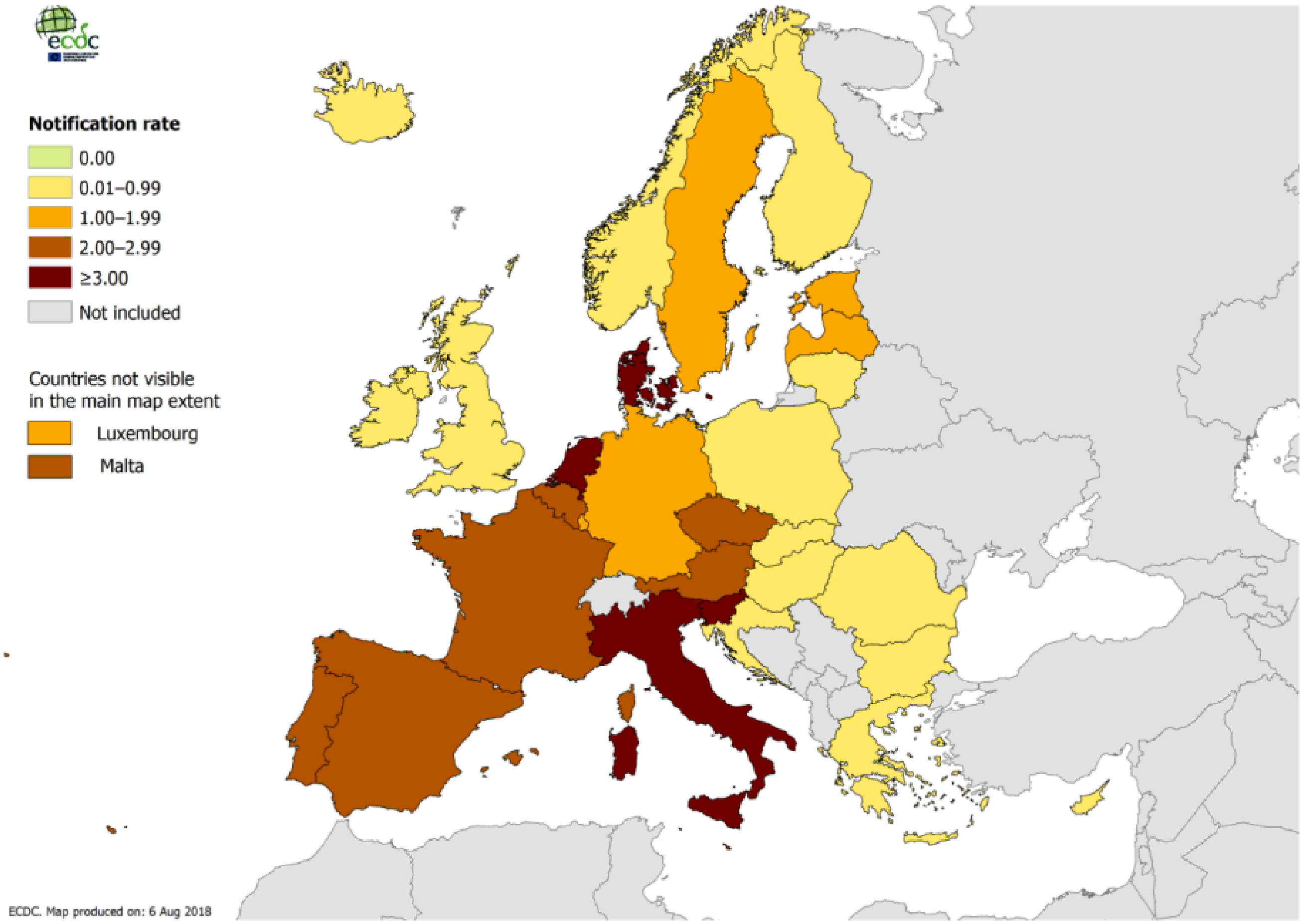


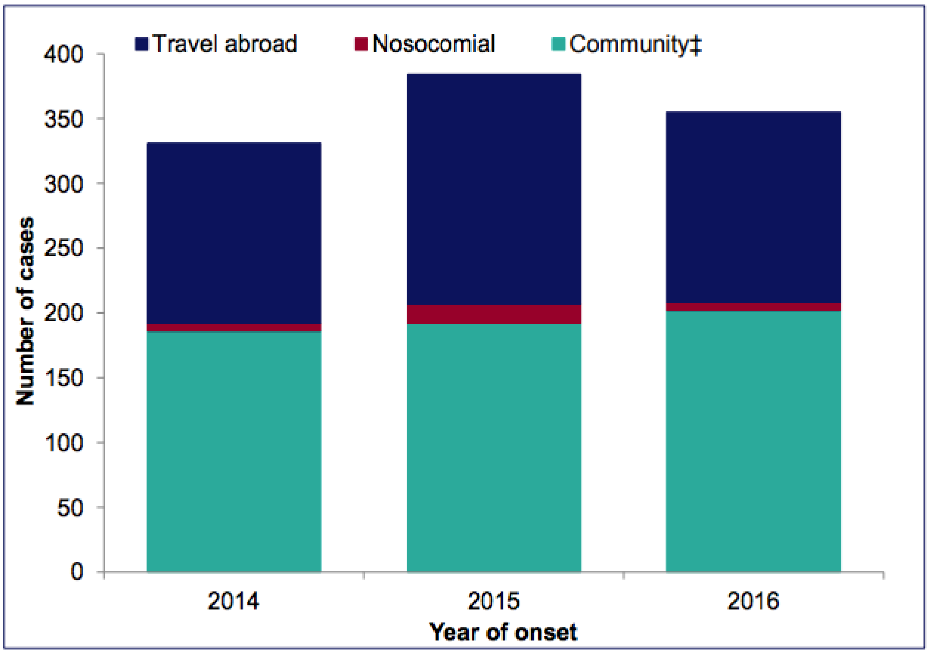
Figure 1.6 Notification Rate of Legionnaires' Disease cases by country in the EU/EEA, 2017. Image reprinted with permission from ECDC. (Citation: European Centre for Disease Prevention and Control. Legionnaires' disease. In: ECDC. Annual epidemiological report for 2017. Stockholm: ECDC; 2019.)

1.6.3 Community, Nosocomial and Travel-Associated Legionnaires' Disease

LD has been a notifiable disease in England since 2010 (Department of Health [DH], 2010). It can occur in the community (including care homes), in hospitals/healthcare settings and in travel settings as single cases, clusters or outbreaks.

A confirmed case of LD is defined by Public Health England (PHE) as a clinical/radiological diagnosis of pneumonia and laboratory confirmed detection. To investigate an environmental source of infection, a 10-day exposure history of the case is collected. For a case to be considered healthcare associated, the individual should have significant exposure to a hospital/healthcare setting for 2 to 10 days prior to symptom onset. For a case to be considered travel-associated, the individual should have stayed/visited accommodation, such as a hotel, cruise ship or campsite, outside the individual's area of residence for 2 to 10 days prior to symptom onset. When 2 or more cases occur in close proximity (up to 6 km for community cases and the same hospital/accommodation site for healthcare- and travel-associated cases), time (up to 6 months for community cases and up to 2 years for healthcare- and travel-associated cases), and share an epidemiological link, it is defined as a cluster. If 2 or more cases have onset of symptoms at a maximum of 28 days apart, share a strong epidemiological link and microbiological evidence of a common source of infection, it is defined as an outbreak (PHE, 2019).

From all confirmed LD cases in England and Wales (2014 to 2016), on average 54.4 % were exposed in the community, 2.4 % in hospital/healthcare settings and 43.2 % during travel abroad (PHE, 2017[a]) (**Figure 1.7**).



‡ includes travel UK cases

Figure 1.7 Confirmed Legionnaires' Disease cases in England and Wales by exposure category (community, nosocomial or travel abroad) and year of onset. Figure taken from PHE, 2017(a): Legionnaires' Disease in residents of England and Wales: 2016. Image reprinted under the terms of the Open Government License v3.0.

1.7 Detection, Diagnosis and Epidemiological Typing

1.7.1 Urinary Antigen Test

The first line method for the diagnosis of legionellosis in England is the urinary antigen test (UAT). The UAT uses monoclonal antibodies against *L. pneumophila* serogroup 1 lipopolysaccharide antigen (LPS). Tests are commercially available in enzyme immunoassay (EIA) and immunochromatographic test (ICT) format (Harrison *et al.,* 1998, Dominguez *et al.,* 1999, Helbig *et al.,* 2001, Diederen *et al.,* 2007). Antigen is detected in urine 2 to 3 days from the initiation of symptoms (Mercante *et al.,* 2015). Antigen detection may be possible up to months after the initiation of symptoms in some cases. Commercially available UATs are specific for the detection of antigen from *L. pneumophila* serogroup 1 (Lp1) which is the dominant species and serogroup. Currently a new commercial UAT test (Sofia [Quidel]) for the detection of serogroup 1, 2, 4 and 6 is available. There is however, reported cross reactivity for the detection of LPS antigen from *L. pneumophila* non-sg1 serogroups (Olsen *et al.,* 2009). These detections vary significantly between different tests. Consequently, the extent of cross-reactivity is unknown (Chen *et al.,* 2015). The UAT is advantageous in that it is rapid, low cost, easy to implement, urine sample collection is non-invasive and appropriate antibiotic treatment can be initiated in a timely manner.

It however has disadvantages in that approximately 8 % of legionellosis patients do not excrete *Legionella* antigen in their urine (Munoz *et al.,* 2009). Additionally, due to the specificity of the UAT for Lp1, the test may be creating a selection bias or blind spot in the diagnosis of LD pneumonia caused by other serogroups and species (Mercante *et al.,* 2015). For example, Australia, New Zealand and Thailand have a greater number of LD cases where *L. longbeachae* has been isolated (Phares *et al.,* 2007, The Institute of Environmental Science and Research Ltd [ESR], 2017, National Notifiable Diseases Surveillance System [NNDS], 2019).

Surveillance data from Scotland from 2015 – 2016 reported the isolation of *L. pneumophila* sg1 from 9 cases and other *L. pneumophila* serogroups and species from 7 cases (Health Protection Scotland [HPS], 2017). Denmark employs a more comprehensive PCR-based LD surveillance system when compared with other EU countries. A study of LD cases over more than a decade from 1993 to 2006 revealed that only 60 % of cases were attributable to *L. pneumophila* serogroup 1 with the others representative of other *L. pneumophila* serogroups and species (St-Martin *et al.,* 2013). It is unknown if Denmark is an outlier in terms of *Legionella* detection profile or if LD cases caused by other serogroups and *Legionella* species are escaping the system of detection in other countries. A recent study from the Raphael *et al*., 2019, reported extensive diversity of *L. pneumophila* strains in Arizona, USA whereby from 236 *L. pneumophila* isolates, 28.2 % belonged to serogroup 6 and 8.9 % to serogroup 8.

1.7.2 Serology and Direct Fluorescent Antibody Testing

Serological analysis and direct fluorescent antibody assays (DFA) of tissue and respiratory specimens are the traditional approaches applied for *Legionella* detection (McDade *et al.,* 1977). Serological analysis is based on measuring IgM and IgG levels. It is not however a timely method for diagnosis for the individual patient or for public health outbreak investigations since seroconversion confirmation requires the collection of acute and convalescent sera 4 to 8 weeks apart to show a 4-fold increase in titre (Edelstein, 1987).

DFA involves the microscopic analysis of an antibody conjugated with a fluorochrome. The method is rapid, cost effective and applicable up to four days after the initiation of antibiotic therapy. It is very specific (99.9 %) for the identification of *L. pneumophila* however sensitivity is low (between 25 and 70 % for culture positive specimens) (WHO, 2007).

1.7.3 Culture

Isolation of *Legionella* is the gold standard for LD diagnosis. Culture is carried out on lower respiratory tract samples, ideally sputum. It can be difficult however for LD patients to produce sputum due to a dry cough which is reported to occur in 65 % of cases (von Baum *et al.,* 2008). Culture can therefore also be carried out on bronchial alveolar lavage fluid (BAL), pleural fluid, bronchial aspirates, lung biopsy or tissue specimens (including post-mortem specimens). In cases of extrapulmonary infection, culture can be carried out on blood, fluid from joints and soft tissue (WHO, 2007, Cunha *et al.,* 2016). For environmental source testing, water, biofilm and soil or sediment samples can be cultured (including swabs).

Culture is carried out on buffered charcoal yeast extract (BCYE) agar supplemented with 0.1 % alpha-ketoglutarate, L-cysteine and ferrous salts. Incubation is carried out at 35 to

37 ^{o}C in a humidified, 2.5 % CO_2 atmosphere (Feeley *et al.,* 1979, Edelstein, 1981). *Legionella* is considered fastidious and a difficult-to-grow pathogen. Isolates typically grow in 3 to 5 days but some take longer (Fields, 2005). In England, after 10 days of no growth, culture is declared negative. Colonies have a grey-white colour and a characteristic cut/ground glass appearance with a pink, blue or green iridescence. The rapid growth of contaminating flora (such as *Pseudomonas aeruginosa* and *Candida* spp.) can cause issues on plated specimens, overwhelming the media and growth of *Legionella* colonies. Contaminants can be controlled by supplementing the media with antibiotics (Edelstein *et al.,* 1981, Dournon *et al.,* 1988) or by exposure to acid or heat treatment. Culture sensitivity is variable (from < 10 % to 80 %) and dependent on laboratory proficiency as well as sample type and quality. For example, an isolate may be more difficult to obtain if a specimen was collected after commencement of antibiotic therapy. In England, a study from 2012 reported that only 64 % of urinary antigen positive specimens could be cultured (Mentasti *et al.,* 2012). Furthermore, some *Legionella* cannot be isolated on culture media. These are termed *Legionella*-like amoebal pathogens (LLAPs). Their isolation is only possible by co-culture with particular protozoan species (Fry *et al.,* 1991). Additionally, seroprevalence studies suggest a human pathogenic role for a number of other LLAPs (McNally *et al.,* 2000, Marrie *et al.,* 2001).

1.7.4 PCR-based Detection

Polymerase chain reaction-based molecular methods for the detection of *Legionella* are rapid, sensitive and specific. When a *Legionella* UAT positive sample is culture negative, PCR can improve the sensitivity of detection from 64 % to > 80 % (Mentasti *et al.,* 2012). There is no general consensus by *Legionella* reference laboratories regarding PCR targets or the usage of PCR to confirm detection, however the majority of laboratories in Europe, including PHE, use the ESGLI qPCR method. The ESGLI method is a 3-plex *Legionella* qPCR to specifically detect and identify *L. pneumophila* by the macrophage infectivity potentiator gene (*mip*) and *L. pneumophila* serogroup 1 strains by the O-antigen ABC transporter permease (*wzm*) gene (Mentasti *et al.,* 2012). Conventional PCR assays based on the *mip* and 16S rRNA gene can also be applied for *Legionella* species identification of *Legionella* species (Wilson *et al.,* 2003, Wilson *et al.,* 2007).

1.7.5 Epidemiological Typing

Epidemiological typing of *L. pneumophila* is necessary to confirm or refute a link between LD cases and environmental sources. A number of genotypic and phenotypic typing methods exist such as pulsed field gel electrophoresis (PFGE) (Lück, *et al.,* 1991), amplified fragment length polymorphism (AFLP) (Valsangiacomo, *et al.,* 1995) and monoclonal antibody (mAb) typing (Helbig *et al.,* 2002). While these methods or a combination of the methods are used in some laboratories, they lack sufficient discriminatory power for *L. pneumophila* strain identification.

The current gold standard for epidemiological typing of *L. pneumophila* is a 7-loci sequence-based typing (SBT) approach developed by the European Study Group for *Legionella* Infection (ESGLI) and analogous to multi-locus sequence-based typing (MLST) (Gaia *et al.,* 2005, Ratzow *et al.,* 2007, Mentasti *et al.,* 2014). Conventional PCR amplification and Sanger sequencing is carried on DNA extracted from a *L. pneumophila* isolate using specially designed primers for the following seven loci: *flaA* (flagellin gene), *pilE* (type IV pilin gene), *asd* (aspartate-B-semialdehyde dehydrogenase gene), *mip* (macrophage infectivity potentiator gene), *mompS* (major outer membrane precursor protein gene), *proA* (zinc metalloprotease gene) and *neuA*/*neuAh* (lipopolysaccharide biosynthesis gene and its homolog). Sequences are then submitted to the *L. pneumophila* SBT database (www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php) where they undergo a quality control step and are compared to a database of numbered allele sequences. An allele number is assigned to each locus sequence if there is a 100 % match. Otherwise, a new allele number is created. Ultimately, a 7-digit combination of allele numbers (allelic profile) corresponding to a sequence type is assigned to the isolate.

On the 19th July 2019, the SBT database hosted 12,935 sample records composed 2,791 sequence types. Of these entries 33 % were from an environmental source and 66 % from clinical sources. The sample entries currently represent 2,738 unique *L. pneumophila* sequence types.

When culture is negative, a nested PCR-based SBT (Ginevra *et al.,* 2009) approach can be carried out on DNA extracted directly from the clinical specimen or environmental sample itself (Scaturro *et al.,* 2011, Mentasti *et al.,* 2016, Quero *et al.,* 2019). PHE relies on the 7-loci SBT for investigations of clusters and outbreaks of LD. In North West Europe, sequence types from clinical specimens submitted to the SBT database are disproportionately represented by ST1, ST23, ST37, ST47 and ST62 (Borchardt *et al.,* 2008, Harrison *et al.,* 2009, Mentasti *et al.,* 2012, David *et al.,* 2016(a)). Owing to this, it is often difficult and sometimes impossible to distinguish to between clusters and confirm an environmental reservoir of infection and LD investigations can remain unresolved.

Presently, epidemiological typing schemes are moving towards the analysis of whole genome data from isolates. Whole genome sequencing (WGS) of bacteria isolates has moved from the proof-of-concept stage to implementation as a routine methodology in public health reference laboratories. PHE has to date established a WGS service for a number of pathogens including Mycobacteria, *Escherichia coli*, *Staphylococcus aureus*, Salmonella, Listeria, Shigella, *Streptococcus pneumoniae* and Campylobacter (PHE, 2018). The move to WGS of pathogens is advantageous in that it provides heightened resolution over traditional genotyping tests which analyse partial genome information.

A pilot study from 2013 demonstrated the utility of WGS of *L. pneumophila* for the investigation of LD isolates during outbreaks (Reuter *et al.,* 2013). Subsequent studies have proven the utility of whole genome sequences during LD investigations (McAdam *et al.,* 2014, Raphael *et al.,* 2016, David *et al.,* 2017[a], Schjorring *et al.,* 2017, Buultjens *et al.,* 2018, Timms *et al.,* 2018).

Recently, core genome MLST schemes have been developed for whole genome data from *L. pneumophila* isolates. In one study a scheme based on the analysis of 1,521 core genes was described (Moran-Gilad *et al.*, 2015). Furthermore, David *et al.,* 2016(b) validated a MLST scheme based on 50 core genes (that can be extended to 100, 500, 1,000 and 1,455 gene schemes) with very high overall discriminatory power.

1.8 *Legionella* Genomics

Legionella pneumophila genomes range in size from 3.3 to 3.5 megabases. They have on average 3,000 protein coding genes and a G+C content of 38 % (Cazalet *et al.,* 2004, Chien *et al.,* 2004). Early genomic studies reported the presence of multiple genes encoding eukaryotic-like proteins, a mechanism through which the microorganism exploits host cell functions (Cazalet *et al.,* 2004, de Felipe *et al.,* 2005, Gomez-Valero *et al.,* 2019). *Legionella* genomes display high plasticity and diversity even within the same species and serogroup with approximately 10 % of the whole genome not represented by the core genome of the species (Cazalet *et al.,* 2004). A 2011 study of *L. pneumophila* sg1 genomes first reported that recombination between strains from eukaryotes and other bacteria to

L. pneumophila play a role in the diversification of the species. These mechanisms also play a role in the relatively rapid replacement of strains over time within defined niches (Gomez-Valero *et al.,* 2011). Furthermore, a study by David *et al.,* 2017(b) reported that recombination within the species has played a major role in the population structure and evolution of *L. pneumophila* and David *et al.* also hypothesised that recombination of multiple regions from a single donor may occur within a single recombination event.

Homologous recombination refers to the importation of DNA from another source, which is generally closely related, to replace a homologous segment of the original DNA. Non-homologous recombination refers to the importation of DNA composed of entirely new genes. Conjugation, transduction and transformation, which are mechanisms of recombination, have all been described for *L. pneumophila* (Dreyfus *et al.,* 1985, Mintz *et al.,* 1987, Stone *et al.,* 1999). Interestingly, contradictory to the recombination dynamics in studied *L. pneumophila* sequence types, a study of ST47 isolates demonstrated high clonality and no recombination (David *et al.,* 2016[a]). The authors discussed that this may be related to the recent emergence of the sequence type or the adaption of the sequence type to a niche with limited opportunities for recombination.

1.9 Mixed Infection by *L. pneumophila* and Other *Legionella* Species

Two key genomic studies of *L. pneumophila* isolates from LD outbreaks in Spain (Coscollá *et al.,* 2014) and Scotland (McAdam *et al.,* 2014) reported the involvement of multiple genotypes of specific sequence types of *L. pneumophila* in clinical cases. This was postulated to be due to the the recombination of *L. pneumophila* and accumulation of mutations over time within environmental populations before dispersal to humans. Other prior studies have reported the presence of mixed *L. pneumophila* serogroups, species and sequence types isolated from LD cases. A summary of cases of mixed infection reported to date in the published literature are detailed in **Table 1.2**. Contrastingly, a study by David *et al*., 2018 reported low genomic diversity of *L. pneumophila* within clinical specimens. The study involved the analysis of ten *L. pneumophila* isolates from ten epidemiologically unrelated individuals.

Table 1.2 Mixed Infection by *Legionella pneumophila* and other *Legionella* species

Mixed Infection	Clinical Specimen	Exposure Category	Country/Year	Reference
L. pneumophila (two different serogroups)	Sputum	Nosocomial	USA/Not stated	Meyer *et al.*, 1980
L. pneumophila sg1 *L. micdadei*	Tracheal aspirate	Not defined	USA/Not stated	Dowling *et al.*, 1983
L. pneumophila *L. micdadei*	Not defined	Not defined	Italy/Not stated	Fumarola *et al.*, 1984
L. pneumophila (multiple serogroups)	Lung tissue	Not defined	Berlin/Not stated	Horbach *et al.*, 1988
L. pneumophila *L. gormanii*	Lung biopsy	Not defined	Germany/Not stated	Buchbinder *et al.*, 2004
L. pneumophila *L. rubrilucens*	Sputum	Community	Japan/Not stated	Matsui *et al.*, 2010
L. pneumophila sg 1 *L. pneumophila* sg 3	Tracheal secretions	Nosocomial	Cyprus/2008	Yiallouros *et al.*, 2013
L. pneumophila sg 1 *L. pneumophila* sg 3	Blood	Nosocomial	Austria/2010	Wewalka *et al.*, 2014
L. pneumophila sg 3 *L. bozemanaii*	Tracheal aspirate	Community	Denmark/2002	Wewalka *et al.*, 2014
L. pneumophila sg1 (two different STs)	Tracheal aspirate	Travel	Denmark/2004	Wewalka *et al.*, 2014
L. pneumophila sg3 *L. pneumophila* sg6	Tracheal aspirate Sputum	Nosocomial	Denmark/2007	Wewalka *et al.*, 2014
L. pneumophila sg6 *L. bozemanaii*	Tracheal aspirate	Community	Denmark/2011	Wewalka *et al.*, 2014
L. pneumophila sg1 (two different STs)	Sputum	Travel	Denmark/2012	Wewalka *et al.*, 2014
L. pneumophila sg1 *L. dumofii*	Pericardial fluid	Nosocomial	Germany/2008	Wewalka *et al.*, 2014
L. pneumophila sg1 (two different STs)	BAL*	Nosocomial	Germany/2009	Wewalka *et al.*, 2014
L. pneumophila sg1 (two different STs)	Respiratory tract samples	Nosocomial	Germany/2010	Wewalka *et al.*, 2014
L. pneumophila sg1 (two different STs)	Respiratory tract samples	Community	Germany/2011	Wewalka *et al.*, 2014

L. pneumophila sg11 *L. longbeachae* sg1	BAL*	Nosocomial	UK/1997	Wewalka *et al.,* 2014
L. pneumophila sg1 *L. pneumophila* sg6	BAL*	Nosocomial	UK/2004	Wewalka *et al.,* 2014
L. pneumophila sg1 *L. pneumophila* sg6	Lung tissue specimen	Community	UK/2006	Wewalka *et al.,* 2014
L. pneumophila *L. bozemanii*	BAL*	Nosocomial	UK/2009	Wewalka *et al.,* 2014
L. pneumophila variants	Sputum	Community	Spain/2008	Coscollá *et al.,* 2014
L.pneumophila sg 1 ST191 subtypes and different STs	Respiratory secretions	Community	Scotland/2012	McAdam *et al.,* 2014

* BAL = bronchoalveolar lavage fluid

1.10 Metagenomic Sequencing

Metagenomic sequencing is the sequencing of all genetic material (microbial communities and host DNA) as it exists in a sample of interest. It is therefore unbiased in representing the genetic composition of the sample when compared, for example, to targeted 16S rRNA approaches which only examine the bacterial community. For metagenomic sequencing to be carried out, an adequate quantity of the sample of interest is collected, nucleic acid is extracted, library preparation and sequencing is carried out and the data undergoes quality control and analysis. To date, metagenomic sequencing has been applied in the investigation of natural environments (freshwater [Oh *et al.,* 2011], marine [Biller *et al.,* 2018], soil [Kroeger *et al.,* 2018], glaciers [Kayani *et al.,* 2018]), wastewater (Gupta *et al.,* 2018), disease vectors (Greay *et al.,* 2018), human and animal biomes (Lloyd-Price *et al.,* 2016, Stewart *et al.,* 2018), food (Leonard *et al.,* 2015, De Filippis *et al.,* 2016), clean rooms (Bashir *et al.,* 2016), forensic cases (Hampton-Marcell *et al.,* 2017) and ancient remains (Chan *et al.,* 2013).

Metagenomes can be interrogated for their functional capabilities, signatures of antibiotic resistance, associations with disease and health states, development of therapeutics as well as the investigation of microbes of clinical and public health importance.

1.10.1 Sequencing Technologies

The first breakthrough DNA sequencing technology was Sanger dideoxy chain terminator sequencing (Sanger *et al.,* 1977). Sanger sequencing is based on the priming of the DNA strands to provide a start point for the initiation of DNA synthesis, the polymerisation of dNTPs onto the strand and strand extension through the incorporation of chain-terminating dideoxy nucleotides. The application of Sanger sequencing prevailed for decades and is still in use today due to the methodology having low sequence error rates. Few reads however are actually produced by the method. The second-generation sequencing platforms (454, Illumina, Ion Torrent, SOLID) addressed the read limitation by scaling read throughput. Sequencing is massively parallelisable on these platforms and millions of short reads (50 to 600 base pairs) with low error rates are produced.

The company Illumina currently dominates the second-generation sequencing market with its multiple sequencing platforms (iSeq, MiniSeq, MiSeq, NextSeq, HiSeq and NovaSeq). The Illumina sequencing technology is based on sequencing-by-synthesis whereby millions of clusters of reads are generated by bridge amplification (Kawashima *et al.,* 2005) and clusters are sequenced simultaneously using fluorescent 'reversible terminator' dNTPs. Despite the advantages in throughput generated by sequencing-by-synthesis platforms, third generation sequencing technology, like PacBio (recently acquired by Illumina), produce very long reads albeit with higher sequencing errors and lower throughput. The third-generation sequencing technology is based on single cell real-time sequencing by a zero-mode waveguide (ZMW) (Levene *et al.,* 2003). Nanopore sequencers are the fourth-generation sequencing technology, the most well-known being Oxford Nanopore Technologies (ONT) (whose platforms include the MinION, GridION, PromethION) (Jain *et al.,* 2015). ONT technology involves the transit of single-stranded DNA molecules through protein pores embedded in a membrane. The nucleotides are read by the effect they individually generate on an electrical current, producing an optical signal. ONT sequencers can sequence very long DNA strands. In addition, the real-time base-calling and portability of the devices have shown efficacy in public health microbiology for outbreak investigations (Quick *et al.,* 2016, Faria *et al.,* 2017).

A major advantage of the use of both second- and third or fourth generation technologies is that a hybrid *de novo* genome assembly approach can be applied to combine the advantages of low error rates and long reads. In recent years, this technique has aided the scaffolding and closure of many bacterial genomes to high accuracy (Wick *et al.,* 2017[a]).

1.10.2 Metagenomics in Clinical and Outbreak Investigations

The development and evolution of sequencing technologies has enabled the rapid investigation of infectious disease cases and outbreaks by metagenomic sequencing without prior isolation. The first reported case to demonstrate the utility of metagenomic sequencing in a clinical setting was the diagnosis of leptospirosis in a case of meningoencephalitis. The diagnosis enabled rapid clinical action (Wilson *et al.,* 2014). One benchmark study investigated metagenomic sequencing of Shiga Toxigenic *Escherichia coli* (STEC) O104:H4 during a 2011 outbreak in Germany (Loman *et al.,* 2014). In the study, near complete STEC genomes were assembled from a number of the investigated faecal specimens. Whilst the study demonstrated the feasibility of metagenomic sequencing it must be acknowledged that faecal material contains an abundance of microbes or may be defined as high microbial biomass. The colon reportedly contains 10^{11} microbial genome copies per millilitre (Whitman *et al.,* 1998). The detection of pathogens from respiratory samples generates additional challenges. In a study of respiratory samples from patients with confirmed bacterial pneumonia, only 1 % of sequenced reads were microbial, with the remaining 99 % constituting human DNA reads (Pendleton *et al.,* 2017). Additionally, in a study examining the direct metagenomic sequencing of culture and smear positive TB samples, *Mycobacterium tuberculosis* reads were detected in the sample however with very low genome coverage. The sample sequences were composed of approximately 99 % human reads (Doughty *et al.,* 2014).

1.10.3 Challenges of Metagenomic Sequencing and Bioinformatic Analysis

Sequencing pathogens from low microbial biomass samples, such as specimens from the lung or freshwater which contain a significantly lower microbial biomass, poses a challenge for a number of reasons. There is less initial starting template and the pathogen frequently constitutes an extremely low proportion of the original specimen. Ideally large quantities of the specimen of interest should be sampled however there are technical constraints to achieving this. Furthermore, the complex background community of host

DNA or DNA from other microorganisms can inhibit detection of the pathogenic organism. Metagenomes from certain sites of the human body may indeed be overwhelmed by human DNA reads (**Figure 1.8**) (National Institute of Health [NIH], 2009, Marotz *et al.,* 2018).

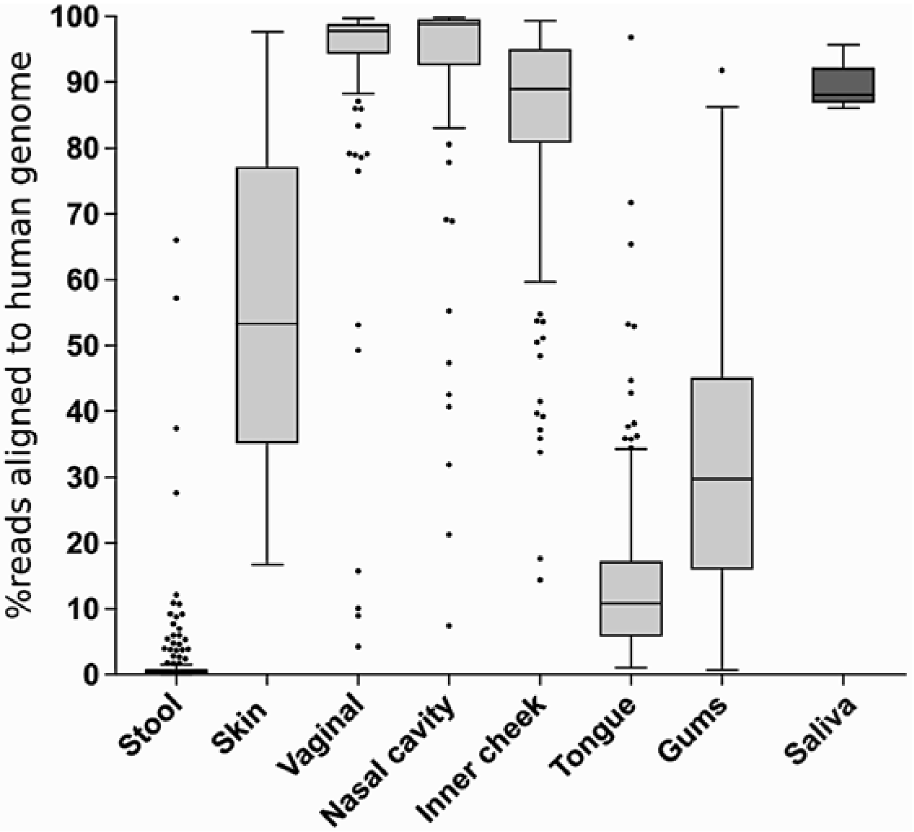


Figure 1.8 Percentage of shotgun metagenome sequencing reads aligning to the human genome varies by samples type. Data for stool, skin, vaginal, nasal cavity, inner cheek, tongue and gums from the NIH Human Microbiome Project (NIH, 2009) of healthy individuals. Saliva data collected by Marotz *et al.,* 2018. Figure reprinted with permission from Marotz *et al.,* 2018.

The challenge of overwhelming human DNA may be addressed by ultra-deep sequencing of the sample although ultra-deep sequencing is currently too expensive to be implemented as a routine method for pathogen detection. A targeted approach can be carried out using RNA baits (as will be discussed in Chapter 5) or a tiling PCR-based method. This firstly requires knowledge of the presence of the microorganism. Due to the levels of host background, untargeted approaches applied to low microbial biomass clinical samples first require the depletion of human DNA. A number of human DNA depletion methods have been developed (summarised in **Table 1.3**) Analysis of a number of these methods have however reported variable results (Marotz *et al.,* 2018) (**Figure 1.9**).

Table 1.3 Human DNA Depletion Methods and Kits

Human DNA Depletion Method/Kit	**Principle**	**Reference**
NEBNext Microbiome DNA Enrichment Kit (New England Biolabs)	Immunomagnetic separation targeting methylated human DNA	Feehery *et al.,* 2013
MolYsis Basic5 Kit (Molzyme)	Differential human cell lysis followed by DNase treatment	NA
QIAamp DNA Microbiome Kit (Qiagen)	Differential human cell lysis followed by DNase treatment	NA
Immunoprecipitation of DNA with inactive methyl-specific restriction endonucleases		Barnes *et al.,* 2014
Immunoprecipitation of DNA with inactive methyl-specific restriction endonucleases		Liu *et al.,* 2016
Osmotic lysis followed by propium monoazide (PMA) treatment		Marotz *et al.,* 2018
Saponin-based lysis followed by DNase treatment		Hasan *et al.,* 2016
Hypotonic lysis and endonuclease digestion with benzonase2		Nelson *et al.*, 2019
Saponin-based lysis followed by HL-SAN treatment		Charalampous *et al.,* 2019

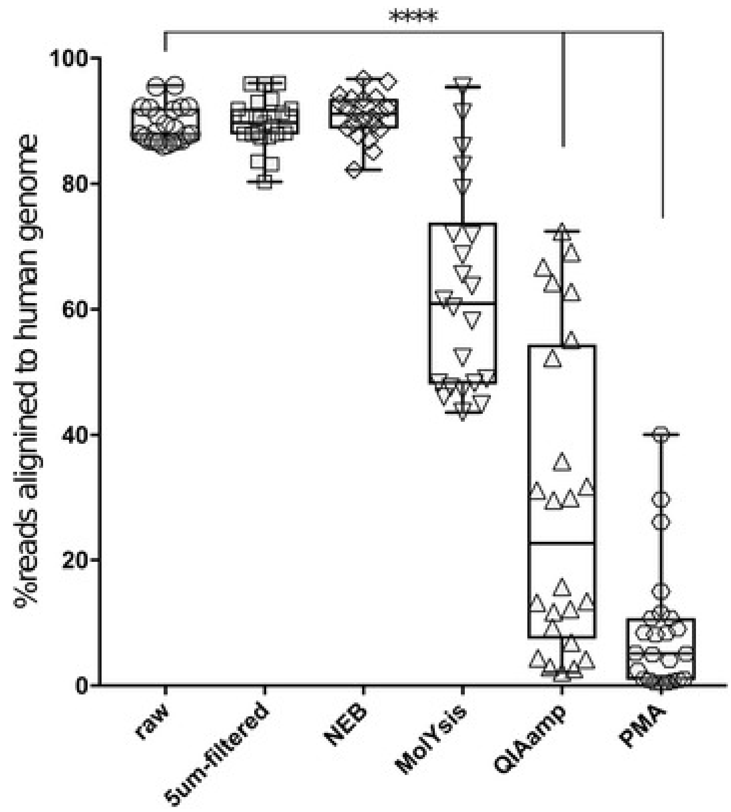


Figure 1.9 Percentage of shotgun metagenome sequencing reads aligning to the human genome before and after human DNA depletion from saliva, a low microbial biomass specimen. The figure demonstrates the effectiveness, ineffectiveness as well as the variability of different methods for human DNA depletion from saliva. PMA = Propium Monoazide treatment. For a description of the methods see **Table 1.3.** Figure reprinted with permission from Marotz *et al.,* 2018.

Human cell depletion by lysis-based methods can lead to the elimination of bacteria without cell walls such as *Mycoplasma pneumoniae* or removal of cell-free nucleic acid from bacteria that have lysed during antibiotic treatment or during sample processing. Furthermore, Charalampous *et al.,* 2019 reported autolysis of *S. pneumoniae* using their method.

In addition to the burden of host DNA in samples, contaminants often make up a larger proportion of the microbial community than that of high microbial biomass specimens (Salter *et al.,* 2014). These can be introduced at any stage during sample processing and sequencing including during treatment to remove human DNA.

Another significant challenge is related to the bioinformatic analysis of metagenomic data. A number of programmes address various steps in metagenomic analysis such as assembly, binning, taxonomic classification and functional evaluation. There is however no gold standard approach or pipeline for end-to-end metagenomic analysis as it is not trivial and introduces both computational and conceptual challenges. Metagenome assembly refers to the overlap and merging of reads into longer genomic contigs. Since metagenomes contain numerous species of unknown abundance, it may not however be possible to *de novo* assemble low abundance species particularly when multiple low abundance species are present. Furthermore, a significant challenge of metagenome assembly lies in the deconvolution of closely related species and particularly strains of the same species (Teeling and Glöckner, 2012). The incorporation of reads from closely related species and strains may result in chimeric contigs. This is particularly problematic when closely related species/strains in a metagenome are present at extremes of high and low abundance. Despite a number of assembly tools being developed to address strain-level assembly such as metaSPAdes (Nurk *et al.,* 2017), MetaVelvet (Namiki *et al.,* 2011) and Meta-IDBA (Peng *et al.,* 2011), the challenges still exist.

The taxonomic classification of metagenomes is based on the alignment or k-mer mapping of metagenomic reads or contigs to a database of microbial whole genomes or marker genes by a classifier. The classifier then assigns a taxon to each read/contig. This is a useful mechanism for binning and assembly of genomes from metagenomic data if closely related genomes are available in the database. Examples of taxonomic classifiers for metagenomic data include Centrifuge (Kim *et al.,* 2016), Kraken (Wood *et al.,* 2014) which are based on whole genome data and MetaPhlan (Segata *et al.,* 2012) which is based on marker gene data. A disadvantage of taxonomic classification is the reliance on the composition of the databases which often do not contain representatives for the entire diversity of the metagenome. Contig binning can also be carried out by unsupervised approaches, based on compositional characteristics of the sequence data such as coverage and nucleotides (Johannes *et al.,* 2014, Kang *et al.,* 2015, Wu *et al.,* 2016). Functional analysis or annotation of metagenomes brings additional challenges as a large proportion of genes are of unknown function. It has been proposed that for the functional study of metagenomes, the clustering and analysis of open reading frames (ORFs) should be carried out. This idea runs adjacent to how operational taxonomic units (OTUs) are clustered and analysed in biodiversity studies (Teeling and Glöckner, 2012).

Despite the caveats, the application of metagenomic sequencing can provide a sequence agnostic method for the investigation of pathogens of clinical and public health relevance.

It has the potential to remove timely culture steps, allowing the rapid discovery of pathogens and initiation of antimicrobial treatment without ambiguity. Furthermore, it can provide evidence of mixed infections, detection of cell-free nucleic acid and non-culturable or difficult-to-culture pathogens, therefore eliminating gaps in epidemiological knowledge created by current diagnostic selection bias.

1.11 Hypothesis and Aims

The hypothesis of this thesis was that the development of methods for the metagenomic sequencing of *Legionella* from clinical and environmental specimens may provide a more timely approach for the detection and identification of *Legionella*, reduce diagnostic selection bias and provide insights into potential mixtures of *L. pneumophila* subtypes, sequence types and *Legionella* species in samples which might aid investigations.
This was investigated in four parts by:

1. Validation of a metagenomic sequencing and analysis approach for the investigation of *Legionella*:
   The sequencing accuracy of mock communities containing *Legionella* species, the sensitivity/limit of detection of *L. pneumophila* in mock samples, the investigation of a method to determine mixed *L pneumophila* strains in mock samples and the contribution of host contamination in clinical samples were investigated.

2. Method development for the depletion of human DNA:
   A number of approaches targeting repetitive regions of human DNA were investigated including the use of biotinylated Cot-1 DNA probes, *Alu* DNA probes, *Alu* RNA probes and an *Alu* PCR to incorporate biotin into *Alu* elements.

3. A pilot study for the targeted capture of *L. pneumophila*:
   An Agilent SureSelect™ approach based on the application of biotinylated RNA baits for the capture of *L. pneumophila* was carried out on clinical and environmental specimens.

4. The investigation of Legionnaires' Disease outbreaks by metagenomic sequencing:
   Two Legionnaires' Disease outbreaks in England were investigated using Agilent SureSelect™ capture of *L. pneumophila* and direct nanopore sequencing of clinical specimens.

Chapter 2.

Materials and Methods

2.1 Overview

This chapter provides a detailed description of:

- The clinical, environmental, and isolate specimens as well as the mock community material studied in this thesis.
- Nucleic acid extraction from the specimens by the phenol-chloroform method.
- General laboratory methods including DNA quantification, DNA purification, DNA size selection and DNA fragment size analysis.
- Methods for hybridisation.
- The methodology for Illumina metagenomic sequencing.
- The methodology for target capture sequencing and associated database design for RNA bait generation.
- Illumina metagenomic data pre-processing/cleaning and the bioinformatic tools used.
- The methodology for Oxford Nanopore metagenomic library preparation, sequencing and data processing.
- The methodology for 16S rRNA gene sequencing and data analysis.

2.2 General Purpose Laboratory Equipment and Consumables

Table 2.1 details the laboratory equipment items and consumables used (and referenced to) for the experiments conducted in this thesis:

Table 2.1. Laboratory Equipment and Consumables

Item	Application	Manufacturer
Heraeus Pico 21 Microcentrifuge	Centrifugation at room temperature	Thermo Fisher Scientific Inc; MA, USA
Mikro 220R microcentrifuge	Centrifugation at 4 °C	Hettich; Tuttlingen, Germany
Rotanta 460 Centrifuge	Centrifugation of 96-well plates at room temperature	Hettich; Tuttlingen, Germany
ViiA™ 7 Real-Time PCR System	Quantitative PCR	Thermo Fisher Scientific Inc; MA, USA
DNA Engine Tetrad 2 Peltier Thermal Cycler	Quantitative PCR	Bio-Rad; CA, USA
Microcentrifuge Tubes (1.5 ml)	Multiple applications	Eppendorf; Stevenage, UK
Nucleic acid-free PCR grade water	Multiple applications	Qiagen; Hilden, Germany
MiliQ® Type 1 Ultrapure water	Multiple applications	Millipore Corporation; MA, USA
Ethanol (EtOH) > 99.8 % - Absolute grade	Purification washes	Fisher Scientific; MA, USA
MicroAMP™ Fast 96-well Reaction Plate (0.1 ml)	qPCR	Applied Biosystems; CA, USA
MicroAMP™ Optical Adhesive Film	qPCR	Applied Biosystems; CA, USA
Dynal® Magnetic bead separation stand	DNA/Library purification	Invitrogen; CA, USA
Magnetic Stand-96	DNA/Library purification	Invitrogen; CA, USA

2.3 Clinical and Environmental Specimens

Ethical approval was granted from the Research Ethics Committee (REC reference number 16/NS/0014) through the Integrated Research Application System (IRAS project ID: 195410) on the 25th January 2016 to allow selection of clinical specimens, both positive and negative for *Legionella* species, from the sample archive of the Respiratory and Vaccine-Preventable Bacteria Reference Unit (RVPBRU), Public Health England (PHE). Environmental specimens were provided by the Food, Water and Environmental Laboratory, PHE. Upon receipt at PHE, samples had been processed according to established in-house protocols within the UKAS accredited laboratory (personal communication Dr. Victoria Chalker).

2.4 Material for Mock Communities

Bacterial type strains and human DNA were used in the preparation of mock communities and series dilution tests. The specific composition of each one is discussed in the relevant chapters.

Legionella species type strains were provided by RVPBRU, Public Health England (PHE). Strains for other bacterial species (*S. pneumoniae, Haemophilus influenzae, Veillonella dispar)* and one *L. longbeachae* strain (Long Beach 4) were purchased from the Leibniz Institute DSMZ (Deutsche Sammlung von Mikroorganismen und Zellulturen GmbH). The identities of the *Legionella* strains were confirmed by 16S rRNA and the macrophage infectivity potentiator (*mip*) gene sequencing, a gene that enables speciation of *Legionella* (Ratcliff *et al.,* 1998). The identities of the other bacterial strains were confirmed by 16S rRNA gene sequencing by the Genomic Medicine Section team at the National Heart and Lung Institute (NHLI), Imperial College London. Information for all bacterial strains used is described in **Table 2.2**.

Stocks of human DNA from controls for a previous Wellcome Trust study undertaken by Professor Miriam Moffatt and Professor William Cookson were used. The blood was taken from healthy individuals who gave informed consent. Ethics for blood collection was approved under REC Reference 01/5/006 NRES Committee East of England, Cambridge South.

Table 2.2 Bacterial Species and Strain Material used in Mock Sample Preparations

Species	Strain Designation	Collection No.	Source
Legionella pneumophila	Philadelphia-1	NCTC 11192	PHE Bacterial Culture Collection, UK
Legionella pneumophila	France 5811	NCTC 12007	PHE Bacterial Culture Collection, UK
Legionella pneumophila	OLDA	NCTC 12008	PHE Bacterial Culture Collection, UK
Legionella longbeachae	NSW150	Unknown	PHE Bacterial Culture Collection, UK
Legionella longbeachae	Long Beach 4	DSM 10572	Leibniz Institute DSMZ, Germany
Legionella anisa	WA-316-C3	NCTC 11974	PHE Bacterial Culture Collection, UK
Legionella cherrii	ORW	NCTC 11976	PHE Bacterial Culture Collection, UK
Legionella feelei	WO-4CC	NCTC 12022	PHE Bacterial Culture Collection, UK
Legionella hackeliae	Lansing-2	NCTC 11979	PHE Bacterial Culture Collection, UK
Legionella micdadei	PPA	NCTC 11372	PHE Bacterial Culture Collection, UK
Streptococcus pneumoniae	SV 1	DSM 20566	Leibniz Institute DSMZ, Germany
Haemophilus influenzae	680 Biotype II	DSM 4690	Leibniz Institute DSMZ, Germany
Veillonella dispar	ERN	DSM 20735	Leibniz Institute DSMZ, Germany

2.4.1 Whole Genome Amplification (WGA)

Due to insufficient DNA (both quantity and concentration), whole genome amplification (WGA) was carried out on the *Legionella pneumophila* and *Legionella longbeachae* strains using the Illustra™ Ready-To-Go™ GenomiPhi™ HY DNA Amplification Kit (GE Healthcare; Illinois, USA). A 25 µl total of 2x Denaturation buffer was added to 2.5 µl (10 ng total) of DNA template after which 22.5 µl PCR-grade water was added to make a final volume of 50 µl. The DNA template was denatured by heating the sample mix to 95 °C for 3 minutes and then cooling to 4 °C on ice. The Ready-To-Go GenomiPhi HY cake was reconstituted in 50 µl of the cooled, denatured DNA template. The wells were sealed with domed caps and the reaction was kept on ice prior to incubation at 30 °C for 4 hours during which the DNA amplification occurred. After 4 hours, the Phi29 DNA polymerase enzyme was inactivated by heating the samples to 65 °C for 10 minutes followed by cooling to 4 °C. Post-WGA, DNA concentrations were measured by the PicoGreen® dsDNA Quantitation as described in Section 2.6.1 below. The amplified samples were stored in a freezer at -20°C until required.

2.5 Phenol-Chloroform DNA Extraction

Clinical, environmental and isolate material was extracted by the phenol-chloroform liquid-liquid DNA extraction method. This method is based on the phase separation of DNA from proteins and other cell lysate material. When phenol-chloroform is added to a biological specimen in lysis buffer, two phases form: an aqueous phase and a phenol phase. When the phases are mixed, phenol is forced into the water phase allowing an emulsion of droplets to form. The proteins present in the water phase are denatured and compartmentalise into the phenol while the DNA remains in the water. Once the phases are separated by centrifugation, the water phase containing the DNA is removed. It is then added to a Polyethylene Glycol (PEG) buffer and allowed to precipitate overnight. The resulting pellet is purified by repeat washes with 70 % ethanol (DeAngelis *et al*., 2009).

2.5.1 Buffer Preparation

2.5.1.1 CTAB Extraction Buffer

A solution of 1M NaCl, sodium chloride (Sigma, Gillingham, UK) was prepared by adding 58.44 g NaCl to 1 litre of *MiliQ* Type 1 Ultrapure H_20. Once the NaCl crystals had dissolved, a CTAB (hexadecytrimethylammonium bromide) 10 % w/v solution was prepared by adding 50 g of CTAB (Sigma) to 500 ml of the 1M NaCl. A 1M NaH_2PO_4 (monobasic phosphate) solution was prepared by adding 11.998 g of monobasic phosphate (Sigma) to 100 ml 1M NaCl. A 1M Na_2HPO_4 (dibasic phosphate) was prepared by adding 70.98 g of dibasic phosphate (Sigma) to 500 ml 1M NaCl. The monobasic phosphate (15.9 ml) solution was added to the dibasic phosphate (284.1 ml) solution and made up to 600 ml with 1M NaCl. The phosphate buffer and CTAB solution were combined 1:1. The completed CTAB extraction buffer solution was sterilised by autoclaving (2100 Classic Portable Sterilizer, Prestige Medical; Blackburn, UK).

2.5.1.2. PEG/NaCl Precipitation Solution

A 500 ml 1.6M NaCl solution was prepared by adding 46.752 g NaCl to 500 ml *MiliQ* H_20. A 30 % (w/v) solution of PEG was prepared by adding 150 g PEG (Sigma) to 500 ml of 1.6M NaCl. The solution was sterilised by autoclaving.

2.5.2 Sample Preparation

2.5.2.1 Sputum Samples

A total of 300 µl of sputasol-treated sputum (1:1) was transferred to a Lysing Matrix E (LME) tube (MP Biomedicals; CA, USA) containing 500 µl of CTAB buffer. Tubes were stored at -20 °C until the extraction protocol was carried out.

2.5.2.2 Environmental Samples

Environmental water samples (1 ml) were centrifuged for 30 minutes at top speed (21,000 x g). A total of 700 µl was removed, and the pellet was resuspended in the remaining 300 µl of water. The 300 µl resuspension was added to a LME tube containing 500 µl of CTAB buffer. Tubes were stored at -20 °C until the extraction protocol was carried out.

2.5.2.3 Bacterial Isolates

A *L. pneumophila* colony was picked from a BCYE (buffered charcoal yeast extract) agar plate at RVPBRU, PHE and suspended in 300µl of *MilliQ* H_2O (work carried out by Dr. Victoria Chalker). The bacterial suspension was transferred to a LME tube containing 500 µl of CTAB buffer. Tubes were stored at -20°C until the extraction protocol was carried out.

2.5.3 Extraction

Samples in LME tubes were removed from the freezer and allowed to defrost. A 50 µl aliquot of filtered 0.1M Aluminium ammonium sulphate ($AINH_4(SO_4)_2.12H_2O$) (Sigma) was added to each LME tube. A total of 500 µl of Phenol:Chloroform:Isoamyl alcohol 25:24:1 pH 8.0 (Sigma) was carefully and immediately added to each tube. LME lids were securely fastened and tubes were transferred to a bead-beater (FastPrep-24™ 5G Instrument; MP Biomedicals) and beat using the settings: Speed: 5.5m/sec, Adapter: Quickprep, Time: 60 seconds, Lysing Matrix: E, Quantity: 1 ml, Cycles: 1, Pause time: 300 seconds. Tubes were centrifuged at 16,000 x g for 5 minutes at 4 °C. All liquid was transferred to pre-spun heavy phase lock gel tube (VWR; PA, USA) and the tubes were kept on ice. Each phase lock tube was centrifuged at 16,000 x g for 5 minutes at 4 °C as the gel forms a barrier between the aqueous and Chloroform:Isoamyl alcohol phases. One volume of Chloroform:Isoamyl alcohol 24:1 (Sigma) was added to each phase lock tube and shaken briefly to mix. Tubes were then centrifuged at 16,000 x g for 5 minutes at 4 °C until the gel formed a barrier. For the second extraction, 500 µl of CTAB, 50 µl of aluminium ammonium sulphate and 500 µl of Phenol:Chloroform Isoamyl alcohol were added to each bead beating tube and the process was repeated.

2.5.4 Precipitation and Purification

The aqueous phase from each tube was transferred to 1.5 ml microcentrifuge tubes containing 1 µl of Linear Polyacrylamide, GenElute-LPA (Sigma) as a DNA carrier. Two volumes of the PEG/NaCl solution were added to each tube and mixed well. Tubes were left overnight at 4 °C to precipitate. All tubes were centrifuged at 16,000 x g for 20 minutes at 4 °C. The PEG/NaCl solutions were carefully aspirated from the pellets. Pellets were washed with 500 µl of ice-cold 70 % EtOH to remove any precipitated salts and then centrifuged at 16,000 x g for 5 minutes. The wash was repeated an additional two times with 200 µl of 70 % EtOH. The pellet was air-dried for 5 minutes and resuspended in 30 µl of low EDTA TE (10 mM Tris, pH 8.0 and 0.1 mM EDTA, Invitrogen; CA, USA). Extracts were stored in tethered O-ring sterile tubes (Starlab; Milton Keynes, UK) at -20°C in a protected space until required.

2.6 dsDNA Quantification

2.6.1 PicoGreen dsDNA Quantification

The PicoGreen ® dsDNA Quantitation assay kit was used for the quantification of dsDNA in solution. PicoGreen is a dye that intercalates with double stranded DNA and this results in the release of a fluorescent signal. The fluorescein emission wavelength is read by a microplate spectrofluorometer reader and a concentration is calculated against a linear standard curve. A Lambda bacteriophage DNA standard (100 ng/µl) was equilibrated to room temperature. Eight 1.5 ml microcentrifuge tubes were labelled 1-8, and 1X Tris Borate EDTA Buffer (TE - freshly prepared) was transferred into them as follows:- Tube 1: 594 µl, Tubes 2-8: 300 µl. A total of 6 µl of DNA standard was added to Tube 1 (100x dilution: 1 ng/µl) and a dilution series was made by transferring 300 µl from one tube into the next and vortexing for 10 seconds, up to and including Tube 7. Tube 8 constituted the "no DNA control". To the wells of column 11 and 12 of a 96-well black fluorometer plate, 100 µl of each DNA standard dilution was transferred for duplicate measurements. A total of 99 µl of 1x TE was added to the remaining wells of the 96-well black fluorometer plates. A 1 µl total of each DNA sample was added to the appropriate well of the fluorometer plates and upon addition, mixed by pipetting up and down 4 times. A 1:200 dilution of Quant-iT™ PicoGreen® dsDNA was prepared in a falcon tube and protected from light by wrapping with aluminium foil. A total of 100 µl diluted Quant-iT™ PicoGreen® dsDNA was added to each well and pipetted up and down carefully 4 times to mix. The final volume in each well was 200 µl. The sample fluorescence was measured using a microplate reader (TECAN Infinite M Plex) capable of excitation at 480 nm and reading emission at 520 nm. A standard curve was plotted and only accepted if the R^2 value of the curve was verified as > 0.998. DNA concentrations were determined using the standard curve.

2.6.2 Qubit® dsDNA BR Assay

A working solution was prepared by diluting the dsDNA BR Reagent 1:200 in dsDNA BR Buffer for two standards and the required number of samples. In each 0.5 ml Qubit assay tube for standard 1 and standard 2, 190 µl of working solution was added to each tube and 10 µl of standard 1 to tube 1 and standard 2 to tube 2. For each sample, 199 µl of working solution and 1 µl of sample was added to a 0.5 ml Qubit assay tube, as required. The tubes were vortexed for 2 to 3 seconds and incubated at room temperature for 2 minutes. A Qubit ® 3.0 Fluorometer was calibrated for dsDNA Broad Range assay by inserting the standards sequentially into the sample chamber and initiating the reader. Samples were read immediately after calibration in the same manner.

2.7 DNA Purification and Size Selection

2.7.1 AMPure XP DNA Purification

Agencourt ® AMPure ® XP (Beckman Coulter, High Wycombe, UK) beads were used for the purification of amplicon, library and genomic DNA. AMPure DNA purification is based on the binding of DNA molecules greater than 100 base pairs in size to solid-phase paramagnetic beads. Wash steps are carried out to remove impurities such as salts and enzymes and oligos or DNA fragments less than 100 base pairs. Purified DNA is then eluted from the beads in the desired volume of TE buffer.

The required volume of AMPure XP beads was added to the DNA sample for purification in a 1.5 ml microcentrifuge tube. The tube containing the sample-bead mix was shaken for 2 minutes at room temperature. This was followed by a further incubation for 5 minutes at room temperature without shaking. The mix was placed on a Dynal ® magnetic stand for 2 minutes until the liquid cleared. The supernatant was carefully removed and discarded using a pipette. The beads were washed twice with 200 µl 70 % EtOH and, after the final wash, any residual 70 % EtOH was removed by pipetting. Beads were left to air-dry on the magnetic stand for approximately 10 to 15 minutes. The tube was then removed from the magnetic stand and the required volume of Low TE buffer (Invitrogen) was added. The sample was shaken for 2 minutes at room temperature. The tube was then incubated at room temperature for a further 2 minutes without shaking. The tube was placed on the magnetic stand for 2 minutes after which the required volume of the clear supernatant was transferred to a new tube. Purified DNA was stored at -20 °C until further use.

2.7.2 Qiagen QIAquick PCR Purification Kit

A total of 5 volumes of Buffer PB (Qiagen) was added to 1 volume of sample for purification and mixed. A QIAquick spin column (Qiagen) was placed in a 2 ml collection tube. To bind the DNA, the sample was applied to the column and centrifuged for 1 minute at 16,000 x g. The flow-through was discarded and the column was placed back into the same tube. A total of 750 µl of wash Buffer PE (Qiagen) was added to the column and centrifuged for 1 minute as before. The flow-through was discarded and the column placed back into the same tube and centrifuged for an additional 1 minute. The column was placed in a clean 1.5 ml Eppendorf DNA LoBind tube. To elute the DNA, the required quantity of Qiagen supplied 1 x Low TE buffer was added to the centre of the column. The column was allowed to incubate for 1 minute at room temperature and was then centrifuged for 1 minute to elute the purified DNA. This was then stored at -20 °C until further use.

2.7.3 NEB Monarch PCR Purification Kit

The sample to be purified was diluted with DNA Cleanup Binding Buffer (NEB) by adding a 2:1 ratio of binding buffer to sample. A column was inserted into a collection tube and the binding buffer-sample solution was loaded onto the column. The column was centrifuged for 1 minutes at 16,000 x g and the flow-through was discarded. The column was re-inserted into the collection tube. A total of 200 µl of DNA Wash Buffer (NEB) was added and the column was centrifuged for 1 minute as before. The wash step was repeated. The column was transferred to a clean 1.5 ml Eppendorf DNA LoBind tube, taking care to ensure that the tip of the column did not come into contact with the flow-through. The required volume of supplied 1 X Low TE buffer was added to the centre of the column matrix. After a 1 minute incubation at room temperature, the column was centrifuged for 1 minute to elute the DNA. This was then stored at -20 °C until further use.

2.7.4 Pippin Prep DNA Size Selection and Purification

A total of 1 µg of the Cot-1 DNA was made up to 30 µl with 1 X Low TE buffer. The 30 µl DNA sample was combined with 10 µl of loading solution (Sage Science). A total of 30 µl of marker mix (Sage Science) was combined with 10 µl of loading solution. Samples were mixed thoroughly by vortexing and centrifuged briefly to collect. A 2 % Pippin Gel Cassette with ethidium bromide (Sage Science) for selecting fragments between 100 and

600 base pairs was removed from its foil packaging and the levels of buffer in the buffer reservoirs were inspected. Gel columns were inspected for breakages and the bottom of the cassette was inspected for bubbles. If no imperfections were observed, the cassette was placed into the Pippin Prep optical nest and adhesive strips were removed from the cassette. Buffer was removed from the elution modules and replaced with 40 µl of fresh electrophoresis buffer. The elution wells were sealed with the adhesive tape strips. A continuity test was performed to measure the current in each separation and elution channel to determine whether they were within the expected values for a successful run. Next, samples were loaded by removing 40 µl of buffer from the sample well and loading 40 µl of sample (or marker), taking care not to pierce the agarose with the pipette tip. A broad range protocol was carried to remove fragments less than 100 base pairs. After run completion, the adhesive strips were removed from the elution wells and 40 µl of size-selected eluate was removed by pipetting and this was transferred into a clean Eppendorf LoBind tube for storage at -20 ^{o}C before use.

2.8 DNA Fragment Size Analysis

2.8.1 Agilent High Sensitivity DNA Chip Assay

A High Sensitivity DNA chip assay (Agilent Technologies; CA, USA) was carried out on an Agilent 2100 Bioanalyzer machine to determine DNA fragment size.

All reagents were allowed to equilibrate to room temperature for 30 minutes before use. A High Sensitivity DNA chip was placed on the chip priming station. A total of 9 µl of gel-dye mix was added to the marked G well. The plunger was positioned at 1 ml and the chip priming station was closed. The plunger was pressed down and after 60 seconds, the plunger was released. After 5 seconds, the plunger was slowly lifted back to the 1 ml position. The chip priming station was opened and 9 µl of the gel-dye matrix was added to the other G wells. A total of 5 µl of High Sensitivity DNA marker was added to each sample well and the ladder well. In each sample well, 1 µl of sample was added and 1 µl of ladder DNA was added to the ladder well. A total of 6 µl of High Sensitivity DNA marker was added to any unused wells. The chip was vortexed for 60 seconds at 2,400 rpm. The chip was placed in the Bioanalyzer machine and run using Agilent 2100 software.

2.8.2 Agarose Gel Electrophoresis

The required quantity of agarose powder (Bioline, London, UK) was weighed and added to 100 ml of 1 X TBE buffer (Sigma) in a microwaveable glass flask. The solution was mixed gently and microwaved for approximately 3 minutes, with occasional stirring, until the agarose completely dissolved. The agarose solution was allowed to cool to approximately 50 °C. A 3 µl volume of GelRed (Biotium, CA, USA) was added and the solution was stirred gently to distribute the dye. The agarose solution was poured slowly into a gel tray with a well comb in place. The poured gel was left at room temperature for 20 to 30 minutes until it solidified completely. Once solidified the gel was placed into the electrophoresis tank (Alpha Laboratories, Eastleigh, UK) and 1 X TBE was added to the tank until the entire gel was covered. GelPilot ® 5X loading dye (Qiagen) was added to each DNA sample at a ratio of 1 volume of dye to 5 volumes of sample. A molecular weight ladder (1 kb ladder [NEB]) was loaded into the first lane of the gel and samples were carefully loaded into the other wells of the gel (one sample per well). The gel underwent electrophoresis at the required voltage using the BIO-RAD power pack 3000 for the required time. To visualise DNA fragments, the gel was placed in a BioDoc-IT2® Imager (UVP), the UV light applied and the gel image captured using VisionWorks ® touch software.

2.9 Methods for Hybridisation Experiments

2.9.1 DNA Digestion by EcoRI

Genomic DNA was digested using EcoRI enzyme (NEB) to generate shorter fragments for hybridisation experiments. For every 1 µg of human genomic DNA, 5 µl of EcoRI Buffer, 8 µl of *EcoRI* Enzyme and nuclease-free water to 50 µl were added. The reaction was incubated at 37 °C for 1 hour in a thermal cycler. The reaction was purified using the NEB Monarch PCR purification kit as described in Section 2.7.3.

2.9.2 Preparation of Hybridisation Buffers

The following hybridisation buffers and stocks were prepared or purchased: SSPE buffer (20X) (sodium chloride-sodium phosphate-EDTA) was prepared by dissolving 175.3 g of NaCl (Sigma), 27.6 g of $NaH_2PO_4 \bullet H_2O$ (Sigma) and 7.4 g of EDTA (Fisher Bioreagents) in 800 ml of *MiliQ* H_2O. The pH was adjusted to 7.4 with NaOH (Sigma). The volume was adjusted to 1 litre with *MiliQ* H_2O. The buffer was sterilised by autoclaving (Autoclave

Prestige Medical). A 10 % SDS stock solution was prepared by dissolving 10 g of SDS (BDH) in 80 ml of *MiliQ* H_2O. The volume was adjusted to 100 ml with *MiliQ* H_2O. The stock was sterilised by autoclaving. SSC buffer (20X) (Saline sodium citrate) was purchased from Sigma.

2.9.3 Preparing Streptavidin-Coated Magnetic Beads for DNA Capture

Binding and Washing Buffer (2X) containing 10 mM Tris-HCl (Fluka), 1 mM EDTA (Fisher Bioreagents) and 2 M NaCl (Sigma) was prepared and autoclaved (Autoclave Prestige Medical). In a separate autoclaved glass container, the 2X Binding and Washing buffer was diluted to 1X concentration and 0.05 % Tween-20 detergent (Sigma) was added. The buffer was mixed thoroughly. Dynabeads™ M-280 Streptavidin (Invitrogen) were resuspended in their vial by vortexing for 30 seconds. M-280 Streptavidin beads are uniform, superparamagnetic beads, 2.8 μm in diameter with a streptavidin monolayer covalently coupled to the surface. The required volume of streptavidin beads was pipetted into a 2 ml Eppendorf DNA LoBind tube. A total of 1 ml of the 1 X Binding and Washing Buffer was added to the beads and the mixture was vortexed for 5 seconds. The tube was placed on a magnetic bead separation stand for 1 minute until the solution was clear and the supernatant was discarded by pipetting. The tube was removed from the magnetic stand and the previous step was repeated after which the washed beads were resuspended in 2X Binding and Washing Buffer at twice the original volume.

2.9.4 Preparing Streptavidin-Coat Magnetic Beads for RNA Capture

For each reaction, 60 μl of Dynabeads M-280 were washed as described in Section 2.9.3. Two additional washing solutions were prepared: Solution A containing Diethyl pyrocarbonate (DEPC)-treated 0.1 M NaOH (Sigma) and DEPC-treated 0.05 M NaCl (Sigma) and Solution B containing DEPC-treated 0.1 M NaCl (Sigma). DEPC was sourced from Sigma. Both solutions were autoclaved (Autoclave Prestige Medical). After the initial washing steps, the beads were washed twice in Solution A for 2 minutes using the same volume as the initial volume of beads taken from the vial. Next beads were washed once with Solution B using the same initial volume. Beads were then resuspended in Solution B using the same volume as initial volume.

2.9.5 Hybridisation Approach for *Alu* RNA: DNA

A drop of mineral oil (Sigma) was added over reactions to prevent evaporation. Denaturation was carried out in a heat block (Grant) at 95 °C for 5 minutes. After denaturation, the tubes were quickly transferred to a heat block at 65 °C. A total of 5 µl of 20 X SSC warmed to 65 °C was added to the reaction and a hybridisation temperature of 65 °C was sustained for 1 hour.

2.9.6 Bead Capture Approach for *Alu* RNA: DNA

Each completed hybridisation reaction was made up to 60 µl with nuclease-free water and added to 60 µl of washed beads. The hyb-bead mixture was incubated (*Whirlmixer*®) for 30 minutes at room temperature. The tubes were centrifuged briefly (1,500 g for 3 seconds) and then incubated on a magnetic stand for 15 minutes. The "microbial" supernatant was harvested and stored. Beads were washed with twice with 200 µl of 1 X Low TE and the supernatant discarded. The beads were resuspended in 15 µl of nuclease-free water and denatured at 95 °C for 5 minutes. After denaturation the tube was incubated on a magnetic stand for 5 minutes. The "human" supernatant was harvested and stored. A purification step for the "human" supernatant was not performed. The "microbial" supernatant was purified using the NEB Monarch PCR purification kit as described in Section 2.7.3 and eluted in 15 µl of nuclease-free water. The purified "microbial" supernatant and "human" supernatant were run on a 1.2 % agarose gel at 120 V for 1 hour and visualised as described in Section 2.8.2.

2.9.7 Quantitative PCR for Bacterial and Human DNA

To quantify bacterial and human DNA, SYBR Green qPCR was carried out. Bacterial 16S rRNA gene standards were diluted from a previously prepared stock of *Vibrio natregens* full length 16S rRNA clones. Standards were prepared for 10^8, 10^7, 10^6, 10^5, 10^4 and 10^3 copies. Human *GAPDH* gene standards were prepared by diluting human genomic DNA. Standards were prepared for 10^5, 10^4, 10^3, 10^2 and 10 copies. Reactions were prepared for bacterial and human qPCR using 2X SYBR Fast qPCR Master Mix (KAPA BioSystems) as follows:-

SYBR FAST qPCR Reaction – components for one reaction

Component	**Volume µl (X1 reaction)**
SYBR FAST qPCR Master Mix (2X)	7.5
Forward Primer (10 µM)	0.3
Reverse Primer (10 µM)	0.3
Nuclease-free water	1.9
Template DNA	5

A no template control was included for each run. qPCR was carried out using a ViiA7 Real-Time PCR System with ViiA7 Software Base v1.1. Cycling conditions were as follows:- 95 °C for 5 minutes followed by 40 cycles of 95 °C for 20 seconds, 50 °C for 30 seconds and 72 °C for 30 seconds. Melt conditions were carried out using standard default parameters. Forward and reverse primer sequences for the 16S rRNA gene were 520F (*5*'- *AYTGGGYDTAAAGNG* -3') and 802R (5'- *TACNVGGGTATCTAATCC* -3') (Kozich *et al.,* 2013). Forward and reverse primer sequences for the *GAPDH* human gene were GAPDH-F (5'– *TACTAGCGGTTTTACGGGCG* -3') and GAPDH-R (5'- *CGAACAGGAGGAGCAGAGAG* -3') (primers designed in-house by members of the Genomic Medicine Section team). Primers were sourced from Eurofins (London, United Kingdom).

2.10 Metagenomic Sequencing

2.10.1 Metagenomic Library Preparation

Library preparation was carried out using the Nextera ® XT DNA Library Preparation Kit (Illumina®, CA, USA). The kit is optimised for 1 ng of input DNA. DNA was first fragmented and adaptor sequences added onto the DNA template by tagmentation. A total of 10 µl of Tagment DNA buffer was added to 5 µl (1 ng) of normalised genomic DNA (0.2 ng/µl) and mixed by pipetting. Next 5 µl amplicon Tagment Mix was added and mixed by pipetting. The plate was centrifuged at 280 × g at 20 °C for 1 minute. The reaction was placed on the thermal cycler and the tagmentation program was run as follows:- one cycle of 55 °C for 5 minutes followed by a hold step at 10 °C. After tagmentation, neutralise tagment buffer (5 µl) was added to the tagmented sample and mixed by pipetting. The plate was centrifuged at 280 × g at 20 °C for 1 minute after which the reaction was incubated at room temperature for 5 minutes. Library amplification was then carried out by adding 5 µl each of Index 1 (i7) and Index 2 (i5) adapters to the tagmented samples followed by 15 µl of Nextera PCR master mix with thorough mixing by pipette. The plate was centrifuged at 280 × g at 20 °C for 1 minute. The plate was then placed on the thermal cycler and the following program run: 72 °C for 3 minutes, 95 °C for 30 seconds followed by 12 cycles of 95 °C for 10 seconds, 55 °C for 30 seconds and 72 °C for 5 minutes followed by a hold step at 10 °C.

2.10.2 Purification of Metagenomic Libraries

Libraries were purified using AMPure XP beads. A total of 30 µl of AMPure XP beads were added to each library. The plate was then shaken at 1,800 rpm for 2 minutes at room temperature followed by a further incubation for 5 minutes at room temperature without shaking. The plate was placed on a magnetic plate stand for 2 minutes until the liquid cleared. The supernatant was carefully removed by pipetting and discarded. The beads were then washed twice with 200 µl 80 % ethanol and, after the final wash, any residual 80 % ethanol was removed by pipetting. Beads were left to air-dry on the magnetic stand for 15 minutes after which the plate was removed from the stand and 52.5 µl of re-suspension buffer (RSB) was added to each library. The plate was shaken at 1,800 rpm for 2 minutes and incubated at room temperature for a further 2 minutes without shaking. The plate was once more placed on a magnetic plate stand for 2 minutes after which 50 µl of the clear supernatant was transferred into a new plate.

2.10.3 Fragment Size Analysis of Metagenomic Libraries

On an Agilent Technology 2100 Bioanalyzer, 1 µl of each library was run using a High Sensitivity DNA chip to measure the dsDNA fragment sizes as described in Section 2.8.1.

2.10.4 Quantification of Metagenomic Libraries

Libraries were quantified by the PicoGreen dsDNA assay as described in Section 2.6.1.

2.10.5 Pooling and Quantification of Pooled Libraries by qPCR

After quantification, a total of 20 ng of each library was pooled into a single 1.5 ml microcentrifuge tube. Two different assay kits for the quantification of pooled libraries by qPCR were used in this thesis. The assay kit used is referenced accordingly in each relevant chapter.

2.10.5.1 KAPA SYBR® FAST qPCR Library Quantification Kit

Quantitative PCR was carried out on the pooled library sample using the KAPA SYBR ® FAST qPCR Library Quantification Kit (KAPA BioSystems Limited, London, UK) for Illumina platforms. The pooled library sample was diluted in PCR-grade water to 1/1,000 and additionally to 1/2,000, 1/4,000 and 1/8,000. To each required well of a MicroAMP Fast 96-well Reaction Plate (0.1 ml), the following was added: 12 µl KAPA SYBR® FAST qPCR Master Mix, 4 µl of PCR-grade Water, 4 µl of Standard (X6)/Non-Template Control/ Sample. The plate was sealed with MicroAmp Optical Adhesive Film (Thermo Fisher Scientific). The reaction was run on an Applied Biosytems Viia 7 under the following cycling conditions: one cycle of 90 °C for 3 minutes followed by 35 cycles of 95 °C for 20 seconds, 55 °C for 30 seconds and 72 °C for 30 seconds, followed by one cycle of 72 °C for 30 minutes and a melt curve step. Data acquisition was carried out at the 72 °C extension step.

2.10.5.2 JetSeq ™ Library Quantification Lo-ROX Kit

Quantitative PCR was carried out on the pooled library sample using the JetSeq ™ Library Quantification Lo-ROX Kit (Bioline). The pooled library sample was diluted in JetSeq Dilution Buffer to 1/50 by adding 1 µ l of pooled library to 49 µl of water. The 1/50 sample was then diluted to 1/5,000 in JetSeq Dilution Buffer and additionally to 1/10,000, 1/20,000, 1/40,000 and 1/80,000. To each required well of a MicroAMP Fast 96-well Reaction Plate (0.1 ml), the following was added: 10 µl JetSeq FAST Lo-Rox Mix, 5 µl of JetSeq Primer Mix and 5 µ l of Standard (x6)/Non-Template Control/Sample. The plate was sealed with a MicroAmp Optical Adhesive Film (Thermo Fisher Scientific). The reaction was run on an Applied Biosystems Viia7 under the following conditions: one cycle of 95ºC for 2 minutes followed by 35 cycles of 95 °C for 5 seconds and 60 °C for 45 seconds and finally a melt curve step. Data acquisition was carried out at the 60 °C extension step.

2.10.6 Quality and Size of Pooled Libraries

On an Agilent Technology 2100 Bioanalyzer, 1 µl of undiluted pooled library was run using a High Sensitivity DNA chip to check library integrity and measure dsDNA fragment sizes as described in Section 2.8.1.

2.10.7 Sequencing of Metagenomic Libraries

A MiSeq v3 Reagent Cartridge (Illumina) containing MiSeq reagents was removed from -20 °C storage and allowed to thaw for approximately 60 to 90 minutes in a water bath at 25 °C. Once the reagents had reached room temperature, the cartridge was removed from the water bath and dried. Reagent positions were inspected to ensure no precipitates were present. The cartridge was tapped gently to remove any air bubbles.

A MiSeq flow cell (Illumina) was carefully removed from the storage buffer container. The flow cell was thoroughly washed with *MiliQ* H_2O to remove any salt crystals. The flow cell was dried using a lint-free lens cleaning tissue (Fisher Scientific). The flow cell was then cleaned with an ethanol tissue (Bollé Safety; Villeurbanne, France) to make sure the glass was free of fingerprints and smudges. Excess ethanol remaining on the flow cell was removed with a lens cleaning tissue. The flow cell was then placed on the flow cell stage of the MiSeq instrument (Illumina).

Dilution and denaturation of the sample and control was carried out as a final step before initiating the MiSeq run. The pooled libraries sample and a PhiX v3 control (Illumina) were denatured and diluted to an 8 pM input concentration. The sequencing reaction was carried out using MiSeq v3 Reagent Kit (Illumina). The PhiX control was prepared by adding 2 µl of PhiX (10M) to 3 µl EBT buffer after which 5 µl of 0.2 N NaOH (Sigma) was added. The mixture was vortexed and centrifuged briefly then incubated at room temperature for exactly 5 minutes. To the solution, 990 µl of ice cold HT1 buffer was added resulting in a 20 pM stock. A 8 pM solution of PhiX was made by adding 600 µl of ice cold HT1 buffer to 400 µl of the 20 pM PhiX. To denature the pooled libraries sample, 10 µl of 0.2N NaOH was added to 10 µl of the library at room temperature. The liquid was vortexed and centrifuged briefly and then incubated at room temperature for 5 minutes. A total of 980 µl of ice cold HT1 buffer was then added. To dilute the pooled libraries samples further to 8 pM, 101 µl of denatured library was added to 898 µl of ice cold HT1 buffer. Finally, to spike in 5% PhiX, 50 µl of the 8 pM prepared PhiX control was added to 950 µl of 8 pM library. A total of 600 µl of the resulting PhiX spiked library was added to the reagent cartridge sample well. The reagent cartridge was loaded onto the MiSeq platform (Illumina) and the sequencing run was carried out as per the manufacturer's guidelines.

2.11 Target Capture for *Legionella pneumophila*

Target capture for the enrichment of *L. pneumophila* species was carried out directly on clinical and environmental DNA extracts.

2.11.1 Database Preparation

A database of *L. pneumophila* genomes was prepared. The database included completed *L. pneumophila* genomes deposited in the NCBI RefSeq ftp server (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/) and *L. pneumophila* PacBio genomes from the Sanger NCTC 3000 project (sanger.ac.uk/resources/downloads/bacteria/nctc/). Two unassembled *L. pneumophila* genomes deposited in the Sanger 3000 server were *de novo* assembled using the `BugBuilder` program (Version 1.0.3b1) (Abbott, 2017). `BugBuilder` is a computational pipeline that assembles and annotates raw sequencing files produced from a number of sequencing platforms as well a hybrid of platform outputs.

Firstly, the PacBio bas.h5 files were converted to FASTQ format using `pbh5tools` (Version 0.8.0) (PacificBiosciences, 2014). The `bash5tools` python script from `pbh5tools` was called to extract read sequences and quality values from the raw bas.h5 files and create FASTQ files. The output FASTQ was then assembled using `BugBuilder`. See Appendix Section 9.1.1 for the full code. Plasmid sequences were removed from all assemblies in the database. The full list of *L. pneumophila* genomes used for bait design is outlined in Appendix Section 9.2.

2.11.2 Bait Design

Biotinylated RNA oligonucleotide baits were designed by Dr. Sunando Roy at the Pathogen Genome Unit, UCL based on the created *Legionella pneumophila* genome database (Section 2.11.1). Baits were designed with a 2x tiling density. This signifies that baits overlap by 50% or two baits cover each base at each interval. The baits were designed with a length of 120 nucleotides as default.

2.11.3 Library Preparation

The SureSelect^{XT} Low Input Target Enrichment System (Agilent) was used for the preparation of target-enriched Illumina paired-end sequencing libraries. The protocol was carried out by the Pathogen Genome Unit team at UCL using an automated system.

2.11.4 Pooling and Sequencing

Library pooling, quantification by qPCR and sequencing was carried out by me at the Genomic Medicine Section, NHLI, Imperial College London. Libraries were pooled for multiplexed sequencing by adding 20 ng total of each library to a 1.5 ml microcentrifuge tube. The pooled libraries were quantified by the JetSeq qPCR assay (Bioline) as described in Section 2.10.5.2. The DNA fragment size of the pooled libraries was then assessed using an Agilent High Sensitivity DNA chip on an Agilent 2100 Bioanalyzer as described in Section 2.8.1. Libraries were diluted and denatured and spiked with 5 % PhiX and Illumina paired-end sequencing was carried out on an Illumina MiSeq as described in Section 2.10.7.

2.12 Metagenomic Data Quality Control and Pre-Processing Pipeline

2.12.1 De-multiplex Paired-End Data Files

Barcode reads from the MiSeq sample sheet were joined together for each sample. The barcodes were then appended to the reads and the reads were processed using the FASTX barcode splitter tool from the `FASTX-Toolkit` (Version 0.0.14) (Gordon, 2009) to split the libraries. Once libraries were de-multiplexed, the appended barcodes were removed using the `BBDuk` script from the `BBTools` suite (Bushnell, 2014) implementing the `ftl` ("force trim left") option. See Appendix Section 9.1.2 for the full code.

2.12.2 Data Quality Control

Once paired-end data was de-multiplexed, the FastQC program (Version 0.3.11) (www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to assess the quality of the sequenced data and to ascertain if the data may be problematic for downstream analyses. FastQC modules report basic statistics, sequence length distribution, per base sequence quality and per tile sequence quality via a phred quality score. A phred score is a number assigned to a nucleotide giving the probability that the base call is erroneous on a logarithmic scale. For example, if a base is assigned a phred score of 20, this means that there is a 1 in 100 probability that the base was called incorrectly. FastQC also determines if adapter sequence was read through during sequencing or if undetermined nucleotides ("N's") are overrepresented in the reads.

For the analyses of metagenomic data in this thesis, the percentage GC content and overrepresented kmer modules were disregarded as the nature of metagenomic data does not support an even percentage GC or kmer distribution. See Appendix Section 9.1.3 for the full code.

2.12.3 Adapter Trimming, Quality Trimming and Quality Filtering

Adapter and Quality Trimming/Filtering was carried out using the BBDuk script from the BBTools suite (Bushnell, 2014). Adapter sequences were trimmed from the right of the sequence using a kmer size of 15 for the Nextera sequencing adapter (CTGTCTCTTATACACATCT) and a kmer size of 33 for SureSelect TruSeq sequencing adapters (AGATCGGAAGAGCACACGTCTGAACTCCAGTCA for read 1 and AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT for read 2). Reads were quality trimmed from the right and bases with a phred score less than 20 were removed. The "trim by overlap" (tbo) and "trim paired end" (tpe) options were used to remove adapter that the ktrim module might have missed. A further 10 bases were force trimmed from the right after quality and adapter trimming. Reads with a length of less than 50 base pairs were removed from the dataset. Histograms pertaining to various aspects of the reads after trimming – base frequency (bhist), quality scores (qhist), average quality (aqhist) and length (lhist) were generated. See Appendix Section 9.1.4 for the code.

2.12.4 Removal of PhiX Reads

The PhiX v3 illumina control was removed from the data by mapping the sample reads to the PhiX genome reference sequence using the `BBduk` script from the `BBTools` suite (Bushnell, 2014). The PhiX reference sequence can be found at the following link: https://www.ncbi.nlm.nih.gov/nucleotide/NC_001422. A kmer size of 31 with a hamming distance of 1 was used for mapping reads to the PhiX genome. A descriptive statistics file of reads mapping to the PhiX and removed was generated. See Appendix Section 9.1.5 for the full code.

2.12.5 Removal of Human Genome Reads

Human genome reads were removed from the data by mapping the sample reads to the human hg38 genome sequence using the `BBMap` script from the `BBTools` suite (Bushnell, 2014). Reference human genomes can be found at the following link: https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/. Coverage statistics (`covstats`) of the mapped sample reads were then generated. See Appendix Section 9.1.6 for the full code.

2.13 Tools for Metagenomic Data Analysis

2.13.1 Taxonomic Classification

Taxonomic classification was carried out using `Centrifuge` (Version 1.0.3) (Kim *et al.,* 2016). `Centrifuge` is a microbial classifier specifically for the classification of metagenomic data that uses a scheme based on the Burrows-Wheeler transform (BWT) and the Ferragine-Manzini (FM) index. The complete bacterial genomes database was downloaded from NCBI RefSeq (O'Leary *et al.,* 2016) on July 1st 2018 and indexed by Dr Lesley Hoyles. Only one distinct classification was assigned to each read. A Kraken-style report was then generated from the centrifuge results file. See Appendix Section 9.1.7 for the full code. A report containing species assignments only was extracted using a custom R script written by Dr Lesley Hoyles (private communication).

2.13.2 *In silico L. pneumophila* Sequence Type Analysis

In silico sequence-based typing (SBT) analysis based on the traditional *L. pneumophila* SBT scheme (Gaia *et al.,* 2005, Ratzow *et al.*, 2007, Mentasti *et al.,* 2014) was carried out to determine the limit of detection of reliable sequence type information. The scheme is based on the analysis of 7 housekeeping and virulence loci: – *flaA*, *pilE*, *mip*, *mompS*, *asd*, *neuA* and the *neuA* homolog, *neuAh*. All *L. pneumophila* allele sequences and allelic profiles were retrieved from the *L. pneumophila* SBT PHE (Health Protection Agency HPA) website (http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php) and used as database and definitions parameters in the `SRST2` (Short Read Sequence Typing for Bacterial Pathogens) program (Version 0.2.0) (Inouye *et al.,* 2014). `SRST2` takes raw Illumina sequenced reads as input, aligns the reads to a FASTA file of all allele sequences and reports the presence of sequence types defined in a file containing the sequence type profiles as combinations of alleles. See Appendix Section 9.1.8 for the full code.

2.13.3 Identification of Mixed *L. pneumophila* Strains

Strain-level analysis using the `StrainEst` program (Version 1.2.2) (Albanese *et al.,* 2017) was carried out to identify either single or mixed *L. pneumophila* strains. A *L. pneumophila* database was created by first downloading the complete list of available bacterial genomes from NCBI RefSeq ftp server (accessed on October 4th, 2018). *L. pneumophila* sequence types were assigned to each downloaded genome based on the ESGLI (European Study Group for *Legionella* Infections) scheme (Gaia *et al.,* 2005, Ratzow *et al.,* 2007, Mentasti *et al.,* 2014). The `mlst` program (Version 2.15.2) (Seeman, 2014) was used to scan the complete and draft genome assemblies against the traditional *L. pneumophila* SBT scheme and a sequence type was assigned to each genome. For a full list of the *L. pneumophila* genomes and their sequence type see Appendix Section 9.3. The `StrainEst` program (Version 1.2.2) (Albanese *et al.,* 2017) was used to align the RefSeq genomes against the reference genome (*L. pneumophila* Philadelphia-1 (ST36) [https://www.ncbi.nlm.nih.gov/nuccore/AE017354]). Ambiguous mappings in alignments were discarded and an output alignment was produced. Positions that were variable in the aligned genomes were then recorded in a single nucleotide variant (SNP) matrix. Then, the number of SNP sites was calculated pairwise between sequences and hierarchical clustering was performed (threshold of 99 %). A distance matrix was then computed. The genome database was reduced whilst maintaining sufficient diversity within the retained genomes to allow SNP detection from the computed matrix. The reduced database was indexed using Bowtie2 (Version 2.3.2) (Langmead *et al.,* 2012). Metagenomes were aligned to the reduced database using Bowtie2. The SAM file containing the mapped reads was converted to a BAM file using SAMtools (Version 1.8) (Li *et al.,* 2009). The BAM file was sorted and indexed. The StrainEst estimation model was used to infer the relative abundance of *L. pneumophila* strains within the metagenomes by Lasso regression. Coverage depth and maximum identity thresholds parameters are specified in the relevant results chapters. See Appendix Section 9.1.9 for the full code.

2.14 Oxford Nanopore Library Preparation and Sequencing

2.14.1 Purification during ONT Library Preparation

AMPure XP beads were resuspended for use by vortexing. From the resuspended beads, the required volume (see each Section of 2.14 hereafter) was added to the reaction. Tubes were mixed gently and incubated with rotation at room temperature for 5 minutes. Following a brief centrifugation step (1,500 x g for 3 seconds), tubes were placed on a magnetic stand until the beads were pelleted. Keeping the tubes on the magnetic stand, the supernatant was removed by pipetting and discarded. Beads were washed with 200 µl of freshly prepared 75 % ethanol without disturbing the pellet. The supernatant was discarded, and the wash step repeated. After removing residual ethanol, the pellet was allowed to air dry for 30 seconds. Tubes were removed from the rack and the pellet resuspended in the required volume of nuclease-free water and incubated for 2 minutes at room temperature. Beads were pelleted on a magnet and the required volume of the clear, colourless eluate removed and stored in a new 1.5 ml Eppendorf DNA LoBind tube.

2.14.2 FFPE DNA Repair

Formalin-fixed paraffin embedded (FFPE) DNA repair was carried to repair nicks and breaks in DNA strands. For each individual reaction, a specific concentration of DNA (as detailed in Chapters 4 and 6) was added to 6.5 µl of FFPE DNA Repair Buffer (NEB) and 2 µl of FFPE DNA Repair Mix (NEB) for a total volume of 60 µl. The tube was mixed gently and centrifuged briefly at 1,500g for 3 seconds. Reactions were incubated in a thermal cycler at 20 °C for 20 minutes. The reaction was purified with 62 µl of AMPure XP beads as described in Section 2.14.1 and DNA was eluted in 45 µl of nuclease-free water and stored in a clean 1.5 ml Eppendorf DNA LoBind tube.

2.14.3 End-repair

End-repair reactions were carried out by combining 45 µl of the FFPE-repaired DNA with 7 µl of Ultra II End-prep reaction buffer (NEB), 3 µl of Ultra II End-prep enzyme mix (NEB) and 5 µl of nuclease-free water. The reaction was mixed gently, transferred to 0.2 ml PCR tubes and incubated in a thermal cycler at 20 °C for 30 minutes followed by 65 °C for 30 minutes. Reactions were kept on ice for 30 seconds and then purified as described in Section 2.14.1 using 60 µl of AMPure XP beads. Purified DNA was eluted in 31 µl of nuclease-free water and stored in a clean 1.5 ml Eppendorf DNA LoBind tube.

2.14.4 PCR Adapter Ligation

PCR adapters were ligated on to the end-repaired DNA template by combining the 30 µl of end-repaired DNA, 20 µl of PCR adapters (Oxford Nanopore Technology [ONT] SQK-LSK108 kit), 40 µl of Ultra II Ligtion Master Mix (NEB) and 1 µl of Ultra II Ligation Enhancer (NEB). Reactions were mixed gently, transferred to 0.2 ml PCR tubes and incubated in a thermal cycler at 20 °C for 20 minutes. Purification was carried out as described in Section 2.14.1 using 91 µl of AMPure XP beads and purified DNA was eluted in 41 µl of nuclease-free water.

2.14.5 Post-PCR End-repair and Adapter Ligation

Adapter ligation on end-repaired DNA templates was carried out using 20 µl of AM1D adapter mix (ONT SQK-LSK108 kit):-

Post-PCR Adapter Ligation – components for one reaction.

Component	**Volume** µl **(X1 reaction)**
End-repaired DNA	30
AM1D (ONT SQK-LSK108 kit)	20
Ultra II Ligation Master Mix (NEB)	40
Ultra II Ligation Enhancer (NEB)	1

Purification was carried out using 40 µl of AMPure XP beads as described in Section 2.14.1. After a 5 minutes incubation with rotation at room temperature, beads were pelleted and 140 µl of ABB Buffer (ONT) was added. The tube lid was closed and the beads were resuspended with the buffer by flicking the tube. The tube was returned to the magnetic rack, beads were allowed to pellet and the supernatant was removed and discarded. This step was repeated. Beads were then resuspended in 15 µl of EBB buffer (ONT) and allowed to incubate at room temperature for 10 minutes. Beads were pelleted by placement on the magnetic rack and 15 µl of purified library were removed and stored in clean Eppendorf tubes.

2.14.6 Oxford Nanopore Minion Sequencing

At room temperature a flow-cell priming mix was prepared containing 576 µl of Running Fuel Buffer (RBF) (ONT SQK-LSK108 kit) and 624 µl of nuclease-free water. A P1000 pipette was set to 200 µl and approximately 20 µl of buffer was carefully withdrawn from the priming port. A total of 800 µl of priming mix was loaded into the flow cell via the priming port (without the introduction of air bubbles) and left for 5 minutes.

During these five minutes the libraries for the test and control were each prepared as follows:-

Oxford Nanopore Library for Loading

Component	**Volume** µl **(X1 library)**
RBF (ONT)	35
Library-loading Beads (LLB)	25.5
Nuclease-free water	2.5
DNA Library	12

The SpotON sample port cover was lifted. A total of 200 µl of priming mix was added to the priming port and 75 µl of the library was added to the flow cell via the SpotON port in a dropwise fashion. The SpotON samples port was replaced and the priming port was closed.

2.15 16S rRNA Gene Library Preparation and Sequencing

This section describes the protocol for dual index 16S rRNA gene-based bacterial community profiling from clinical samples on the Illumina MiSeq.

2.15.1 Quadruplicate 16S rRNA Gene PCR

Quadruplicate PCR was carried out on samples of interest, a negative control from extraction, a water control and an in-house-prepared mock community.

This was performed to amplify the 16S rRNA gene and to add indexing barcodes to samples for sequencing. A total of 5 µl of each indexed 16S forward sequencing primer (1.5 µM) and 5 µl of each indexed 16S reverse sequencing primer was pipetted into the required well of a PCR strip tube. A total of 1 µl (4 ng) of each sample/control was added to four corresponding PCR tubes of the strip. Q5 Master Mix was diluted with Qiagen water (50:6) and 14 µl of diluted Q5 Master Mix was added to each well. Strip tubes were sealed and centrifuged briefly. The tubes were placed on a PCR block and the following cycling conditions were carried out: 95 °C for 2 minutes, 34 cycles of 95 °C for 20 seconds, 50 °C for 20 seconds and 72 °C for 5 minutes and hold at 10 °C.

2.15.2 Sample Pooling and Contamination Check

A PCR contamination check was carried out by preparing a 1.2 % agarose gel (1.2 g of agarose in 100 ml of 1 X TBE) as described in Section 2.8.2. All samples were loaded into the wells of the gel and 5 µl of 100 base pair ladder was loaded next to the last sample. Electrophoresis was carried out at 120 V for 40 minutes. The gel was visualised under UV to confirm correct amplicon size in samples and positive control and no amplification in negative control. Corresponding replicate PCR reactions were pooled and a gel was prepared, run and visualised as described above.

2.15.3 Sample Purification, DNA Quantification and Equimolar Pooling

A total of 80 µl of pooled sample/control was transferred to a round bottom plate (Thermo Scientific) and samples were purified using AMPure XP beads at a 0.7:1 µl beads:PCR product. Purification was carried out as described in Section 2.7.1. The pellet was resuspended in 31 µl of Low EDTA TE buffer and incubated for 5 minutes at room temperature. It was then placed on a magnetic rack and the liquid was removed and retained in 1.5 ml LoBind DNA Eppendorf tubes. Purified amplicons were quantified by

PicoGreen as described in Section 2.6.1. After sample quantification, pooling was carried out by adding 20 ng total of each sample to a 1.5 ml LoBind DNA Eppendorf tube. The pooled library was purified again by AMPure XP magnetic beads at a 0.7:1 µl beads:library volume as described in Section 2.7.1. The pellet was resuspended in 31 µl of Low EDTA TE buffer, incubated for 5 minutes at room temperature, placed on a magnetic stand and 30 µl of liquid was removed and retained in a 1.5 ml LoBind DNA Eppendorf tube.

A gel purification of the concentrated pooled library was carried out by preparing a 1.8 % gel (150 ml) with 7.5 µl Gel Red dye as described in Section 2.8.2. A total of 10 µl of loading dye was mixed with the 30 µl library. The whole library was carefully loaded into a lane on the gel and 5 µl of 100 bp ladder was added to a nearby well. The gel was run at 80 V for 65 minutes. The gel was visualised under UV light in a dark room and the bright band at approximately 350 bp was excised using a gel-cutting tip on a Gilson P1000 pipette. Next, the gel slice was purified using NEB Monarch DNA Gel Extraction kit. The gel slice was weighed and dissolved in Gel Dissolving Buffer at a 1:4 gel weight:buffer volume. It was then incubated in an Eppendorf on a heating block at 50 °C for 10 minutes. The tube was centrifuged briefly every 1 – 2 minutes. The sample was loaded on to a spin column and centrifuged for 1 minute at 13,000 x g. The flow-through was discarded and 200 µl of DNA wash buffer was applied, centrifuged for 1 minute and flow-through discarded. The last step was repeated again. The column was transferred to a clean 1.5 ml LoBind DNA Eppendorf tube. A total of 30 µl of Low EDTA TE buffer was added to the spin column, incubated at room temperature for 1 minute and then centrifuged for 1 minute. The flow-through was collected.

2.15.4 Library Quality Check, Quantification and Sequencing

At this stage, the purified library should contain a single peak of approximately 350 bp. To assess this, a total of 2 µl of the purified pooled library was diluted 1 in 10 and 1 µl of the diluted library was run on a Bioanalyzer High Sensitivity DNA chip as described in Section 2.8.1. A qPCR was carried out on the pooled library using the JetSeq Library Quantification Kit as described in Section 2.10.5.2. The final library was spiked with 20 % PhiX. Before loading the library to the Illumina MiSeq V2 reagent cartridge, 4 µl of each sequencing primer (Kozich *et al.,* 2013) were loaded into wells labelled 12, 13 and 14. Sequencing was carried out on an Illumina MiSeq as described in Section 2.10.7.

Chapter 3.

Evaluating a Metagenomic Sequencing and Analysis Pipeline for *Legionella* Species Detection

3.1 Introduction

Metagenomic sequencing in the context of *Legionella spp.* detection may be useful in determining the epidemiological type of strains in a timely manner since it excludes an isolation step of at least 3 days which is required for *Legionella* growth and time to carry out sequence-based typing. It may also be useful when an isolate fails to grow and in determining if cases or outbreaks are caused by a mixture of *L. pneumophila* sequence types or *Legionella* species (Coscollá *et al.,* 2014, Wewalka *et al.,* 2014). There is however no standardised approach for metagenomic sequencing and the analysis of metagenomic data for public health purposes. Moreover, the complexity of the approach increases when dealing with low microbial biomass specimens such as sputum and water, the standard samples in which the presence of *Legionella* is investigated.

In considering a metagenomics sequencing approach, a number of steps can be taken to validate a laboratory workflow before sequencing real specimens. First, the accuracy of the sequencing approach can be investigated to ensure technical biases and errors are not introduced during the library preparation and sequencing steps. A number of studies have highlighted the importance of validating metagenomic laboratory workflows with mock communities (Bowers *et al.,* 2015, Jones *et al.,* 2015, Schlaberg *et al.,* 2017[a]). A mock community, a mixture of known bacterial species at defined proportions/copy numbers, is not only important for the validation of library preparation and sequencing protocols but also for the *in silico* validation of downstream data analysis tools.

The analytic sensitivity of pathogen detection can be investigated to determine the limit of detection of the pathogen. Sensitivity of detection is influenced by many variables such as genome size, depth of sequencing, and the presence of host DNA or microbial flora that overwhelm pathogen detection. Additionally, in cases where the pathogen constitutes an extremely low proportion of the original specimen, the signal from contaminating microorganisms alone may overwhelm pathogen signal. Contaminating microorganisms may be introduced from the environment or reagents during extraction, sample manipulation and library preparation (Salter *et al.,* 2014). Ideally to increase sensitivity of detection, large quantities of the specimen of interest should be sampled however there are technical constraints to achieving this. For example, it may be difficult for an ill patient to produce a sufficient volume of sputum. Another approach would be to sequence at a high depth of coverage although this approach introduces additional costs.

Once a species is identified, disentangling the species to the strain-level is vital to pathogen surveillance networks for the epidemiological resolution of cases.

For *L. pneumophila* surveillance, a 7-loci sequence-based typing (SBT) scheme exists to determine the sequence type of a given specimen (Gaia *et al.,* 2005, Ratzow *et al.*, 2007, Mentasti *et al.,* 2014). This scheme can be validated *in silico* on sequenced mock communities before applying it to metagenomes from real specimens. Due to the fragmented nature of metagenomes, it may not be possible to determine a full or even partial sequence type based on the 7-loci scheme. A number of tools have been described recently for the analysis of metagenomic data to the strain-level based on coding gene presence/absence, marker gene SNPs and core genome SNPs (Hong *et al.,* 2014, Ahn *et al.,* 2015, Luo *et al.,* 2015, Nayfach *et al.,* 2016, Scholz *et al.,* 2016, Quince *et al.,* 2017). These tools rely on the construction of species databases, which are incomplete strain-level representations of the species, however they may be useful in identifying a closely related sequence type and therefore resolving the phylogeny of the case in question.

3.2 Aims and Objectives

1. Accuracy: Determine the accuracy of sequencing mock communities composed of defined proportions of bacterial and human DNA as a means of pipeline validation.
    a. Design, prepare and sequence mock communities composed of *Legionella* strains and other microbial species in a human DNA background.
    b. Determine the accuracy of sequencing microbes at their actual versus sequenced proportions.

2. Sensitivity: Evaluate the sensitivity of detection of *L. pneumophila* in a human DNA background.
    a. Design, prepare and sequence sensitivity tests composed of human DNA spiked with a *L. pneumophila* strain at high to low proportions.
    b. Investigate relative abundance of *L. pneumophila* in the sample versus percentage coverage of the reference genome and depth of coverage.
    c. Determine the proportional cut-off relative to the quantity of background DNA in generating reliable *in silico* traditional sequence type information.

3. Identification of Mixed Strains: Investigate the ability of a bioinformatic workflow to identify mixed *L. pneumophila* strains:
    a. Prepare a database of publicly available complete and draft *L. pneumophila* genome assemblies representing as many sequence types as possible.
    b. Build a core single nucleotide polymorphism (SNP) matrix from the genomes.
    c. Map metagenomes to the *L. pneumophila* database and extract SNPs for strain prediction.

4. Host DNA Contamination: Investigate the influence of host contaminating human DNA on the ability to detect and sequence *Legionella* from sputum DNA extracts.
    a. Sequence known *L. pneumophila* positive and negative sputum DNA extracts.
    b. Determine what proportion of total sequenced reads are human DNA.
    c. Determine what proportion of total sequenced reads are microbial in origin.
    d. Investigate the presence of *Legionella* reads in the sequenced sample.

3.3 Methods

3.3.1 Material Preparation for Mock Communities and Sensitivity Tests

The following genomic material was used in the preparation of mock communities and sensitivity tests: human genomic DNA, *L. pneumophila* Philadelphia-1, *L. pneumophila* France 5811, *L. longbeachae*, *S. pneumoniae*, *H. influenzae* and *V. dispar*. The strain designation, source and ethical considerations regarding the genomic material is detailed in Chapter 2, Section 2.4. Due to insufficient concentrations, isothermal whole genome amplification (WGA) was carried out on the *L. pneumophila* and *L. longbeachae* strains as described in Section 2.4.1. Whole genome amplified products were purified using AMPure XP beads as described in Section 2.7.1. All DNA concentrations were measured by the PicoGreen dsDNA Quantitation as described in Section 2.6.1. The amplified samples were stored in a freezer at -20°C until required.

3.3.1.1 Mock Community Design

A total of nine mock communities were prepared by mixing bacterial type strain DNA at varying proportions with human DNA (**Table 3.1** and **Figure 3.1**).

Copy number was calculated using the following formula:

$$\textbf{No. of copies} = (\textbf{amount (in ng)} * [6.022 \times 10^{23}]) / (\textbf{length(in bps)} * [1 \times 10^{9}] * 650)$$

where:

6.022×10^{23} molecules/mole = Avogadro's constant

650 Daltons = the assumed average weight of a base pair

Table 3.1 Mock Community Design and Composition

Mock Community	Composition: % (Genome Copies)
Mock 1-1	*L. pneumophila* Philadelphia-1: 9 % (2.45×10^{4}) *L. pneumophila* France 5811: 81 % (2.21×10^{5}) Human genomic DNA: 10 %
Mock 1-2	*L. pneumophila* Philadelphia-1: 45 % (1.23×10^{5}) *L. pneumophila* France 5811: 45 % (1.31×10^{5}) Human genomic DNA: 10 %
Mock 1-3	*L. pneumophila* Philadelphia-1: 81 % (2.21×10^{5}) *L. pneumophila* France 5811: 9 % ($2.45 X 10^{4}$)

	Human genomic DNA: 10 %
Mock 2-1	*L. pneumophila* Philadelphia-1: 10 % (2.73 x 10^4) *L. pneumophila* France 5811: 40 % (1.05 x 10^5) *L. longbeachae*: 40 % (9.09 x 10^4) Human genomic DNA: 10 %
Mock 2-2	*L. pneumophila* Philadelphia-1: 30 % (8.18 x 10^4) *L. pneumophila* France 5811: 30 % (8.18 x 10^4) *L. longbeachae*: 30 % (6.82 x 10^4) Human genomic DNA: 10 %
Mock 2-3	*L. pneumophila* Philadelphia-1: 80 % (2.18 x 10^5) *L. pneumophila* France 5811: 5 % (1.36 x 10^4) *L. longbeachae* : 5 %(1.14 x 10^4) Human genomic DNA: 10 %
Mock 3-1	*L. pneumophila* Philadelphia-1: 10 % (2.73 x 10^4) *L. pneumophila* France 5811: 16 % (4.36 x 10^4) *L. longbeachae*: 16 % (3.64 x 10^4) *S. pneumoniae*: 16 % (7.24 x 10^4) *H. influenzae*: 16 % (7.46 x 10^4) *V. dispar*: 16 % (3.5 x 10^4) Human genomic DNA: 10 %
Mock 3-2	*L. pneumophila* Philadelphia-1: 50 % (1.36 x 10^5) *L. pneumophila* France 5811: 8 % (2.18 x 10^4) *L. longbeachae*: 8 % (1.82 x 10^4) *S. pneumoniae*: 8 % (3.62 x 10^4) *H. influenzae*: 8 % (3.73 x 10^4) *V. dispar*: 8 % (1.75 x 10^4) Human genomic DNA: 10 %
Mock 3-3	*L. pneumophila* Philadelphia-1: 80 % (2.18 x 10^5) *L. pneumophila* France 5811: 2 % (5.45 x 10^3) *L. longbeachae*: 2 % (4.54 x 10^3) *S. pneumoniae*: 2 % (9.06 x 10^3) *H. influenzae*: 2 % (9.33 x 10^3) *V. dispar*: 2 % (4.37 x 10^3) Human genomic DNA: 10 %

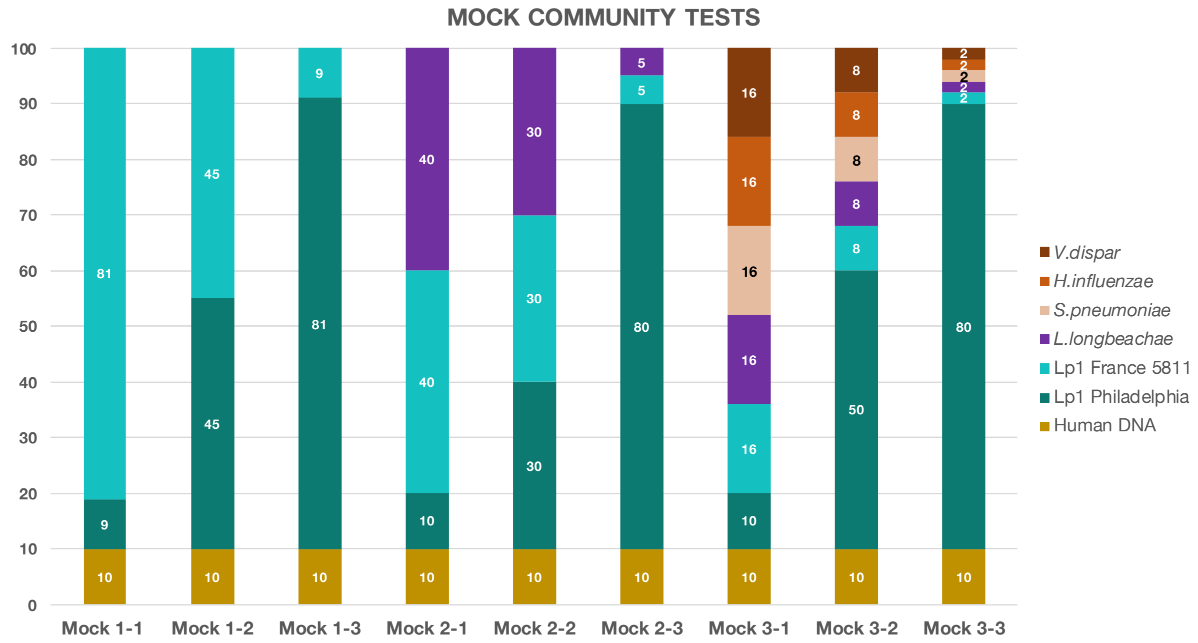


Figure 3.1. Compositions (in %) of the nine mock communities. Lp1 = *L. pneumophila* serogroup 1. Percentage is indicated on the y-axis and mock community name is indicated on the x-axis. Percentages of each component of the mock communities are displayed within the stacked bars.

3.3.1.2 Sensitivity Tests Design

Tests to determine the sensitivity of metagenomic sequencing were prepared by spiking human DNA with *L. pneumophila* Philadephia-1 DNA. Thirteen tests were prepared ranging from high relative abundance of *L. pneumophila* (99.99% of the total) to low relative abundance of *L. pneumophila* (0.001 % of the total) (**Table 3.2** and **Figure 3.2**):

Table 3.2 Sensitivity Test Design and Composition

Mock Community	Composition: % (Genome Copies)
S1	*L. pneumophila*: 99.99 % (2.65 x 10^5) Human genomic DNA: 0.01 %
S2	*L. pneumophila*: 99.9 % (2.64 x 10^5) Human genomic DNA: 0.1 %
S3	*L. pneumophila*: 99 % (2.62 x 10^5) Human genomic DNA: 1 %
S4	*L. pneumophila*: 95 % (2.51 x 10^5) Human genomic DNA: 5 %
S5	*L. pneumophila*: 90 % (2.38 x 10^5)

	Human genomic DNA: 10 %
S6	*L. pneumophila*: 75 % (1.99 x 10^5) Human genomic DNA: 25 %
S7	*L. pneumophila*: 50 % (1.32 x 10^5) Human genomic DNA: 50 %
S8	*L. pneumophila*: 25 % (6.64 x 10^4) Human genomic DNA: 75 %
S9	*L. pneumophila*: 10 % (2.65 x 10^4) Human genomic DNA: 90 %
S10	*L. pneumophila*: 5 % (1.32 x 10^4) Human genomic DNA: 95 %
S11	*L. pneumophila*: 1 % (2.65 x 10^3) Human genomic DNA: 99 %
S12	*L. pneumophila*: 0.01 % (2.65 x 10^2) Human genomic DNA: 99.9 %
S13	*L. pneumophila*: 0.001 % (2.65 x 10^1) Human genomic DNA: 99.99 %

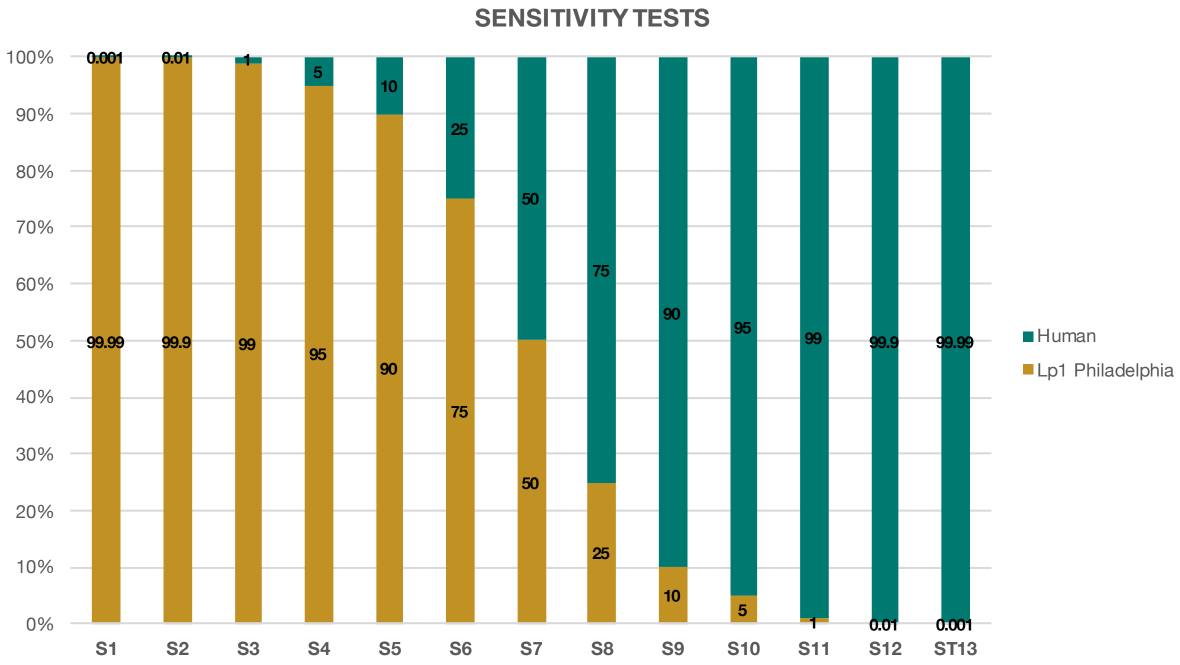


Figure 3.2. Composition of the Thirteen Sensitivity Tests. Lp1 = *L. pneumophila* serogroup 1. Percentage is indicated on the y-axis and mock community name is indicated on the x-axis. Percentages of each component of the mock communities are displayed within the stacked bars.

3.3.2 Metagenomic Sequencing

Libraries were prepared for each of the 9 mock communities (Mock 1-1 to Mock 3-3) and 13 sensitivity tests (S1 to S13) with the Nextera® XT DNA Library Preparation Kit (Illumina®) and sequencing was carried out on a MiSeq using version 3 chemistry to produced paired-end reads of 300 base pairs as described in Chapter 2, Sections 2.10.1 to 2.10.7 and using the KAPA SYBR ® FAST qPCR Library Quantification Kit as described in Chapter 2, Section 2.10.5.1.

3.3.3 Mock Community and Sensitivity Test Data Analysis

3.3.3.1 Data Cleaning and Quality Control

Sequenced reads were demultiplexed into individual libraries as described in Chapter 2, Section 2.12.1. All data cleaning and quality control steps (adapter trimming for removal of the Illumina Universal adapter sequence, quality filtering, PhiX removal and human DNA removal) were carried out as described from Sections 2.12.2 to 2.12.5, Chapter 2. Read numbers before and after quality control and after human DNA removal were investigated.

3.3.3.2 Taxonomic Classification

Taxonomic classification of mock communities was carried out using `Centrifuge` (Version 1.0.3) (Kim *et al.,* 2016) and the complete bacterial genomes database from RefSeq (O' Leary *et al.,* 2016) as described in Chapter 2, Section 2.13.1. Only one distinct classification was assigned to each read.

3.3.3.3 Alignment-Based Analysis of Sensitivity Tests

Sequenced reads from Sensitivity Tests (S1 to S13) were aligned to the *L. pneumophila* Philadelphia-1 reference genome (https://www.ncbi.nlm.nih.gov/nuccore/AE017354) to investigate the percentage of reference bases covered and mean depth of coverage across the range of high to low abundances. Mapping was carried out using the BBMap script from `BBTools` (Bushnell *et al.,* 2014). Duplicates were removed from the mapped reads using the `dedupe` script from `BBTools`. Deduplicated reads were mapped back to the reference sequence to generate mapping and coverage statistics. See Appendix Section 9.1.10 for full code.

3.3.3.4 *In silico L. pneumophila* Sequence Type Analysis of Sensitivity Tests

In silico sequence-based typing (SBT) analysis based on the traditional *L. pneumophila* SBT scheme (Gaia *et al.,* 2005, Ratzow *et al.,* 2007, Mentasti *et al.,* 2014) was carried out on sensitivity tests (S1 to S13) to determine the limit of detection of reliable sequence type information. The *L. pneumophila* SBT database was retrieved from the HPA website ((http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php) and `SRST2` (Version 0.2.7) and was used to carry out the analysis (Inouye *et al.,* 2014). All steps undertaken are detailed and explained in Chapter 2, Section 2.13.2.

3.3.3.5 Identification of *L. pneumophila* Strains

All completed and draft genome assemblies submitted as *L. pneumophila* in the RefSeq database were retrieved as described in Chapter 2, Section 2.13.3. The distance between the RefSeq genomes was estimated against a reference sequence with `Mash` (Version 2.1) (Ondov *et al.,* 2016). The `Mash` programme applies the MinHash dimensionality-reduction technique for simple distance estimation between genomes. The reference genome chosen was the type strain of the species, *L. pneumophila* Philadelphia-1, as defined on the NCBI website (https://www.ncbi.nlm.nih.gov/nuccore/AE017354). First a sketch was made of the reference genome and then all query genomes. The mash distance between the reference and each query genome was calculated. See Appendix Section 9.1.11 for the full code.

Strain-level analysis using the `StrainEst` program (Version 1.2.2) (Albanese *et al.,* 2017) was carried out to validate the identification and abundance of mixed *L. pneumophila* strains in the mock communities and the sensitivity tests S8 to S10 as single strain controls. All steps undertaken are detailed in Chapter 2, Section 2.13.3. A variant was only considered if it had a coverage depth of 10 or greater. It was determined that a maximum identity threshold of 99 % ensured a sufficient prediction specificity.

3.3.4 Metagenomic Sequencing of Clinical Specimens

3.3.4.1 Sample Preparation, *L. pneumophila* Identification and Sequencing

Sputum specimens were treated and extracted at Public Health England (PHE) Colindale using a MagnaPure Compact machine. Three samples (C1, C3 and C5) were confirmed positive for the presence of *L. pneumophila* serogroup 1 by urinary antigen testing and the ESGLI qPCR protocol based on the detection of the *L. pneumophila mip* and *wzm* genes. Three samples (C2, C4 and C6) tested negative for the presence of *L. pneumophila* by urinary antigen testing, qPCR and culture. Sputum treatment, extraction and all routine diagnostic testing was carried out by the Respiratory and Vaccine-Preventable Bacteria Reference Unit (RVPBRU) team at PHE. Library preparation (in triplicate) and sequencing was carried out by me at the Genomic Medicine Section, NHLI, Imperial College London as described Chapter 2, Sections 2.10.1 to 2.10.7 and using the KAPA SYBR® FAST qPCR Library Quantification Kit as described in Chapter 2, Section 2.10.5.1.

3.3.4.2 Clinical Specimen Data Analysis

Sequenced reads were demultiplexed back into individual libraries as described in Chapter 2, Section 2.12.1. All data cleaning and quality control steps (adapter trimming for removal of the Illumina Universal adapter sequence, quality filtering, PhiX removal and human DNA removal) were carried out as described from Section 2.12.2 to 2.12.5, Chapter 2. Read numbers before and after quality control and after human DNA removal were reported. Insert sizes of paired-end reads were calculated. Taxonomic classification was carried out on each sample using the Centrifuge classifier (Kim *et al.,* 2016) and the RefSeq bacterial genomes database (O' Leary *et al.,* 2016) as described in Chapter 2, Section 2.13.1.

3.4 Results

3.4.1 Sequencing Run Output for Mock Communities and Sensitivity Tests

Mock communities and sensitivity tests were sequenced simultaneously. Consequently, basic statistics relating to the run output are reported in **Table 3.3**, including total number of paired reads before and after quality trimming/filtering and total number of paired reads after removal of reads mapping to the human genome. The total number of paired reads sequenced for each library ranged from 1,582,468 reads (Mock2-3) to 3,562,414 reads (Mock2-1). The total number of paired reads retained after quality trimming and filtering for each library ranged from 90 % (S11) to 95 % (Mock3-3). The mean insert size ranged from 180 bps (S8) to 230 bps (Mock3-1) for each sequenced library.

Table 3.3 Sequenced Reads Output for Mock Communities and Sensitivity Tests.

Library	Total Number of Reads (Paired)	Total Number of Reads After QC	Total Number of Reads After Human DNA Filtering
Mock 1-1	2,500,310	2,360,652 (94 %)	2,106,768 (89 %)
Mock 1-2	2,192,836	2,010,098 ((91 %)	1,771,762 (88 %)
Mock 1-3	2,063,246	1,915,948 (93 %)	1,658,116 (86 %)
Mock 2-1	3,562,414	3,376,310 (94 %)	2,959,798 (87 %)
Mock 2-2	2,283,984	2,150,318 (92 %)	1,880,838 (87 %)
Mock 2-3	1,582,468	1,451,646 (92 %)	1,251,126 (86 %)
Mock 3-1	1,606,344	1,500,696 (93 %)	1,329,778 (87 %)
Mock 3-2	2,196,594	2,070,604 (94 %)	1,807,960 (87 %)
Mock 3-3	2,118,978	1,988,056 (94 %)	1,705,788 (86 %)
S1	2,540,134	2,408,568 (95 %)	2,407,486 (99.99 %)

S2	2,528,658	2,374,756 (94 %)	2,373,898 (99.9 %)
S3	1,622,968	1,485,626 (92 %)	1,469,496 (99 %)
S4	2,878,544	2,689,192 (93 %)	2,538,318 (94 %)
S5	2,609,240	2,463,058 (94 %)	2,219,790 (90 %)
S6	2,225,444	2,095,342 (94 %)	1,582,920 (76 %)
S7	1,868,118	1,692,096 (91 %)	858,836 (51 %)
S8	2,792,728	2,568,364 (92 %)	674,622 (26 %)
S9	2,923,334	2,729,114 (93 %)	292,912 (11 %)
S10	2,441,872	2,262,792 (93 %)	125,134 (6 %)
S11	2,234,536	1,997,040 (90 %)	23,968 (1 %)
S12	1,981,918	1,805,650 (91 %)	2,300 (0.1 %)
S13	2 ,412,830	2,263,244 (94 %)	3,118 (0.1 %)

3.4.2 Accuracy of Sequencing Mock Communities

Paired-end library reads were mapped to the human genome and non-human reads were extracted and taxonomically classified to the species level. The actual input (relative abundance) of human DNA and microorganisms in each mock community was then compared to the sequenced output (**Figure 3.3**). There was good agreement between the proportion of input DNA and the output (the proportion of reads sequenced and classified as input species).

3.4.3 Sensitivity of Detection of *L. pneumophila* in a Human DNA Background

3.4.3.1 Relative Abundance of *L. pneumophila* Versus Coverage of the Reference Genome

The 13 sensitivity tests containing defined proportions of *L. pneumophila* (99.9 % for S1 to 0.001 % for S13) were mapped to the Philadelphia-1 reference genome. The percentage of reference bases covered, and depth of coverage were investigated (**Figure 3.4**).

A total of 100 % of reference bases were covered from S1 to S8 (99.99 % to 25 % *L. pneumophila* DNA). This dropped to 99 % and 98 % for S9 and S10 (10 % and 5 % *L. pneumophila* DNA, respectively) and 7 % to 3 % for S11 to S13. Depth of coverage varied from 134 x to 0.04 x.

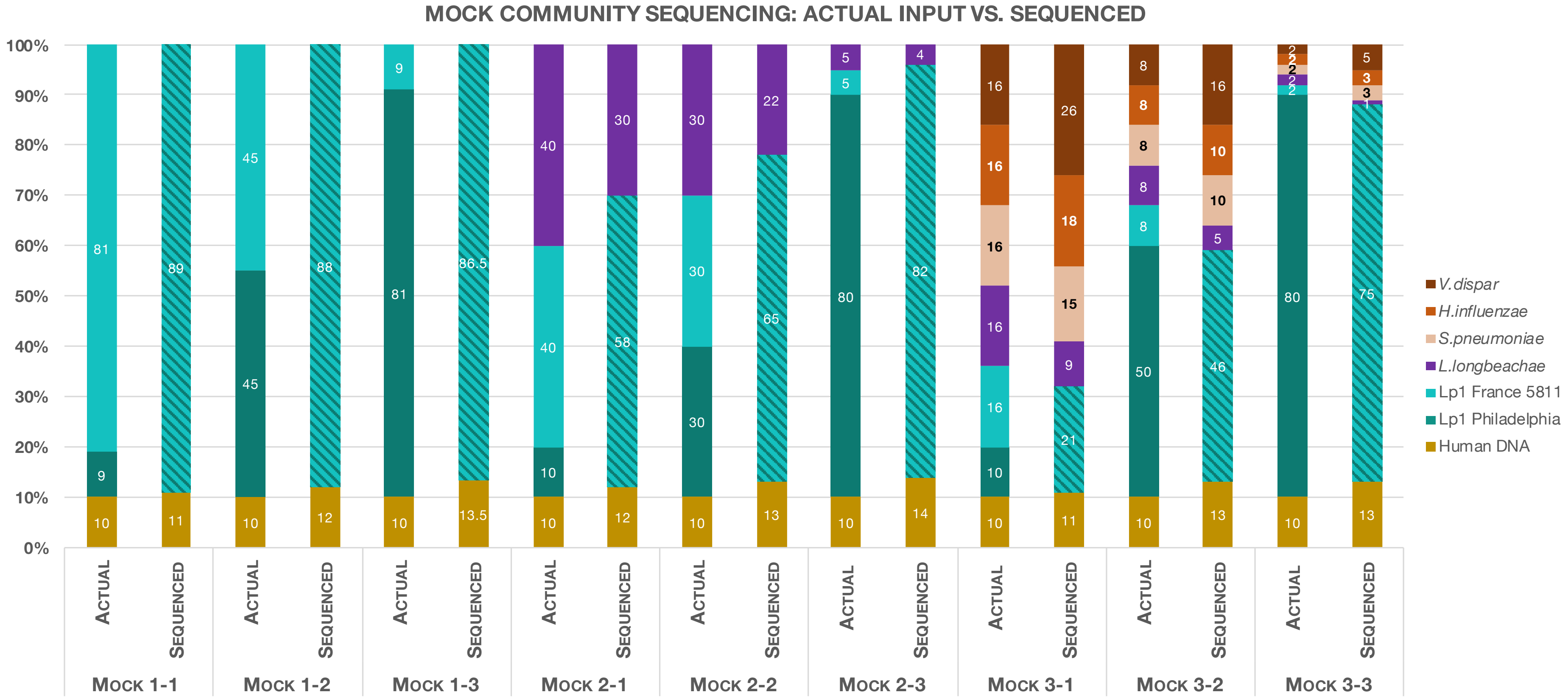


Figure 3.3 Sequencing and taxonomic classification validation against mock communities. Stacked bar charts showing the actual input (relative abundance) versus percentage of sequenced reads assigned at the species taxonomic level by the Centrifuge classifier for microorganisms and mapped to the human genome for human read classification. Striped *L. pneumophila* bars represent the relative abundance of the mixture of two strains. Percentage is indicated on the y-axis and mock community name is indicated on the x-axis. Percentages of each component of the mock communities are displayed within the stacked bars.

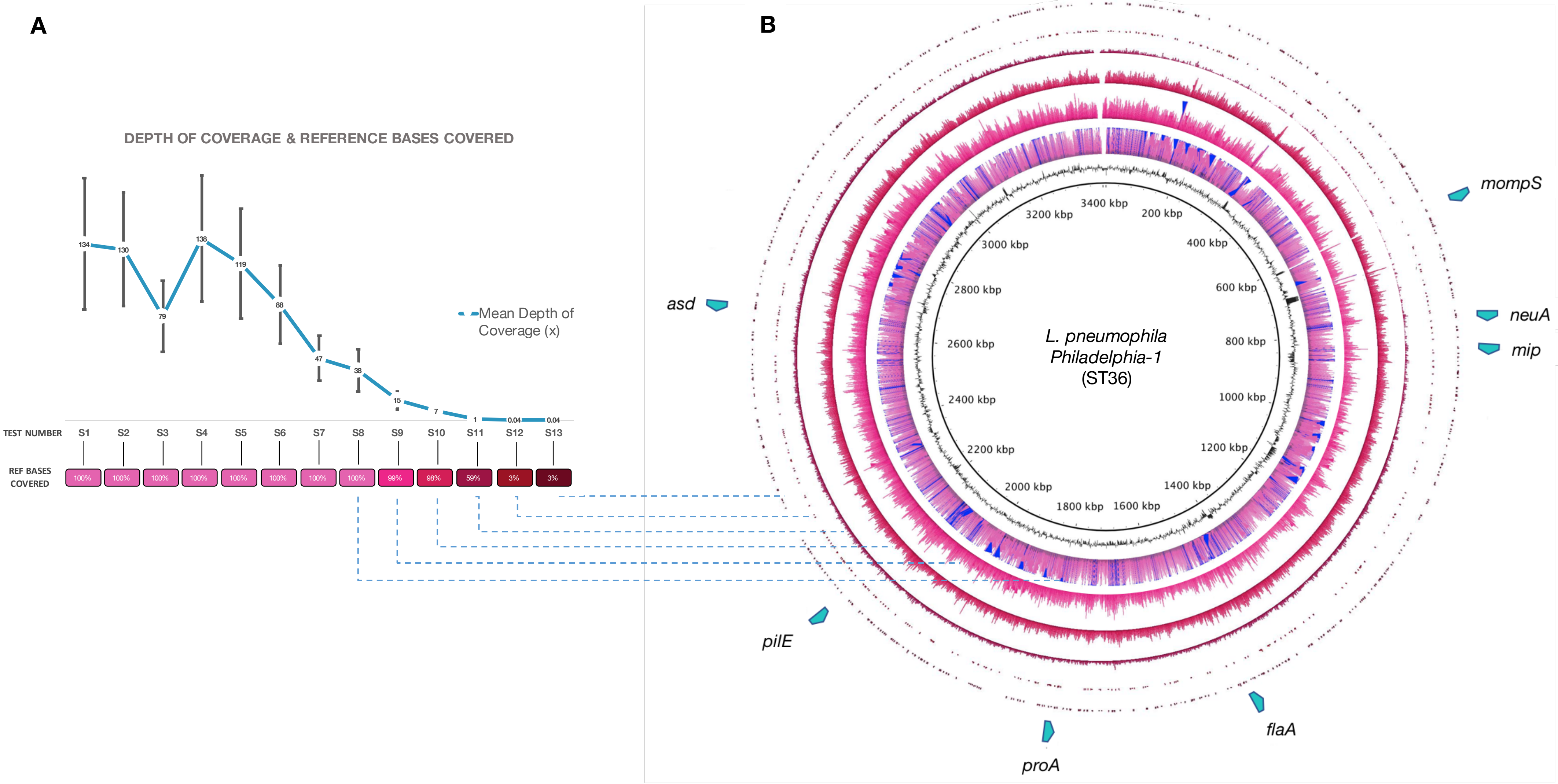


Figure 3.4 Sensitivity of detection of *L. pneumophila* in a background of human DNA. The relative abundance of *L. pneumophila* in 13 sequenced sensitivity tests containing defined proportions (99.9 % for S1 to 0.001 % for S13) versus genome coverage when sequenced tests were mapped to the Philadelphia-1 reference genome was investigated. **(A)** shows the depth of coverage of sensitivity tests containing *L. pneumophila* Philadelphia-1 from high abundance (S1: 99.99 %) to low abundance (S13: 0.001 %) and percentage of reference bases covered. **(B)** Circle plot representing sensitivity tests S8 (100 % of reference bases covered) to S13 (3 % of reference bases covered). Blue areas represent spikes in coverage depth. Green annotated arrows represent regions of the genome where the seven loci from the sequence-based typing scheme are located.

3.4.3.2 Limit of Detection of *L. pneumophila* Sequence Type Information

Quality filtered reads from each sample were aligned to the *L. pneumophila* European Study Group for *Legionella* Infections (ESGLI) allele database (http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php) for sequence type profiling. The sensitivity tests contained *L. pneumophila* Philadelphia-1 which is a sequence type 36 (ST36) strain represented by the allelic profile: *flaA*-3, *pilE*-4, *asd*-1, *mip*-1, *mompS*-14, *proA*-9 and *neuA*-1. The SRST2 program and the ESGLI allele database were used here to investigate if the correct allele numbers could be determined, determined with uncertainty or could not be determined for the sensitivity tests (**Figure 3.5**). Uncertainty signifies that the best scoring allele had a low-depth of coverage across part of or the whole allele or there were missing bases. All allele numbers representing ST36 could be determined with certainty from tests S1 to S7, representing relative abundances of 99.9 % to 50 %. The average depth of coverage of alleles ranged from 110 x to 33 x for these 7 tests. For test S8, 6 alleles were determined with certainty and 1 allele (*neuA*) was determined with uncertainty. For test S9, 6 alleles were determined with certainty and 1 allele (*flaA*) was determined with uncertainty. For test S10, 4 alleles (*pilE*, *asd*, *mompS*, and *proA*) were determined with uncertainty and 3 alleles (*flaA*, *mip* and *neuA*) were undetermined. For test S11, 2 alleles (*asd* and *mompS*) were determined with uncertainty and 5 alleles (*flaA*, *pilE, mip, proA* and *neuA*) were undetermined. For tests S12 and S13, all alleles were undetermined. Overall, when *L. pneumophila* was present in these samples at a relative proportion of less than 50 % and above, all alleles and therefore the sequence type profile could be determined with certainty. When present down to a 10 % proportion of the total sample content (including human DNA) the sequence type could be determined however, one allele was determined with uncertainty. *L. pneumophila* sequence type in test samples with a proportion below 10 % could not be determined with certainty or remained undetermined.

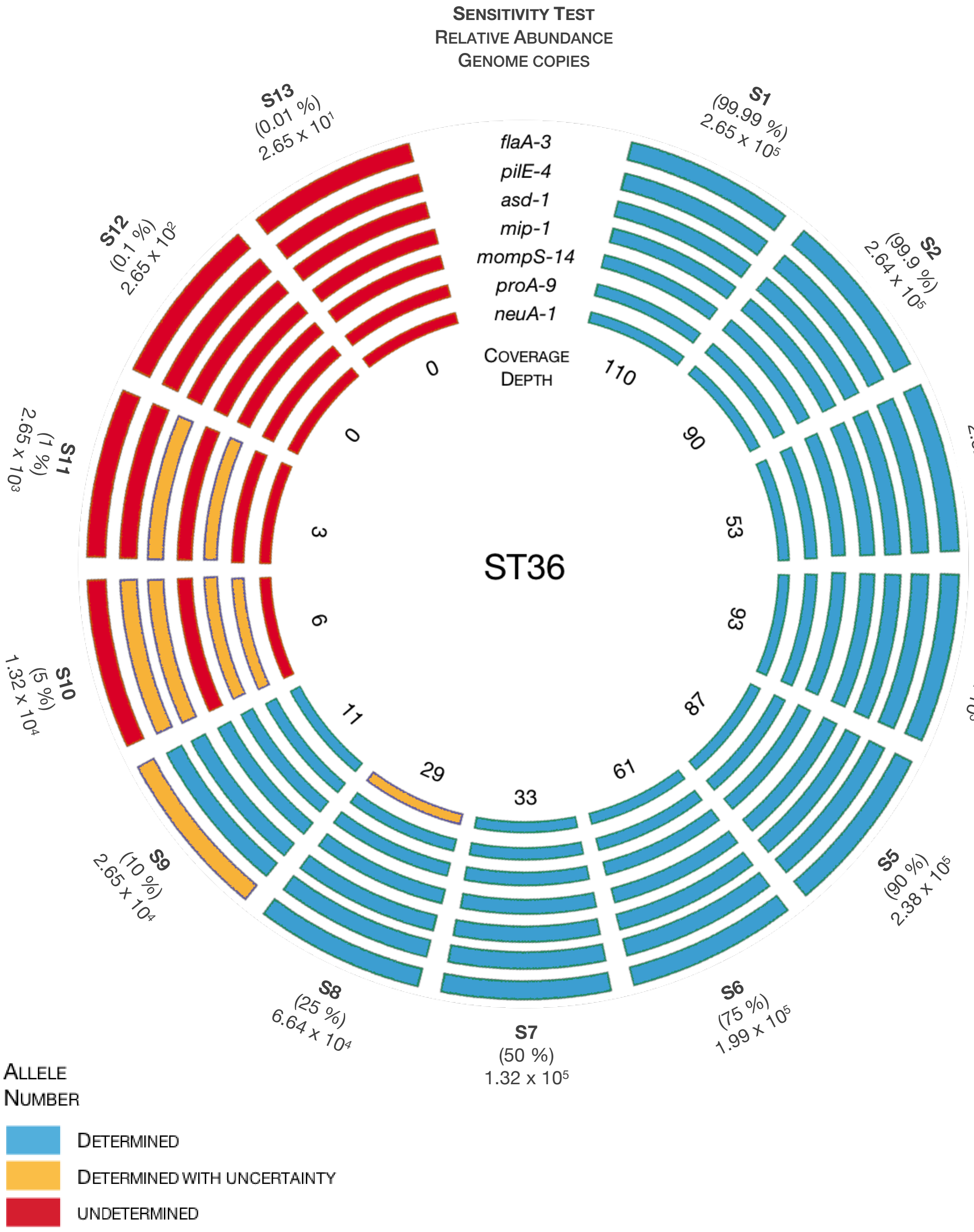


Figure 3.5 Circle plot representing each sensitivity test from high abundance to low abundance presence of *L. pneumophila* (ST36). The SRST2 program and the ESGLI allele database were used to investigate if correct allele numbers could be determined (blue), determined with uncertainty (yellow) or could not be determined (red) for the sensitivity tests. Uncertainty signifies that the best scoring allele had a low depth of coverage across part of or the whole allele or there were missing bases. Each segment of the circle plots shows the test name, percentage of the sample composed of *L. pneumophila* (ST36) and the corresponding number of *L. pneumophila* genome copies.

3.4.4 Strain-Level Identification of Mixed *L. pneumophila* Strains

3.4.4.1 *L. pneumophila* Strain Database

A *L. pneumophila* strain database was built from complete and draft genome assemblies deposited in the NCBI RefSeq server on the 5th of October 2018 (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/). At that time, the server contained a total of 593 genomes. *In silico* sequence type analysis carried out on the assemblies revealed that 492 genomes had an allele profile previously reported in the ESGLI database representing 90 different sequence types. A sequence type was not assigned to 101 of the genomes: 32 genomes had allele profiles not reported in the ESGLI database, 3 genomes had 6 known alleles and 1 novel allele sequence not reported in the database, 11 genomes had 6 known alleles and one allele with only a partial match to a known allele in the database, 4 genomes had 6 known alleles and 1 allele not present in the assembly. Finally, 51 genomes had multiple copies of the *mompS* locus. The presence of multiple non-identical *mompS* copies, usually two, in *L. pneumophila* genomes has been defined previously (Moran-Gilad *et al.,* 2015) and is known to create difficulties when attempting to call a traditional sequence type from genomic data. Whilst a bioinformatic solution (Gordon *et al.,* 2017) exists to address this issue, the tool requires both raw reads as well as the assembly. Therefore, it was decided to define the sequence type of these assemblies as "undetermined" for the purpose of this study. The RefSeq assembly reference number, sequence type and allele profile for each genome are detailed in Appendix Section 9.3.

The Mash distance was calculated between the reference genome, the type strain of the species, *L. pneumophila* Philadelphia-1, and all other genomes. The genome furthest from the reference had a mash distance of 0.0759, equivalent to an average nucleotide identity (ANI) of 92.41 %. The threshold for species inclusivity has been defined as a 95 to 96 % ANI (Richter *et al.,* 2009). However, the genomes falling below 95 % ANI (mash distance of 0.05) in the current study were found to represent three known subspecies of *L. pneumophila* other than subspecies *pneumophila*. These subspecies were *fraseri*, *pascullei* and the recently described *raphaeli* (Kozak-Muiznieks *et al.,* 2018). Also included within the genomes was an ST2186 strain representing either an atypical subspecies *fraseri* strain or a novel subspecies, as previously reported (Kozak-Muiznieks *et al.,* 2018). Since all these subspecies are represented in the traditional *L. pneumophila* typing scheme, all genomes were retained.

3.4.4.2 Identification of *L. pneumophila* strains in Mock Communities and Sensitivity Tests

StrainEst analysis was carried out on the 9 mock communities and 3 of the sensitivity tests (ST8 to ST10) as single strain controls. The mock communities each contained varying relative abundances of two *L. pneumophila* strains: *L. pneumophila* Philadelphia-1 (ST36) and *L. pneumophila* France 5811 (ST1) (**Table 3.1** and **Figure 3.1**) and the model reported a list of top hits or abundances up to 100 % based on the *L. pneumophila* component only, disregarding other microorganisms. The single strain controls (S8 to S10 – **Table 3.2** and **Figure 3.2**) contained *L. pneumophila Philadelphia-1* (ST36) only. The model predicted all mock communities to contain the two strains that corresponded to the actual strains in terms of sequence type designation (**Figure 3.6**). False positive strains were predicted, each with an abundance of less than 3 % in the mock communities. It was noted that these false positive predictions fell within the same phylogenetic clusters as one or both of the true positive predicted strains. The model accurately predicted the ST36 strain in the single strain controls for S8, S9 and S10. Similar to the mock predictions, false positive strains falling within the same phylogenetic cluster as ST36 were predicted but this time at a relative abundance of less than 0.001 % each. Predicted strains and abundances for each mock community and sensitivity test are detailed in Appendix Section 9.4.

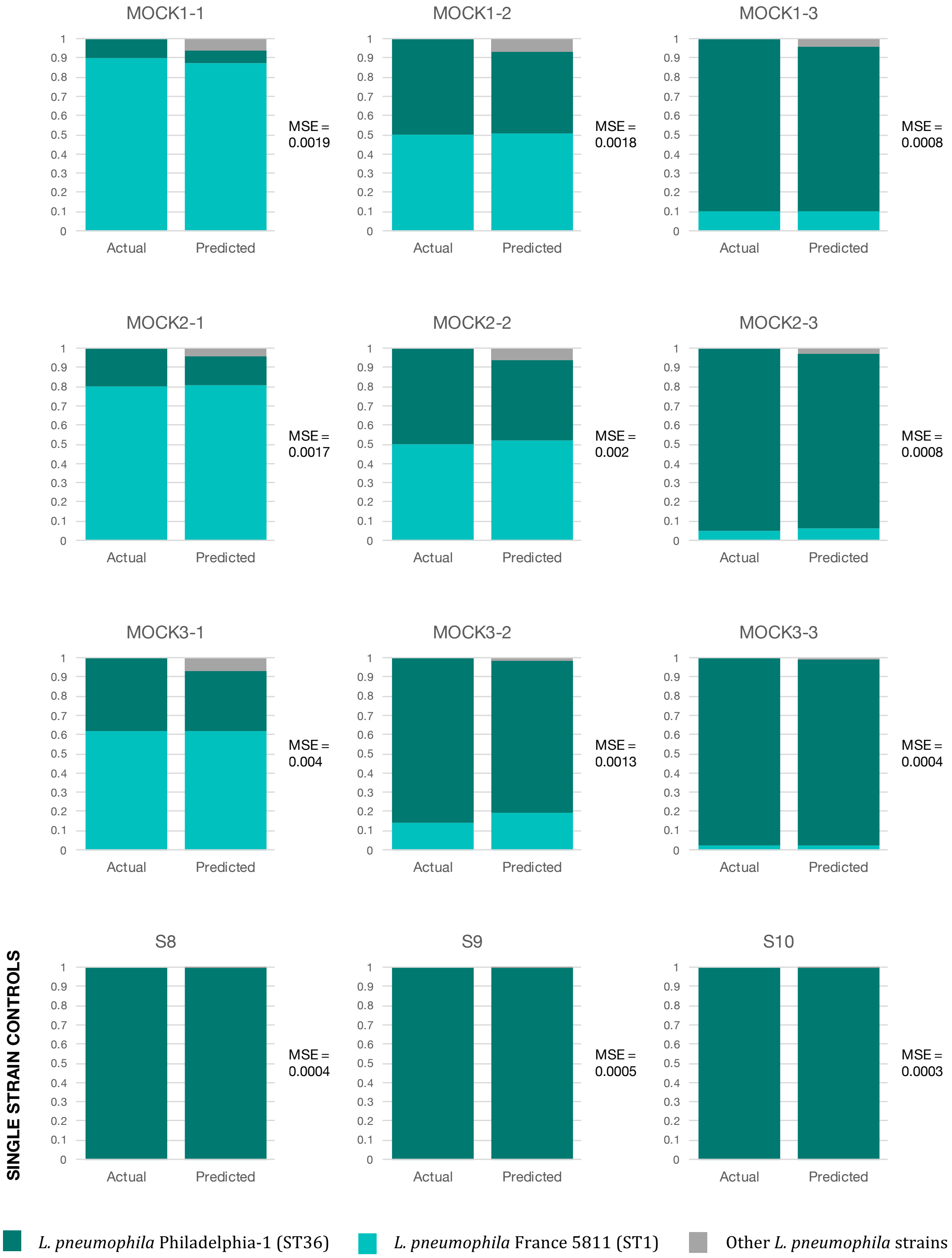


Figure 3.6 StrainEst analysis for prediction of multiple *L. pneumophila* strains in 9 mock communities and 3 sensitivity tests as single strain controls. The model reported a list of top hits or abundances based on the *L. pneumophila* component only. A mean squared error (MSE) was calculated by StrainEst to represent the average squared distance between the estimated values and the actual values. Percentage is displayed on the y-axis and actual composition versus predicted composition is displayed on the x-axis.

3.4.5 Host DNA Contamination in Clinical Specimens

3.4.5.1 Sequencing Run Output for Clinical Specimens

The total number of paired reads sequenced for the clinical specimen libraries ranged from 1,397,702 (C4-C) to 3,868,338 (C4-B). One library failed during sequencing (C3-B) and no reads were generated. The number of paired reads sequenced for the negative control libraries ranged from 4,144 (NEG4) to 9,868 (NEG2). The total number of paired reads retained after quality filtering and trimming for the clinical specimen libraries ranged from 80 % (C6-A) to 96 % (C1-A). The number of reads retained after quality trimming and filtering the negative control libraries ranged from 15 % (NEG2) to 20 % (NEG3) (**Table 3.4**). The mean insert size for the clinical specimen libraries ranged from 200 bps (C1-B) to 600 bps (C1-A). Mean insert size for negative control libraries ranged from 130 bps (NEG5) to 150 bps (NEG6). Reads retained after human DNA filtering are investigated in Section 3.4.5.2.

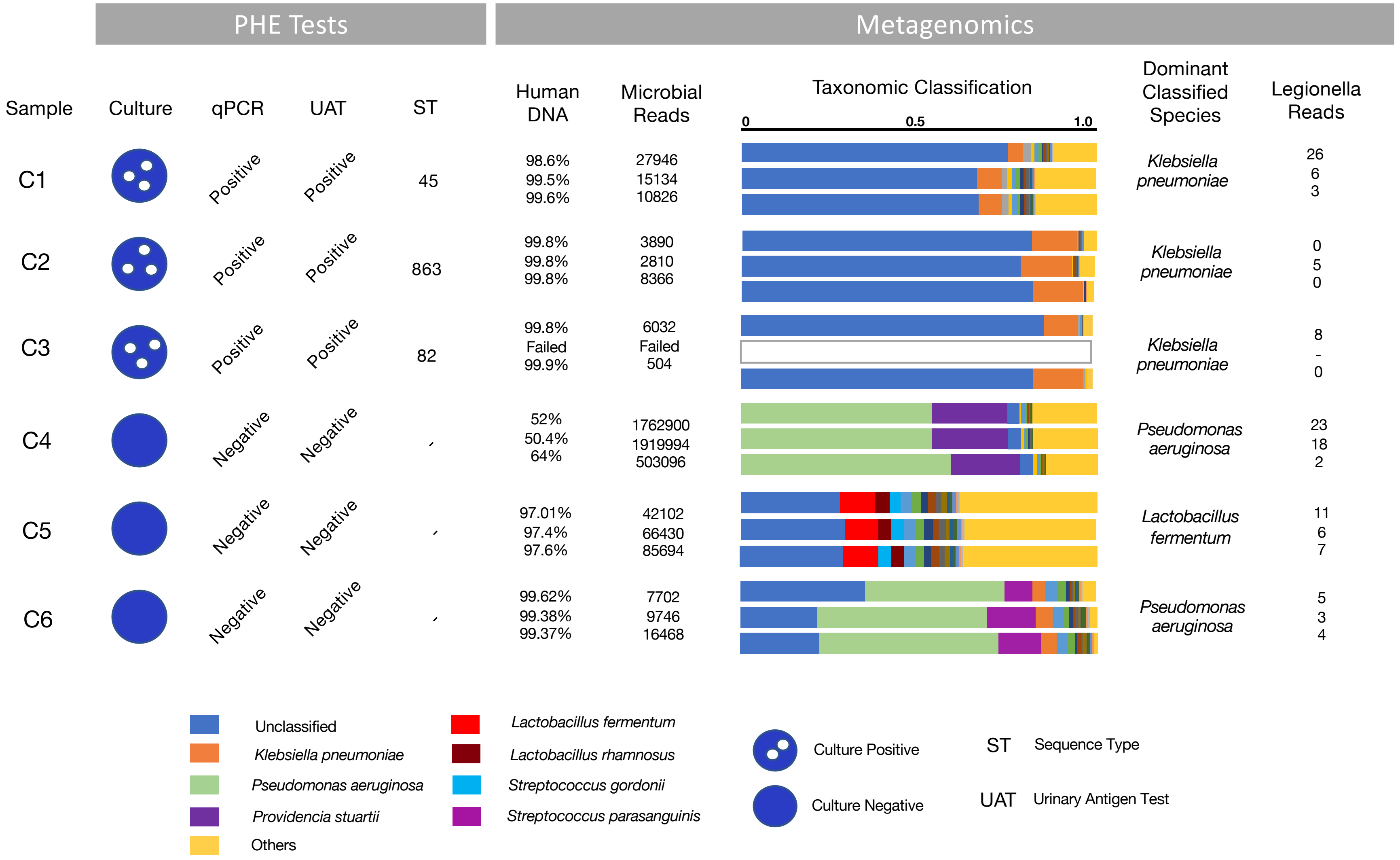
Table 3.4. Total number of paired-end reads sequenced for each library. Each library was prepared in triplicate (A-C) and 6 negative controls (NEG1 to NEG6), reads after adapter trimming, quality trimming and filtering, and reads remaining after human DNA filtering.

LIBRARY	TOTAL NUMBER OF READS (PAIRED)	TOTAL NUMBER OF READS AFTER QC	TOTAL NUMBER OF READS AFTER HUMAN DNA FILTERING
C1-A	2,060,330	1,997,180 (97 %)	2,746 (1.4 %)
C1-B	3,238,800	2,856,800 (88 %)	15,134 (0.5 %)
C1-C	2,688,450	2,341,454 (87 %)	10,826 (0.5 %)
C2-A	2,086,370	1,867,662 (89 %)	3,890 (0.2 %)
C2-B	1,504,436	1,363,266 (90 %)	2,810 (0.2 %)
C2-C	4,819,006	4,261,260 (88 %)	8,366 (0.2 %)
C3-A	3,396,246	3,024,572 (89 %)	6,032 (0.2 %)

C3-B	Failed	Failed	Failed
C3-C	2,420,952	2,145,284 (89 %)	504 (0.02 %)
C4-A	3,673,250	3,098,758 (84 %)	1,762,900 (57 %)
C4-B	3,868,338	3,342,646 (80 %)	1,919,994 (57 %)
C4-C	1,397,702	1,117,802 (85 %)	503,096 (45 %)
C5-A	1,418,602	1,276,512 (86 %)	42,102 (3.2 %)
C5-B	2,525,718	2,150,884 (85 %)	66,430 (3 %)
C5-C	3,558,120	3,072,280 (86 %)	85,694 (3 %)
C6-A	2,021,874	1,801,956 (89 %)	7,702 (0.4 %)
C6-B	1,560,252	1,349,298 (86 %)	9,746 (0.7 %)
C6-C	2,628,342	2,296,856 (87 %)	16,468 (0.7 %)
NEG1	8,448	1,546 (18 %)	192 (12.4 %)
NEG2	9,868	1,504 (15 %)	274 (18.2 %)
NEG3	6,198	1,246 (20 %)	438 (35 %)
NEG4	4,144	744 (18 %)	66 (9 %)
NEG5	4,852	902 (18 %)	66 (7.3 %)
NEG6	4,726	896 (18 %)	60 (6.7 %)

3.4.5.2 The Effect of Host DNA Contamination on the Ability to Sequence Microorganisms from Clinical Specimens

Three previously determined *L. pneumophila* positive samples and three *L. pneumophila* negative samples were sequenced in triplicate at a shallow depth (24 x). The aim of sequencing these samples was to understand the proportion of host DNA in the sputum specimens available for this study. Samples C1, C2 and C3 were confirmed *L. pneumophila* positive at PHE by culture, which is the gold standard, qPCR and urinary antigen testing. A sequence type was also generated for these specimens. C4, C5 and C6 were confirmed negative for *L. pneumophila* by the same techniques. Metagenomic sequencing of the specimens revealed that human DNA constituted the majority of reads sequenced for samples C1, C2, C3, C5 and C6 (**Figure 3.7**). Human DNA constituted approximately half of the reads sequenced for sample C4. Whilst the classification of microbial reads was investigated, the majority of reads for *Legionella* positive samples remained unclassified. This is likely due to the presence of viral and fungal DNA reads, which were not included in the reference database here. From the classified reads, *Legionella* was never the dominant classified species sequenced in these specimens. The dominant classified species was *Klebsiella pneumoniae* for the three positive *L. pneumophila* samples. The dominant classified species for *L. pneumophila* negative samples were *Pseudomonas aeruginosa* (C4 and C6) and *Lactobacillus fermentum* (C5). *Legionella* reads belonging to the *L. pneumophila* species as well as other *Legionella* species were identified in at least one of each of the clinical specimen libraries however it was determined that these could not reliably indicate presence of the bacteria as the sequenced negative controls NEG3 and NEG5 also contained a number of *Legionella* reads, 2 and 6 reads respectively.



Sample	Culture	qPCR	UAT	ST	Human DNA	Microbial Reads	Dominant Classified Species	Legionella Reads
C1	⦿	Positive	Positive	45	98.6% 99.5% 99.6%	27946 15134 10826	*Klebsiella pneumoniae*	26 6 3
C2	⦿	Positive	Positive	863	99.8% 99.8% 99.8%	3890 2810 8366	*Klebsiella pneumoniae*	0 5 0
C3	⦿	Positive	Positive	82	99.8% Failed 99.9%	6032 Failed 504	*Klebsiella pneumoniae*	8 - 0
C4	●	Negative	Negative	-	52% 50.4% 64%	1762900 1919994 503096	*Pseudomonas aeruginosa*	23 18 2
C5	●	Negative	Negative	-	97.01% 97.4% 97.6%	42102 66430 85694	*Lactobacillus fermentum*	11 6 7
C6	●	Negative	Negative	-	99.62% 99.38% 99.37%	7702 9746 16468	*Pseudomonas aeruginosa*	5 3 4

Figure 3.7 Metagenomic sequencing of 6 clinical specimens to investigate the degree of host DNA contamination on microbe sequencing. Three previously determined *L. pneumophila* positive samples and three *L. pneumophila* negative samples were sequenced in triplicate. Human DNA constituted the majority of reads sequenced for samples C1, C2, C3, C5 and C6 and half of reads sequenced for sample C4. From the classified reads, *Legionella* was never the dominant classified species sequenced in these specimens.

3.5 Discussion

The initial aim of this chapter was to assess the ability of the pipeline to accurately sequence mock communities at their species-level input proportions as a means of validation before approaching a panel of clinical and environmental specimens. Biases and errors can occur at any level of sample or data manipulation including during library preparation, sequencing, control of data quality and analysis (Bowers *et al.,* 2015, Jones *et al.,* 2015, Schlaberg *et al.,* 2017[a]). This may lead to the observation of an inaccurate community distribution in comparison to the actual species distribution within the mock community.

All microorganisms included in the mock communities were identified to the species-level. It is acknowledged that taxonomic classifiers cannot accurately identify microbes to the strain level of a species due to the high similarity between some strains as well as database inaccuracies such as misnomers and strain misplacement (Federhen, 2015, Kim *et al.,* 2016). Since the mock communities were composed of two *L. pneumophila* strains, both *L. pneumophila* strains were considered at the species taxonomic level only. The assigned proportions of the various bacteria and human reads were generally very close to the real proportions present in the mock communities and no major deviations from actual species input was observed. Whilst these communities do not represent real microbiome specimens, they are useful for technical validation here specifically for *Legionella* sequencing as it is of primary interest to the study.

The next aim involved an analysis of sensitivity of detection of *L. pneumophila* in a background of human DNA. A total of 100 % of reference bases were covered from sensitivity tests S1 to S8 which contained a relative abundance of 25 % *L. pneumophila* DNA at an average depth of 38 x. Whilst 99 % and 98 % of reference bases were covered for sensitivity tests S9 and S10 with an average depth of 15 x and 7 x, respectively, coverage fell thereafter for sensitivity tests S11 to S13. It was also noted that as the relative abundance of the microbial component fell, the technical background or contaminating microorganisms represented an increasing fraction of the microbial reads, as previously discussed (Salter *et al.,* 2014). This emphasises the difficulties generated by sample contamination from reagents or the laboratory environment which may then contribute to a greater fraction of the microbial component of the low microbial biomass specimen. To further confound metagenomic sequencing efforts for *Legionella* detection, contamination with *Legionella* species has been reported in commercial extraction kits (van der Zee *et al.,* 2002, Evans *et al.,* 2003) and commercial purified water (Shen *et al.,* 2006).

Identification of *L. pneumophila* at the strain-level is paramount for the epidemiological investigation of LD cases, particularly to link patients and detect clusters, whether community or travel related and to determine an environmental source of transmission. The traditional 7-loci *L. pneumophila* sequence typing scheme developed by ESGLI represents the current "gold standard" for molecular typing of *L. pneumophila* strains (Gaia *et al.,* 2005, Ratzow *et al.,* 2007, Mentasti *et al.*, 2014). As of 19th July 2019, a total 2,791 unique sequence types have been submitted to the ESGLI database. In this study, full sequence type information based on the ESGLI scheme could be achieved from direct metagenomic sequencing when *L. pneumophila* was present at a relative abundance of 10 % (S9) in the sample, here representing 2.65 x 10^4 genome copies. A caveat here however is that these sensitivity tests did not contain other microorganisms. A 16S rRNA microbiome study of sputum specimens from legionellosis patients by Mizrahi *et al.*, 2017 reported that *Legionella* was never the dominant genus and was always accompanied by other respiratory pathogens. The relative abundance of *L. pneumophila* within the microbial communities ranged from 0.004 % to 2.88 % in the Mizrahi *et al*. study. This highlights the significant challenge of directly sequencing of a low abundance *Legionella* genome moreover since community studies target microorganisms only and do not need to account for the additional burden of host DNA. Currently a 50-loci scheme is in development by ESGLI for isolate data (Moran-Gilad *et al*., 2015, David *et al*., 2016[b]). When available, it will be informative to apply this scheme to metagenomes containing *L. pneumophila* to investigate if a sensitivity threshold can be established.

Current microbial techniques for epidemiological studies focus on the identification and/or isolation of one strain. Microbial communities however may be composed of a number of strains from the same species. Whilst rare, previous studies have reported the presence of mixed strains of *L. pneumophila* identified by traditional techniques (Coscollá *et al.,* 2014, Wewalka *et al.,* 2014). For the investigation of mixtures of strains in metagenomes, a number of bioinformatic tools have been developed (Hong *et al.,* 2014, Ahn *et al.,* 2015, Luo *et al.,* 2015, Nayfach *et al.,* 2016, Scholz *et al.,* Quince *et al.,* 2017, 2016). These tools generally report the presence of the highest abundance strain only (Scholz *et al.,* 2016) or require specific conditions to be satisfied such as a depth of coverage (Luo *et al.,* 2015) which might not be achievable in some studies. StrainEst, a recently developed tool for strain profiling (Albanese *et al.,* 2017) alongside a database of public *L. pneumophila* genome assemblies was validated in the current study against the mock communities and sensitivity tests with the aim of employing its usage on real specimens in further chapters. Whilst mixed strains were correctly predicted, a number of false positives belonging to the same clusters as true positives were predicted at low levels. In light of these findings, to mitigate incorrect strain predictions when analysing real specimens (in Chapter 5), a threshold will be established to disregard strains predicted with a relative abundance of less than 3 %.

The final aim of the study was to examine the burden of host DNA when sequencing directly from a real clinical specimen at an affordable, shallow sequencing depth. Sequenced *L. pneumophila* positive samples and two of the negative samples were comprised primarily of human DNA reads, with up to 99.9 % of reads (C3-C) attributable to the host genome. This supports the results of a previous studies attempting to sequence directly from lower respiratory tract specimens (Doughty *et al.,* 2014, Pendleton *et al.,* 2017). The large quantity of human DNA reads in sputum specimens may be the product of neutrophils present in the airway which may propagate chromatin structures known as neutrophil extracellular traps (NETs) during infection (Brinkmann *et al.,* 2004). Additionally, the release of contents from the cells during cell death may contribute to the abundance of host DNA (Wartha *et al.,* 2007). Notably, the human genome is approximately one thousand times larger than the average bacterial genome. Even if only a few human cells are present in the low microbial biomass sample, host DNA can rapidly create a disproportionate noise to signal ratio.

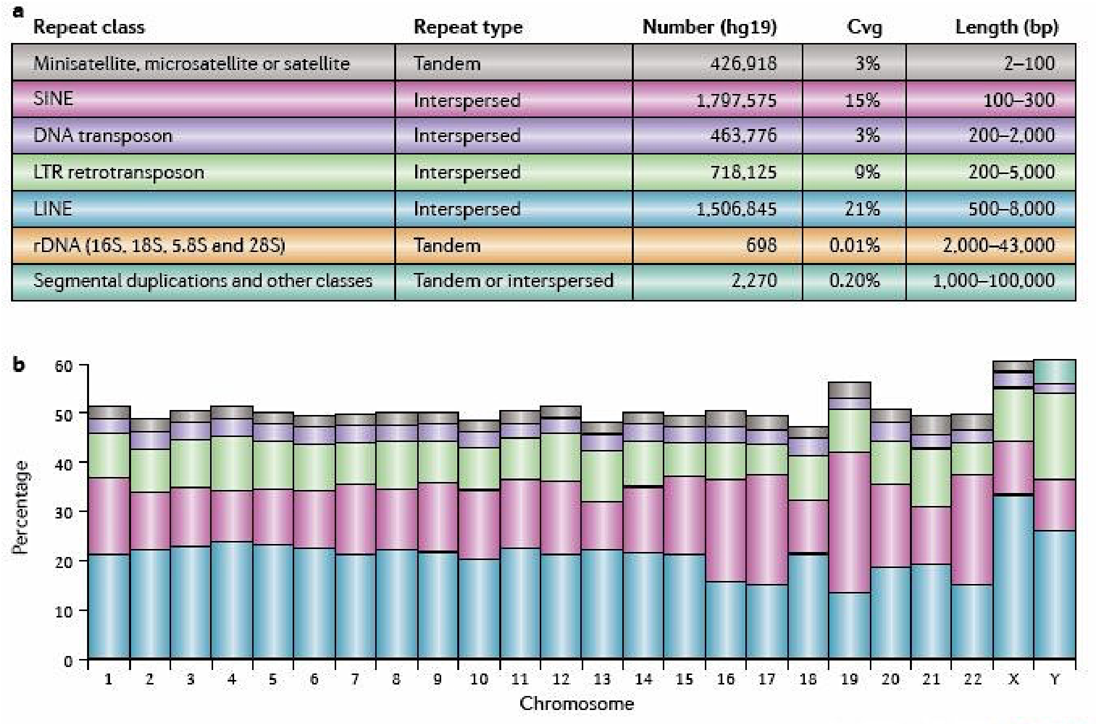
Few microbial reads were sequenced in the 5 clinical specimens in this present study compared to human DNA, leading to some difficulty in classifying the microorganisms present, particularly for the *Legionella* positive specimens. A number of *Legionella* reads were detected in *L. pneumophila* positive and negative samples however positivity could not be determined on this result alone as two of the negative controls contained a number of *Legionella* reads. This may be indicative of barcode cross-contamination within or between sequencing runs or contamination during processing in the laboratory although is more likely due to the noise to signal ratio because of the human (host) DNA (Salter *et al.,* 2014, Strong *et al.,* 2014). It also indicates that for low abundance pathogen detection a threshold needs to be established. Additionally, in the case of *L. pneumophila* positive samples, other species dominated the bacterial component, *K. pneumoniae* in the three samples studied here. A caveat of the current study is that sequencing was carried out at a shallow depth with samples multiplexed 24 times on an Illumina MiSeq platform. As depth of sequencing requirements for samples rise, so too does cost. Sequencing a low microbial abundance metagenome at an adequate depth to generate strain-level information can be prohibitively expensive. It is therefore necessary to apply a method to either capture *Legionella* genomic regions directly from a specimen of interest or to remove the contaminating host material that overwhelms the microbial component. These methods and their application will be addressed in the next chapters.

Chapter 4.

Method Development for Human DNA Depletion

4.1 Introduction

Human DNA sequences often comprise the majority of reads when a lower respiratory sample is sequenced. The human genome is approximately 3 billion base pairs in length per haploid genome, organised over 23 paired chromosomes. It is estimated that approximately 66 to 69 % of the genome is composed of repetitive DNA elements (de Koning *et al.,* 2011). Repetitive elements are represented by a variety of classes and have a differential distribution within the chromosomes (Treangen *et al.,* 2011) (**Figure 4.1**). Repetitive elements can be tandemly arranged, such as microsatellite DNA (Cooke *et al.,* 1979, Jefferys *et al.,* 1985) or are interspersed (Singer *et al.,* 1982), depending on their amplification mechanism. Interspersed repetitive elements arose from modes of transposition that resulted in identical or unidentical copies of the sequences being produced and inserted into the genome. In the case of LTR (long terminal repeat) and non-LTR elements, such as LINEs and SINEs, transposition occurred by a “copy-and-paste” mechanism which required an RNA intermediate obtained after transcription of the element (Rogers, 1985, Weiner *et al.,* 1986).



a

Repeat class	Repeat type	Number (hg19)	Cvg	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426,918	3%	2–100
SINE	Interspersed	1,797,575	15%	100–300
DNA transposon	Interspersed	463,776	3%	200–2,000
LTR retrotransposon	Interspersed	718,125	9%	200–5,000
LINE	Interspersed	1,506,845	21%	500–8,000
rDNA (16S, 18S, 5.8S and 28S)	Tandem	698	0.01%	2,000–43,000
Segmental duplications and other classes	Tandem or interspersed	2,270	0.20%	1,000–100,000

Figure 4.1 Classes of Repetitive DNA in the human genome. The table in panel **a** shows various named classes of repeat in the human genome, along with their pattern of occurrence (shown as 'repeat type' in the table; this is taken from the RepeatMasker annotation). The number of repeats for each class found in the human genome, along with the percentage of the genome that is covered by the repeat class (Cvg) and the approximate upper and lower bounds on the repeat length (bp). The graph in panel **b** shows the percentage of each chromosome, based on release hg19 of the genome, covered by repetitive DNA as reported by RepeatMasker. The colours of the graph in panel b correspond to the colours of the repeat class in the table in panel **a**. Image reprinted with permission from Springer Nature, Nature Reviews Genetics (License Number: 4630180719524). (Citation: Treangen *et al.*, 2011).

The most well studied interspersed repeat is the *Alu* element, a member of the SINE (short interspersed nuclear element) family (**Figure 4.2** [Deininger, 2011]). The *Alu* element is reported to be ancestrally derived from the 7SL RNA gene which is abundant in the cytoplasm and functions in protein secretion as a component of the signal recognition particle (Walter *et al.,* 1982). The *Alu* element is approximately 280 to 300 base pairs in length, formed by two diverged dimers which are separated by a short A-rich region. *Alu* elements are estimated to represent 11 % of the human genome i.e. the genome contains approximately one million copies (Batzer *et al.,* 2002).

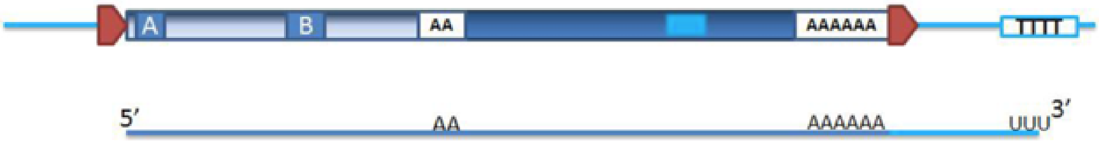


Figure 4.2 The structure of an *Alu* element. The top portion shows a genomic *Alu* element between two direct repeats formed at the site of insertion (red arrowheads). The *Alu* ends with a long A-run, often referred to as the A-tail, and it also has a smaller A-rich region (indicated by AA) separating the two halves of a diverged dimer structure. *Alu* elements have the internal components of a RNA polymerase III promoter (boxes A and B), but they do not encode a terminator for RNA polymerase III. They utilize whatever stretch of T nucleotides is found at various distances downstream of the *Alu* element to terminate transcription. A typical *Alu* transcript is shown below the genomic *Alu*, showing that it encompasses the entire *Alu*, including the A-tail, and has a 3' region that is unique for each locus. Image reprinted with permission from Springer Nature, Genome Biology (License Number: 4630780586117). (Citation: Deininger, 2011).

The presence of repetitive elements was first proposed in a study by Waring and Britten (Waring & Britten, 1966) on the re-association dynamics of mouse embryo DNA. They reported that mouse satellite DNA exhibited rapid annealing dynamics which more closely resembled the kinetics of bacteriophage and bacterial DNA renaturation. After further studies on the renaturation kinetics of DNA from different species, three distinct kinetic classes of DNA were described in complex organisms: the fast annealing, highly repetitive fraction, the intermediate, moderately repetitive fraction and the slow annealing, single copy fraction (Britten *et al.,* 1968) (**Figure 4.3**). Additionally, a mathematical explanation for the kinetics of re-association was given by the term “Cot” (Britten *et al.,* 1968). Cot refers to the rate at which heat-denatured DNA sequences in solution re-associate:

Cot = DNA concentration (mol/L) x renaturation time (seconds) x a buffer factor that accounts for the effect of cations on the speed of renaturation

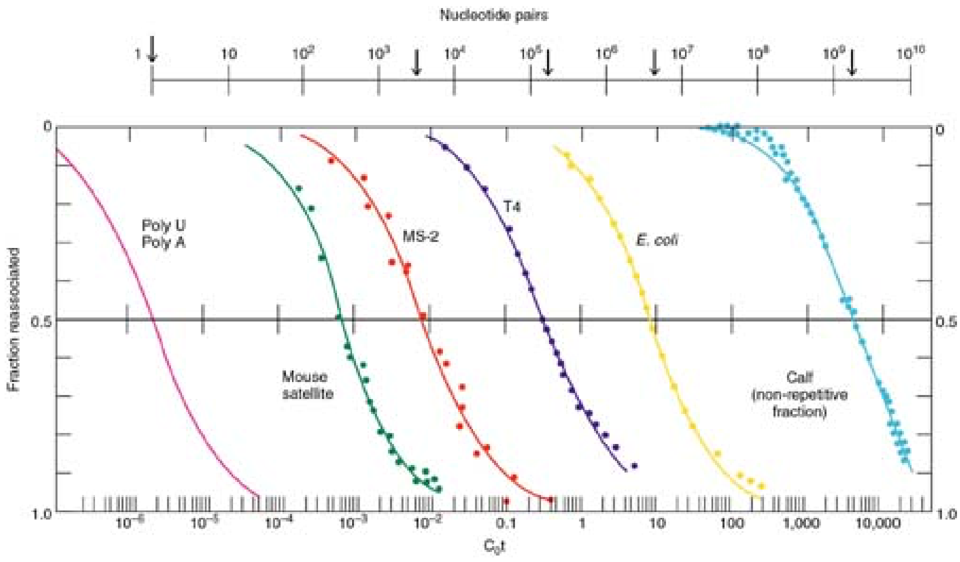


Figure 4.3 Re-association of nucleic acids from various sources. Image reprinted with permission from The American Association for the Advancement of Science, Science Journal (License Number: 4630210537706). (Citation: Britten *et al.,* 1968).

The rate at which a particular sequence re-associates is proportional to the number of times it is found in the genome (**Figure 4.4**). Therefore, repetitive sequences or those that occur more than once in a genome re-associate at a lower Cot value than unique or single copy sequences in a genome.

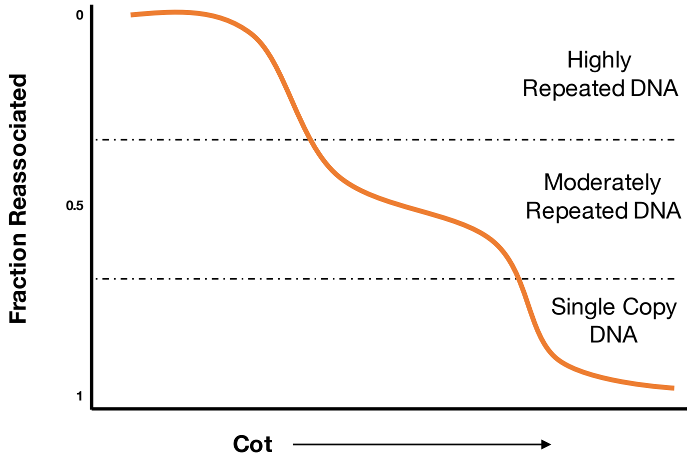


Figure 4.4 Cot curve representing fraction re-association versus Cot value for highly repeated, moderately repeated and single copy DNA.

Repetitive DNA is commonly used as a genetic tool for hybridisation applications. Cot-1 DNA, derived from human male placental DNA, is enriched for repetitive DNA sequences between 50 and 300 base pairs in length. Cot-1 is applied as a repeat suppressor to block non-specific hybridisation of human DNA probes that may have some complementarity to repetitive sequences. Additionally, inter-*Alu* PCR has been used for decades for the analysis of human loci flanked by *Alu* elements. This has allowed the mapping of previously unknown polymorphisms and mutations in human coding sequences (Nelson *et al.,* 1989).

The current study hypothesised that the repetitive elements in the human genome could be targeted to remove human DNA fragments from sputum DNA extracts in a timely manner, whilst preserving the microbial component for sequencing. Sputum specimens are considered to be of low microbial biomass overwhelmed by the presence of human DNA. A number of methods have been described for human DNA depletion from clinical specimens (reviewed in Chapter 1, Section 1.10.2) however these may not preserve the original structure of the microbial community and introduce biases. Furthermore, they may not be robust enough to remove a sufficient quantity of human cells or DNA sequences.

The aim of this chapter was to investigate a number of methods for the removal of human DNA fragments from mock and real specimens by targeting interspersed repetitive sequences present in the human genome.

4.2 Aims and Objectives

Investigate the removal of human DNA from mock and real samples by:

1. DNA: DNA hybridisation with biotinylated Cot-1 Human DNA probes.
    a. Biotinylate Cot-1 Human DNA by random primed labelling.
    b. Hybridisation test and control experiments on mock material.
    c. Quantify by qPCR the human DNA and microbial DNA in each test and control hybridisation.
2. DNA: DNA hybridisation with biotinylated *Alu* DNA probes.
    a. Prepare biotinylated *Alu* DNA probes by PCR.
    b. Hybridisation test and control experiments on digested human DNA.
    c. Identify depleted DNA qualitatively by gel electrophoresis.
3. RNA: DNA hybridisation with biotinylated *Alu* RNA capture probes.
    a. Prepare T7-*Alu* DNA templates by PCR.
    b. T7 RNA synthesis to generate biotinylated *Alu* RNA probes.
    c. Hybridisation tests and controls on human genomic DNA and bacterial genomic DNA.
    d. Identify depleted DNA qualitatively by gel electrophoresis.
    e. Sequence mock community and real specimens before and after depletion.
4. Investigate the depletion of human DNA from Illumina libraries.
    a. Prepare Illumina libraries of mock community and real specimens with paired hybridisation controls.
    b. Hybridisation experiments using the biotinylated *Alu* RNA, biotinylated Cot-1 DNA probes and a single cycle *Alu* PCR approach with biotin.
    c. Sequence the Illumina libraries before and after depletion.
5. Investigate the depletion of human DNA from an Oxford Nanopore library.
    a. Prepare Oxford Nanopore libraries of real specimens.
    b. Single cycle *Alu* PCR to incorporate biotin into the human DNA fragments of the test sample.
    c. Sequence the Oxford Nanopore libraries before and after depletion.

4.3 Methods & Results

4.3.1 DNA: DNA hybridisation with Biotinylated Cot-1 Human DNA Probes

4.3.1.1 Cot-1 DNA Size Selection

4.3.1.1.(i) Methods

Cot-1 DNA was sourced from Roche. For each size selection reaction, 1 µg of the Cot-1 DNA was taken and DNA fragments less than 100 base pairs were removed by use of the Pippin Prep (Sage Science, MA, USA) as described in Chapter 2, Section 2.7.4. Success of size selection was assessed by visualisation of fragment sizes on an Agilent Bioanalyzer performed as described in Chapter 2, Section 2.8.1. Cot-1 DNA was size selected as required.

4.3.1.1.(ii) Results

Pippin Prep size selection efficiently removed DNA fragments less than 100 base pairs from Cot-1 DNA. The DNA size range before (**A**) and after size selection (**B**) are shown in **Figure 4.5**.

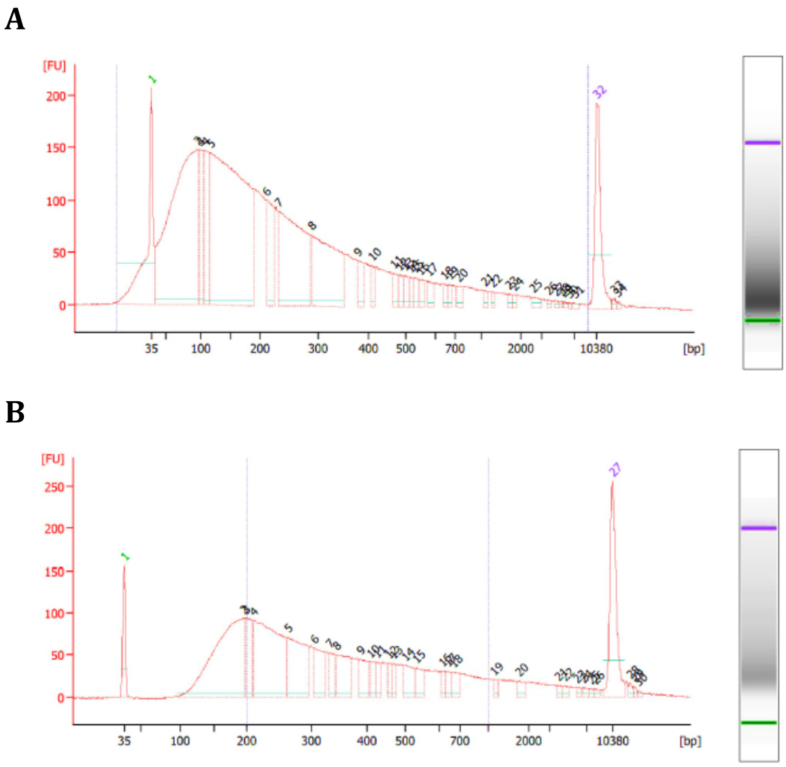


Figure 4.5 Bioanalyzer trace results. (**A**) Cot-1 DNA fragment lengths before size selection and (**B**) after size selection with Pippin Prep showing successful removal of fragments less than 100 base pairs in size.

4.3.1.2 Biotinylation of Cot-1 DNA by Random Primed Labelling

4.3.1.2.(i) Methods

The size selected Cot-1 DNA was biotinylated by random primed labelling using the Biotin-High Prime kit (Roche). Random primed labelling is a technique whereby the complementary DNA strand is synthesised by the Klenow polymerase using the 3'OH termini of the random oligonucleotides as primers. Biotin-16-dUTP is incorporated into the newly synthesised complementary DNA strand. In brief, 1 µg of the size selected Cot-1 DNA was added to distilled nucleic-acid free PCR grade water to a final volume of 16 µl. The DNA was denatured at 95 °C for 10 minutes in a thermal cycler and immediately chilled on ice. A total of 4 µl of the Biotin-High Prime master mix was added to the denatured DNA, mixed gently and centrifuged briefly. The reaction was incubated for 12 hours at 37 °C in a thermal cycler. The reaction was stopped by heating to 65 °C for 10 minutes. The reaction was purified to remove unincorporated biotin moieties using AMPure XP purification as described in Chapter 2, Section 2.7.1. A 5 µl aliquot was added to washed Dynabeads™ M-280 Streptavidin (Invitrogen). Bead washing was performed as described in Chapter 2, Section 2.9.3. After 30 minutes, the tube containing the mix was applied to a magnet for 10 minutes and the concentration of the supernatant was measured by PicoGreen assay as described in Chapter 2, Section 2.6.1.

4.3.1.2.(ii) Results

Biotinylation of the size selected Cot-1 DNA was carried out and the concentration after biotinylation and purification was 16.7 ng/µl. After incubation with 5 µl of streptavidin-coated magnetic beads, the concentration of the supernatant was 1.3 ng/µl, indicating that approximately 86 % of the Cot-1 DNA had been biotinylated and therefore captured on the beads and removed.

4.3.1.3 Determining Genomic DNA Fragment Sizes from Clinical Material

4.3.1.3.(i) Methods

The DNA fragment size profile of genomic DNA extracted (either with or without bead beating) from respiratory samples (throat swab, sputum – extracted samples from other projects within the Section of Genomic Medicine) were visualised by running aliquots on an Agilent Bioanalyzer as described in Chapter 2, Section 2.8.1.

4.3.1.3.(ii) Results

The fragment sizes for clinical material ranged from 1.5 kb to 7 kb for sputum samples and 4 kb for a throat swab that underwent bead beating followed by extraction. The fragment size for a sputum sample that underwent kit-based extraction without bead beating was 5 kb (**Figure 4.6**).

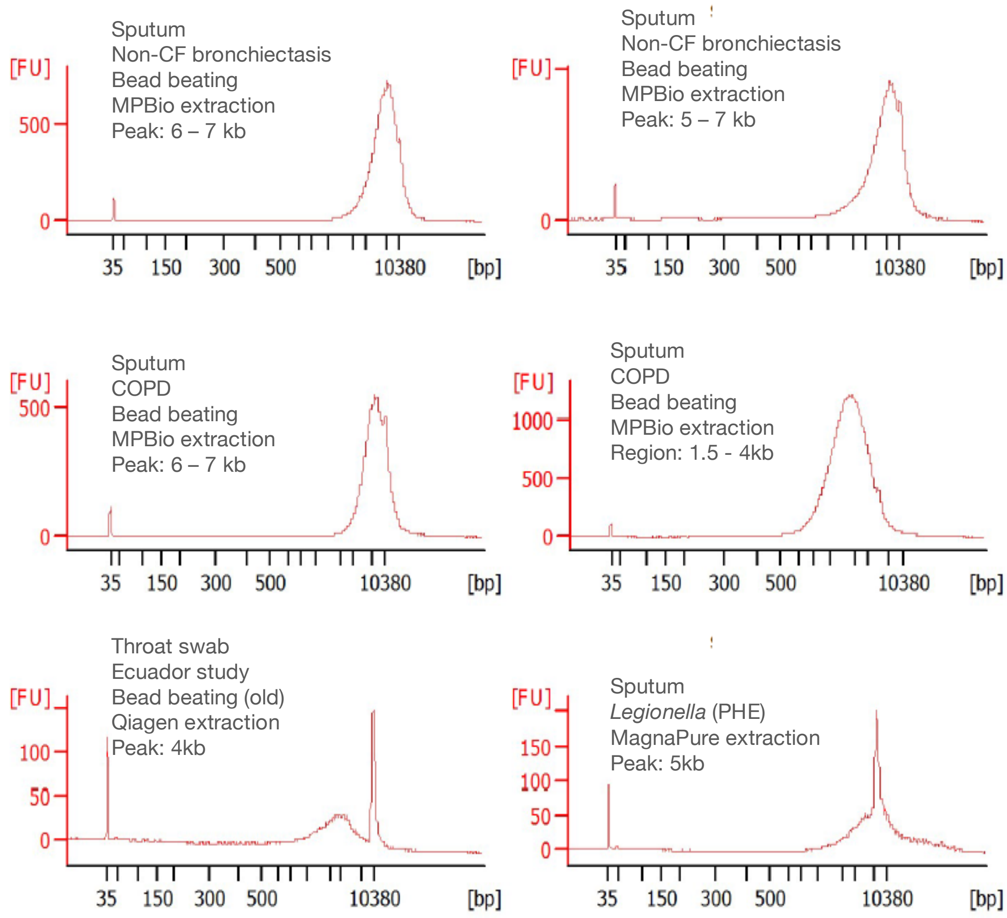


Figure 4.6 Profile of genomic DNA fragments extracted from a number of different patients or types of clinical material. On each trace size in kb of the peak is indicated. The fragment sizes for clinical material ranged from 1.5 kb to 7 kb for sputum samples and 4 kb for a throat swab that underwent bead beating followed by extraction. The fragment size for a sputum sample that underwent kit-based extraction without bead beating was 5 kb.

4.3.1.4 Preparation of a Mock Sample Representative of Genomic DNA

4.3.1.4.(i) Methods

A mock genomic DNA sample composed of 90 % human DNA and 10 % *Legionella pneumophila* Philadelphia-1 DNA (see Chapter 2, Section 2.4 for further details regarding material) was prepared as required. A high molecular weight community was stored for hybridisation experiments (approximately 30 - 50 kb in length). Size selection by QIAquick column purification was carried out as described in Chapter 2, Section 2.7.2.

The size selected material (approximately 10 kb in length) was stored for subsequent hybridisation experiments.

4.3.1.4.(ii) Results

The size selection step was carried out to ensure an approximate representation of fragment sizes based on those observed from the profiling of previously extracted clinical material (see Section 4.3.1.3.[ii]). Size selected *L. pneumophila* DNA represented average fragment lengths of 2 to 4 kb and size selected human DNA represented an average fragment length of 10 kb (**Figure 4.7 (A)** and **(B)**).

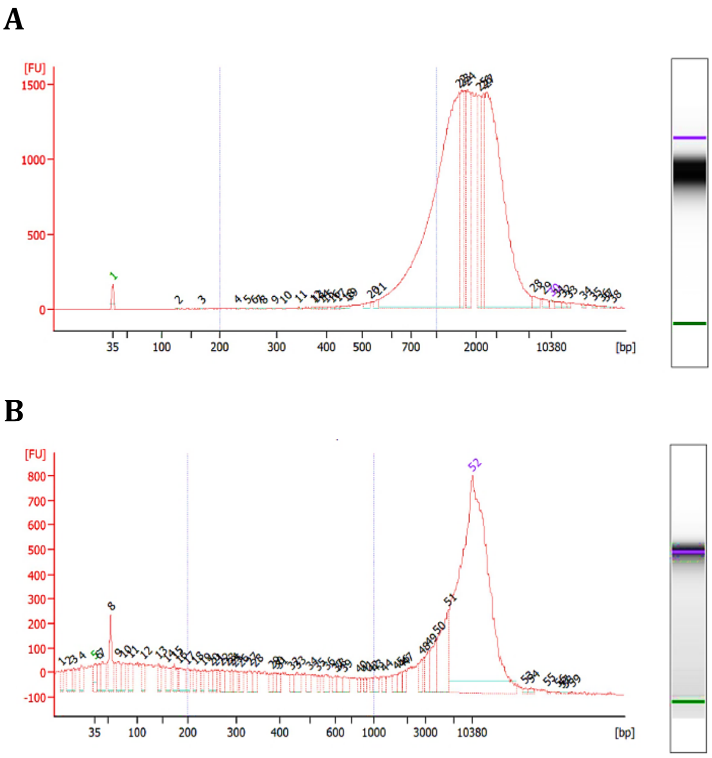


Figure 4.7 Bioanalyzer trace results. **(A)** size selected *L. pneumophila* DNA (2 to 4 kb) and **(B)** human DNA to represent fragment lengths of ~10 kb.

4.3.1.5 Hybridisation Experiments

4.3.1.5.(i) Methods

Taking the mock sample and the biotinylated Cot-1 DNA a series of hybridisation experiments were carried out as detailed in **Table 4.1**. For each hybridisation experiment a mock hybridisation control without a probe was also run. Hybridisation buffers and stocks were prepared or purchased as described in Chapter 2, Section 2.9.2. Each hybridisation reaction was added to washed streptavidin-coated beads as described in Chapter 2, Section 2.9.3. An equal volume of the sample containing the biotinylated nucleic acids was added to the washed beads and incubated at room temperature for 30 minutes with shaking (Vortex Genie 2, Scientific Industries). The tube containing the mix was applied to a magnet for 15 minutes and the supernatant ("microbial pool") was harvested and stored at -20 °C until further use. Beads were then washed in 200 µl of 1X Low TE buffer, the tube was applied to the magnet for 5 minutes and the supernatant ("wash pool") was harvested and stored as per the "microbial pool". The beads were re-suspended in 15 µl of double distilled H_2O and heated to 95 °C for 10 minutes. The tube was allowed to cool and then applied to a magnet. The supernatant ("human pool") was stored as per the other two pool samples. The pools were purified using the NEB Monarch PCR purification kit as described in Chapter 2, Section 2.7.3. To quantify bacterial DNA (16S rRNA gene) and human DNA (*GAPDH* gene) for each post-capture test and control hybridisation, SYBR Green qPCR was carried out as described in Chapter 2, Section 2.9.7. The experimental process can be visualised in **Figure 4.8**.

Table 4.1 Hybridisation experiments with biotinylated Cot-1 DNA on a size selected mock community.

	Experiment Conditions							
Experiment	**Cot-1 (ng)**	**Sample (ng)**	**Cot-1 fragment size**	**Sample fragment size**	**Buffer**	**Denaturation**	**Hybridisation**	**Variable changed**
1	50	50	> 100 bp	~30 -50 kb	2XSSPE, 0.2% SDS	95 °C - 5 min	65 °C - 60 min	N/A
2	50	50	> 100 bp	~30 - 50 kb	2XSSPE, 0.2% SDS	95 °C - 5 min	65 °C - 4 hrs	Incubation time increased to 4 hours
3	50	50	> 100 bp	~10 kb	2XSSPE, 0.2% SDS	95 °C - 5 min	65 °C - 60 min	Size selection for fragments less than or equal to 12 kb
4	100	100	> 100 bp	~10 kb	2XSSPE, 0.2% SDS	95 °C - 5 min	65 °C - 60 min	Increased concentration of probe and sample
5	100	100	> 100 bp	~10 kb	2X SSPE + 1% SDS added to reaction	95 °C - 5 min	65 °C - 60 min	Increased percentage of SDS
6	100	100	> 100 bp	~10 kb	10X SSPE + 1% SDS added to reaction	95 °C - 5 min	65 °C - 60 min	Changed SSPE buffer concentration to 10X
7	100	100	> 100 bp	~10 kb	10X SSC + 1% SDS added to reaction	95 °C - 5 min	65 °C - 60 min	Changed buffer to SSC
8	100	100	> 100 bp	~10 kb	2XSSPE, 0.2% SDS	95 °C - 5 min	65 °C - 60 min	Hybridisation buffer added after denaturing Cot-1 and sample

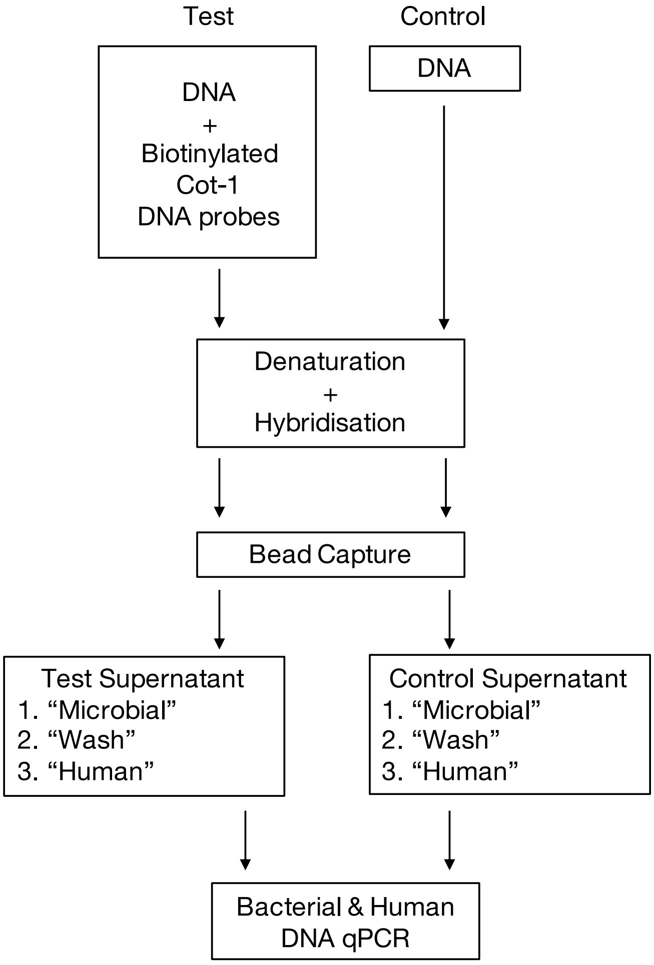


Figure 4.8 Experimental flow chart for Cot-1 DNA hybridisation tests and negative controls.

4.3.1.5.(ii) Results

After hybridisation and bead-capture, the initial supernatant ('Microbial'), the supernatant from the washing step ('Wash') and the supernatant after bead denaturation ('Human') from each hybridisation test and control was quantified. The wash step was not carried out on beads for Experiment 2, therefore no data is shown on the result graphs.

Quantitative PCR assays were carried out on seven of the eight experiments (**Figure 4.9** to **Figure 4.15**). Experiment 1 was not quantified due to insufficient material. Human *GAPDH* copy number was quantified for Experiments 2 to 8. Bacterial 16S rRNA gene copy number was quantified for post-capture Experiments 2, 5, 6, 7 and 8. Bacterial quantification was not carried out for Experiments 3 and 4 due to insufficient material. An increase in human DNA quantified from the denatured beads and a decrease in human

DNA quantified from the 'Microbial' supernatant would signify the success of the assay. If the control represented a similar dynamic or if increased human DNA was quantified from the 'Wash' supernatant, this would signify non-specific binding of DNA to beads. An increase in bacterial DNA quantified from the 'Wash' and 'Human' supernatant and a decrease in bacterial DNA quantified in the 'Microbial Test' versus 'Microbial Control' would also signify non-specific binding of DNA to beads.

Based on all experiments, there was limited evidence that human DNA was depleted from the mock samples. There was evidence of minor non-specific binding of bacterial DNA in Hybridisation Test 6 (**Figure 4.13**). It must also be acknowledged that loss of DNA may have been incurred during purification steps.

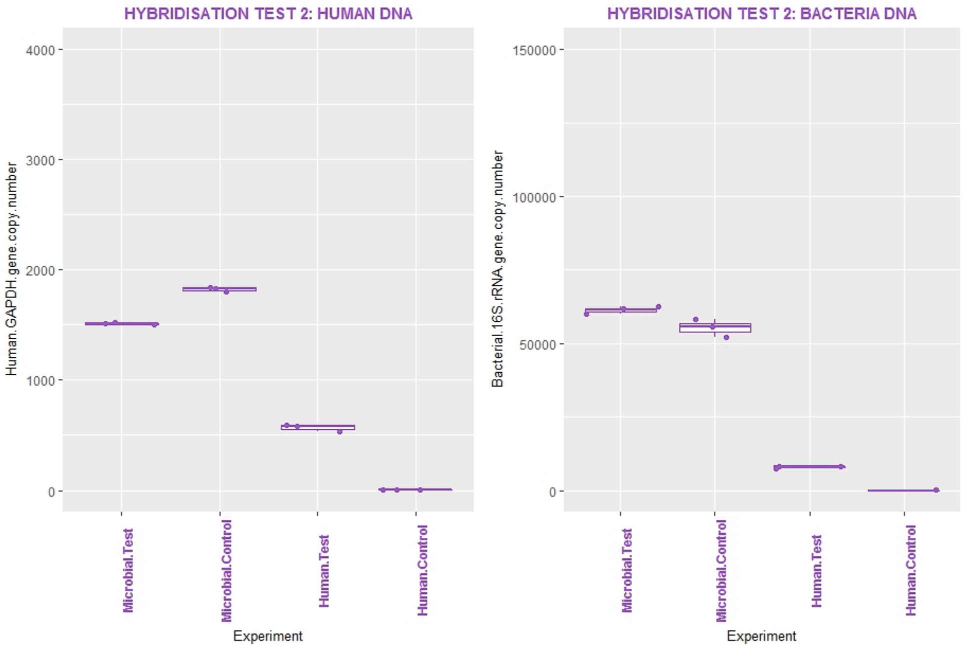


Figure 4.9 Boxplots of qPCR Results for Experiment 2. Human DNA presence determined by *GAPDH* copy number with Bacterial DNA presence determined by 16S rRNA copy number. Microbial, Wash and Human Test/Control = initial supernantant, supernatant from the washing step and supernatant after bead denaturation, respectively.

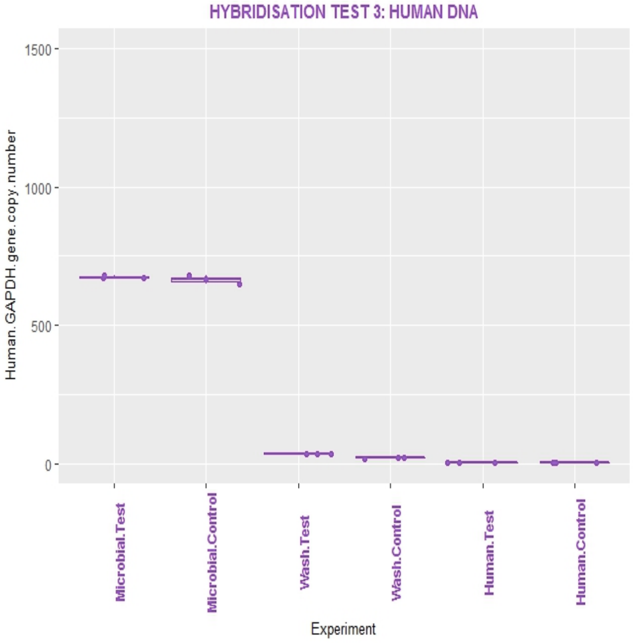


Figure 4.10 Boxplot of qPCR Results for Experiment 3. Human DNA presence determined by *GAPDH* copy number. Microbial, Wash and Human Test/Control = initial supernantant, supernatant from the washing step and supernatant after bead denaturation, respectively.

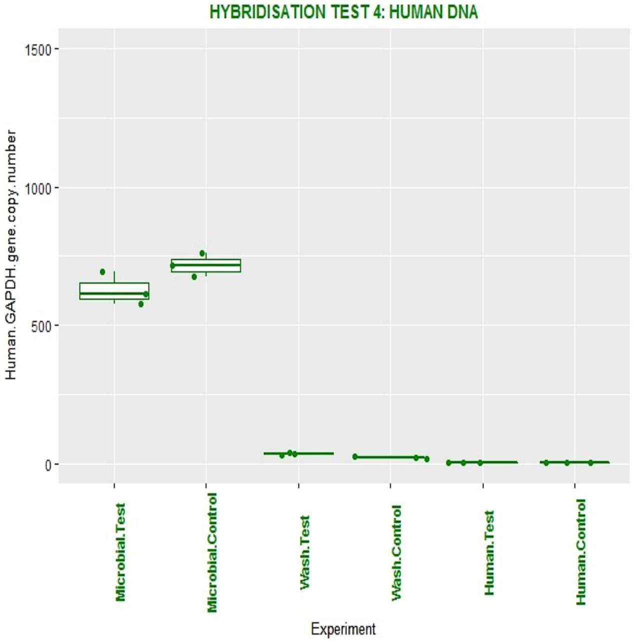


Figure 4.11 Boxplot of qPCR Results for Experiment 4. Human DNA presence determined by *GAPDH* copy number. Microbial, Wash and Human Test/Control = initial supernantant, supernatant from the washing step and supernatant after bead denaturation, respectively.

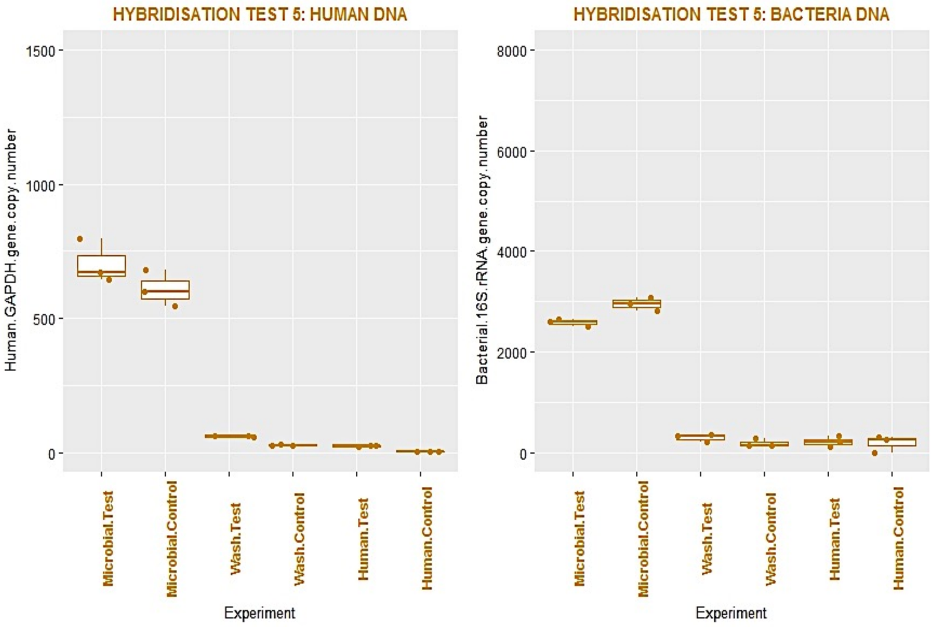


Figure 4.12 Boxplots of qPCR Results for Experiment 5. Human DNA presence determined by *GAPDH* copy number with Bacterial DNA presence determined by 16S rRNA copy number. Microbial, Wash and Human Test/Control = initial supernantant, supernatant from the washing step and supernatant after bead denaturation, respectively.

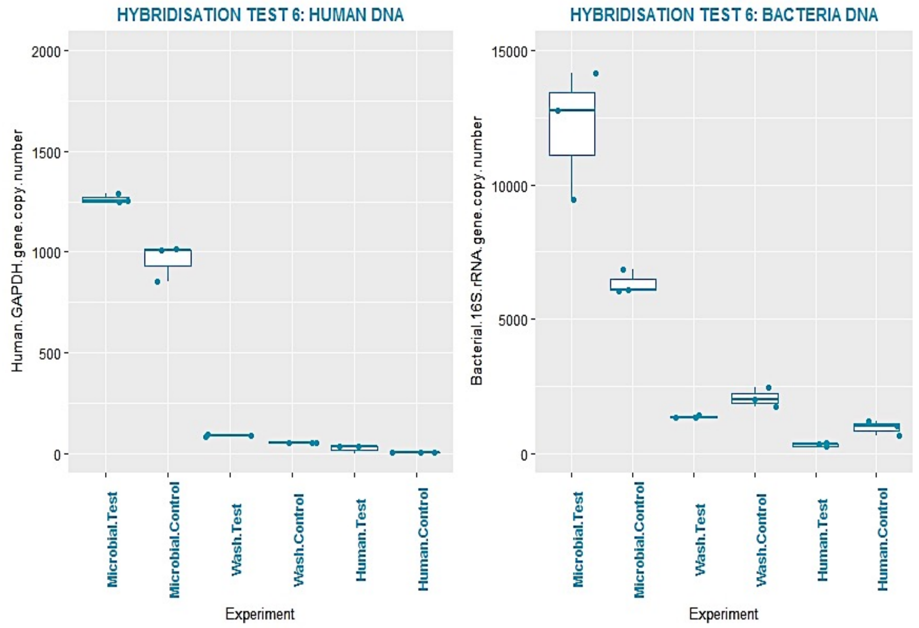


Figure 4.13 Boxplots of qPCR Results for Experiment 6. Human DNA presence determined by *GAPDH* copy number with Bacterial DNA presence determined by 16S rRNA copy number. Microbial, Wash and Human Test/Control = initial supernantant, supernatant from the washing step and supernatant after bead denaturation, respectively.

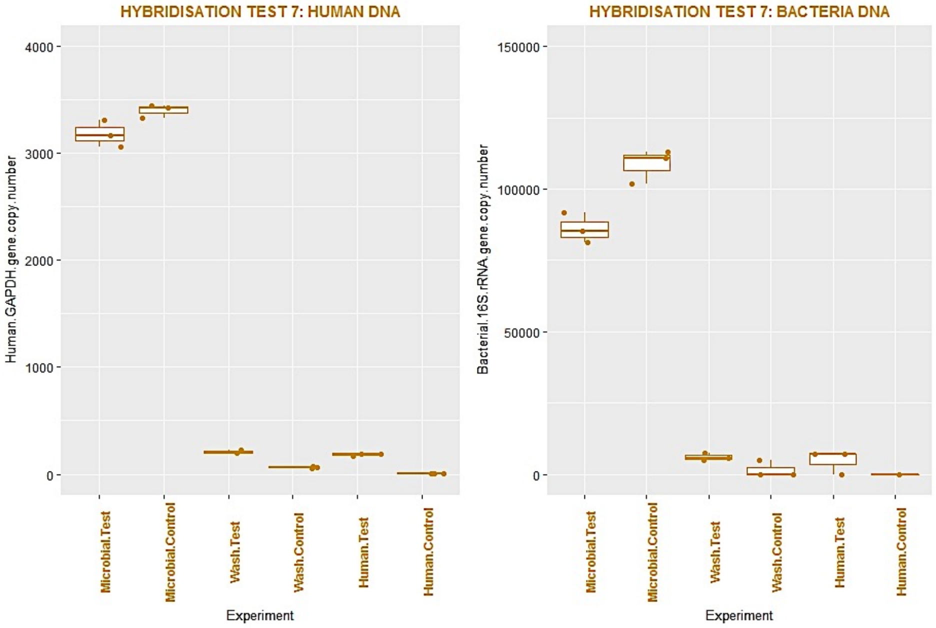


Figure 4.14 Boxplots of qPCR Results for Experiment 7. Human DNA presence determined by *GAPDH* copy number with Bacterial DNA presence determined by 16S rRNA copy number. Microbial, Wash and Human Test/Control = initial supernantant, supernatant from the washing step and supernatant after bead denaturation, respectively.

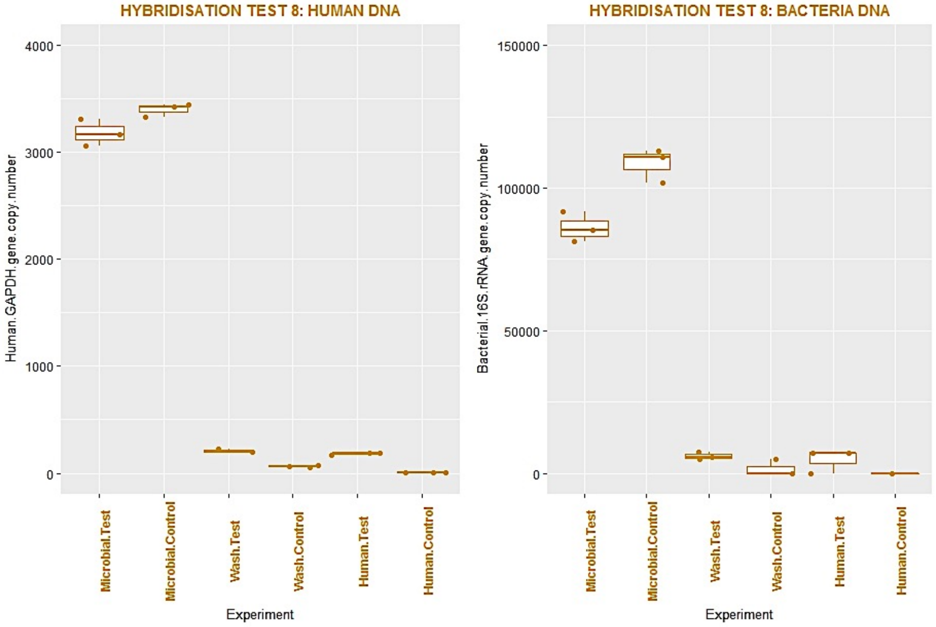


Figure 4.15 Boxplots of qPCR Results for Experiment 8. Human DNA presence determined by *GAPDH* copy number with Bacterial DNA presence determined by 16S rRNA copy number. Microbial, Wash and Human Test/Control = initial supernantant, supernatant from the washing step and supernatant after bead denaturation, respectively.

4.3.2 DNA: DNA Hybridisation Using Biotinylated *Alu* DNA Probes

4.3.2.1 *Alu* DNA Probe Generation and Biotinylation

4.3.2.1.(i) Methods

A primer pair designed by Lou *et al.*, 2014 for the amplification of an approximately 200 base pair region of the *Alu* element was chosen to generate the *Alu* hybridisation probe amplicon. The region spans the A-rich sequence that separates the dimers of the *Alu* element. The sequence for the forward primer was *S1F* (5'- *AGACCATCCTGGCTAACACG* - 3') and the sequence for the reverse primer was *A1F* (5'- *AGACGGAGTCTCGCTCTGTC* - 3'). Primers were sourced from Eurofins. All other PCR reagents were sourced from Sigma. Ethical considerations regarding usage of the template human genomic DNA and source are detailed in Chapter 2, Section 2.4. Each reaction was prepared as detailed in **Table 4.2**.

Table 4.2 *Alu* PCR - components for one reaction.

Component	Volume µl (X1 reaction)	Final Concentration in Reaction
10X PCR Buffer	10	1X
$MgCl_2$ (25mM)	6	1.5 mM
dGTP (10 mM)	3	300 µM
dATP	3	300 µM
dCTP	3	300 µM
dTTP	3	300 µM
Primer S1F (10 µM)	2	0.2 µM
Primer A1F (10 µM)	2	0.2 µM
Nuclease-free water	66.5	-
Template DNA (100 ng)	1	1 ng/µl
Taq DNA Polymerase (5 units/µl)	0.5	2.5 units

PCR was carried out in a thermal cycler under the following cycling conditions: 94 °C for 4 minutes, 35 cycles of (94 °C for 1 minute, 59 °C for 45 seconds, 68 °C for 1 minute) and a final extension of 68 °C for 5 minutes. The PCR reaction was purified using the NEB

Monarch PCR purification kit as described in Chapter 2, Section 2.7.3. and eluted in a final volume of 12 µl of 1 X Low TE. The concentration of the amplicon was measured using PicoGreen as described in Chapter 2, Section 2.6.1. Reactions were run on a 2 % agarose gel at 100 V for 1 hour and products visualised as described in Chapter 2, Section 2.8.2. To biotinylate the *Alu* DNA probes, 10 % biotin-16-dUTP (Roche) was added to the PCR reaction replacing 10 % of the dTTP. Each reaction was prepared as described in **Table 4.3**.

Table 4.3 *Alu* PCR with biotin - components for one reaction.

Component	**Volume µl (X1 reaction)**	**Final Concentration in Reaction**
10X PCR Buffer	10	1X
$MgCl_2$ (25mM)	6	1.5 mM
dGTP (10 mM)	3	300 µM
dATP (10 mM)	3	300 µM
dCTP (10 mM)	3	300 µM
dTTP (10 mM)	2.7	300 µM
Biotin-16-dUTP (1 mM)	3	30 µM
Primer S1F (10 µM)	2	0.2 µM
Primer A1F (10 µM)	2	0.2 µM
Nuclease-free water	66.5	-
Template DNA (100 ng)	1	1 ng/µl
Taq DNA Polymerase (5 units/µl)	0.5	2.5 units

PCR thermal cycling conditions were the same as for the *Alu* PCR without biotin (see above). The PCR reaction was purified using the NEB Monarch PCR purification kit as described in Chapter 2, Section 2.7.3. and eluted in 21 µl of nuclease-free water. The concentration of the amplicon was measured using the PicoGreen assay as described in Chapter 2, Section 2.6.1. Reactions were run on a 2 % agarose at 100 V for 1 hour and products visualised as described in Chapter 2, Section 2.8.2. A non-biotinylated *Alu* amplicon control and a negative control without template DNA were included at the PCR set up step and run alongside the biotinylated *Alu* on the agarose gel.

4.3.2.1.(ii) Results

The *Alu* DNA probe of approximately 200 base pairs was successfully generated by PCR (**Figure 4.16** duplicates run).

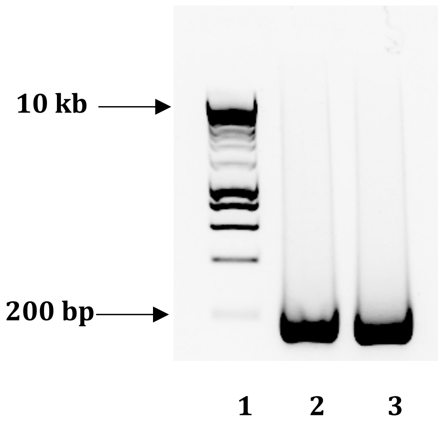


Figure 4.16 Non-biotinylated *Alu* amplicon size. Lane 1 = ladder, 2 = *Alu* amplicon, 3 = *Alu* amplicon.

A biotinylated probe containing 10 % biotin-16-dUTP was also successfully generated using the same PCR conditions.

4.3.2.2 *Alu* DNA: DNA Hybridisation Experiment 1

4.3.2.2.(i) Rationale

To qualitatively assess the depletion of human DNA fragments from a digested human genomic DNA sample using biotinylated *Alu* DNA probes and streptavidin-coated magnetic bead capture.

4.3.2.2.(ii) Methods

Human genomic DNA was digested as described in Chapter 2, Section 2.9.1. A hybridisation test composed of 1 µg of digested human DNA and 2 µg of biotinylated *Alu* DNA probes and a control composed of 1 µg of human DNA and 2 µg of non-biotinylated *Alu* DNA probes were prepared as follows:

Hybridisation Test

Component	Volume
Human genomic DNA (1 µg)	10 µl
Biotinylated *Alu* DNA probes (2 µg)	15 µl
Hybridisation Buffer (20X SSC)	6 µl

Hybridisation Control

Component	Volume
Human genomic DNA (1 µg)	10 µl
Non-biotinylated *Alu* DNA probes (2 µg)	15 µl
Hybridisation Buffer (20X SSC)	6 µl

Denaturation of both hybridisation test and hybridisation control was carried out in a thermal cycler at 95 °C for 5 minutes followed by a -0.1 °C ramp down to 65 °C. A temperature of 65 °C was maintained for 1 hour to enable hybridisation. A total of 100 µl of Dynabeads M-280 were washed for each reaction as described in Chapter 2, Section 2.9.3 and resuspended in 100 µl of 2 X Binding and Washing buffer. Each 31 µl reaction was made up to 100 µl with nuclease-free water and added to 100 µl of washed beads. The hyb-bead mixtures were incubated rotating for 30 minutes at room temperature. The tubes were centrifuged briefly and then incubated on a magnetic stand for 15 minutes. The post-capture supernatant was harvested and stored. The post-capture supernatant was purified using the NEB Monarch PCR purification kit as described in Chapter 2, Section 2.7.3 and eluted in 15 µl of nuclease-free water. Beads were washed twice with 200 µl of 1 X Low TE and supernatant was discarded. The beads were resuspended in 15 µl of nuclease-free water and denatured at 95 °C for 5 minutes.

After denaturation the tube was incubated on a magnetic stand for 5 minutes. The nuclease-free water from the denatured beads was harvested and not purified. The purified post-capture supernatant and the denatured bead supernatant were run on a 1.2 % agarose gel for 1 hour at 120 V and visualised as described in Chapter 2, Section 2.8.2. All steps were replicated for the hybridisation control. The harvesting process for post-capture supernatant and denatured beads supernatant is shown in **Figure 4.17**. A flowchart of the hybridisation and capture process for *Alu* DNA: DNA Hybridisation Experiment 1 is shown in **Figure 4.18**.

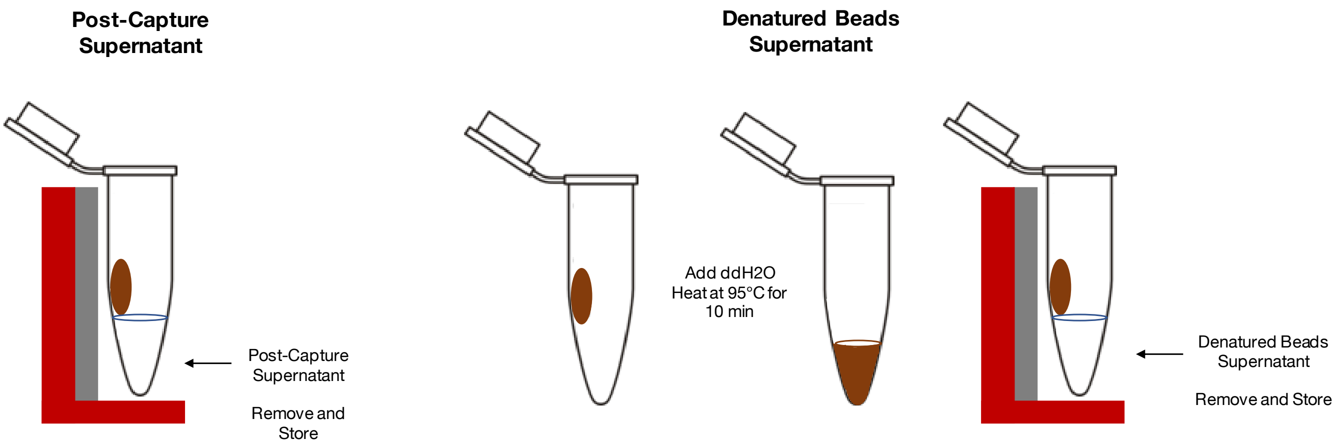

Post-Capture Supernatant
Post-Capture Supernatant
Remove and Store
Denatured Beads Supernatant
Add ddH2O Heat at 95°C for 10 min
Denatured Beads Supernatant
Remove and Store

Figure 4.17 Supernatant harvesting and storage process. Post-Capture supernatant harvesting and storage. Bead denaturation and denatured bead supernatant harvesting and stored.

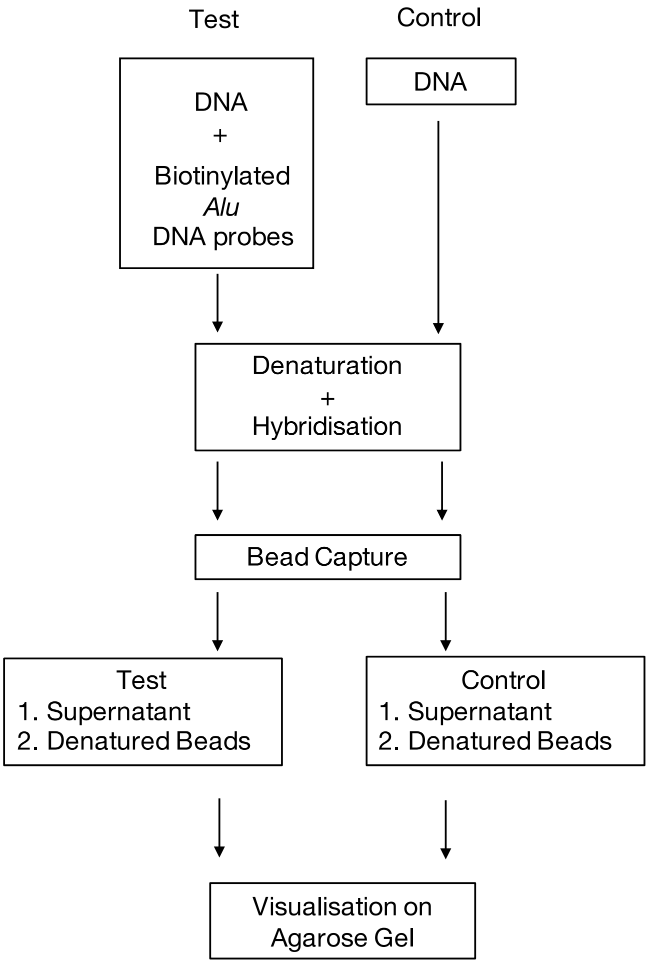

Test
Control
DNA + Biotinylated *Alu* DNA probes
DNA
Denaturation + Hybridisation
Bead Capture
Test 1. Supernatant 2. Denatured Beads
Control 1. Supernatant 2. Denatured Beads
Visualisation on Agarose Gel

Figure 4.18 Process flowchart of *Alu* DNA: DNA Hybridisation Experiment 1.

4.3.2.2.(iii) Results

A gel electrophoresis approach was carried out to qualitatively assess the depletion of human DNA from *Alu* test hybridisations due to the accumulating costs associated with quantitative PCR analysis.

Hybridisation Experiment 1 was carried out on digested human genomic DNA for a test (with biotinylated probe) and a control (non-biotinylated probe). **Figure 4.19** shows the 'Post-Capture Test' in lane 2, 'Post-Capture Control' in lane 3, the 'Denatured Bead Test' in lane 4 and 'Denatured Bead Control' in lane 5. Whilst there was no evidence of non-specific binding (no DNA in lane 5 of the control), there is limited evidence for DNA depletion. Lane 4 shows some depletion of short human DNA fragments however it is evident from lane 2 ('Post-Capture Test') that a visible quantity of the *Alu* probe is either not biotinylated (*Alu* probes at approximately 200 base pairs remaining in the supernatant post-capture) or that the streptavidin-coated magnetic beads had become saturated to capacity with biotinylated *Alu* elements and could not bind additional biotin moieties.

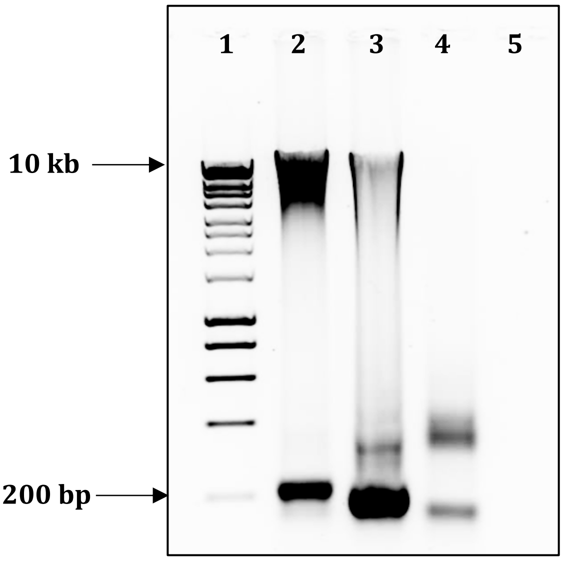


Figure 4.19 *Alu* DNA: DNA Hybridisation Experiment 1: Lane 1 = ladder, 2 = Post-Capture Test, 3 = Post-Capture Control, 4 = Denatured Beads Test and 5 = Denatured Beads Control.

4.3.2.3 *Alu* DNA: DNA Hybridisation Experiment 2

4.3.2.3.(i) Rationale

To test the hypothesis that beads were becoming saturated with biotinylated *Alu* DNA probes, sequential bead captures were carried out on a test (with biotinylated *Alu* DNA probes) and a control (with non-biotinylated *Alu* DNA probes).

4.3.2.3.(ii) Methods

Hybridisation Experiment 2 was carried out using 1 µg of human genomic DNA and using 1 µg of *Alu* DNA probes. This time three sequential streptavidin-coated bead captures were carried out on the "microbial" supernatant. A hybridisation control composed of 1 µg of human genomic DNA and 1 µg of non-biotinylated *Alu* DNA probes was also prepared:

Hybridisation Test

Component	Volume
Human genomic DNA (1 µg)	10 µl
Biotinylated *Alu* DNA probes (2 µg)	7.5 µl
Hybridisation Buffer (20X SSC)	6 µl
Nuclease-free water	7.5 µl

Hybridisation Control

Component	Volume
Human genomic DNA (1 µg)	10 µl
Non-biotinylated *Alu* DNA probes (2 µg)	7.5 µl
Hybridisation Buffer (20X SSC)	6 µl
Nuclease-free water	7.5 µl

A total of 100 µl of streptavidin beads for each reaction were washed as described in Chapter 2, Section 2.9.3 and three sequential captures were carried out: Capture 1 was carried out as described in Section 4.3.1.5.(i) above and the post capture and denatured beads supernatant were harvested. The Post-Capture Test 1 supernatant was split into two 100 µl aliquots. One 100 µl of aliquot was purified using the NEB Monarch PCR purification kit as described in Chapter 2, Section 2.7.3 and stored for subsequent gel visualisation. The other 100 µl aliquot was brought forward for a second capture reaction. Capture 2 was carried out as previously described in Section 4.3.1.5.(i) above, on 100 µl of Post-Capture Test 1 supernatant. The Post-Capture Test 2 supernatant and Denatured Beads Test 2 supernatant were harvested. The Post Capture Test 2 supernatant was split into two 100 µl aliquots. One 100 µl of aliquot was purified as described in Chapter 2, Section 2.7.3 and stored and the other was brought forward for a third capture reaction. Capture 3 was carried out again as previously described in Section 4.3.1.5.(i) above, on 100 µl of Post-Capture Test 2 supernatant. The Post-Capture Test 3 supernatant and Denatured Beads Test 3 supernatant were harvested. The Post-Capture Test 3 supernatant was purified and eluted in 15 µl of nuclease-free water as described in Chapter 2, Section 2.7.3.

For all three capture reactions, the same process was carried out for a hybridisation control reaction. Each 15 µl reaction was run on 1.2 % agarose gel at 120 V for 1 hour and visualised as described in Chapter 2, Section 2.8.2. The experimental steps are shown in **Figure 4.20**.

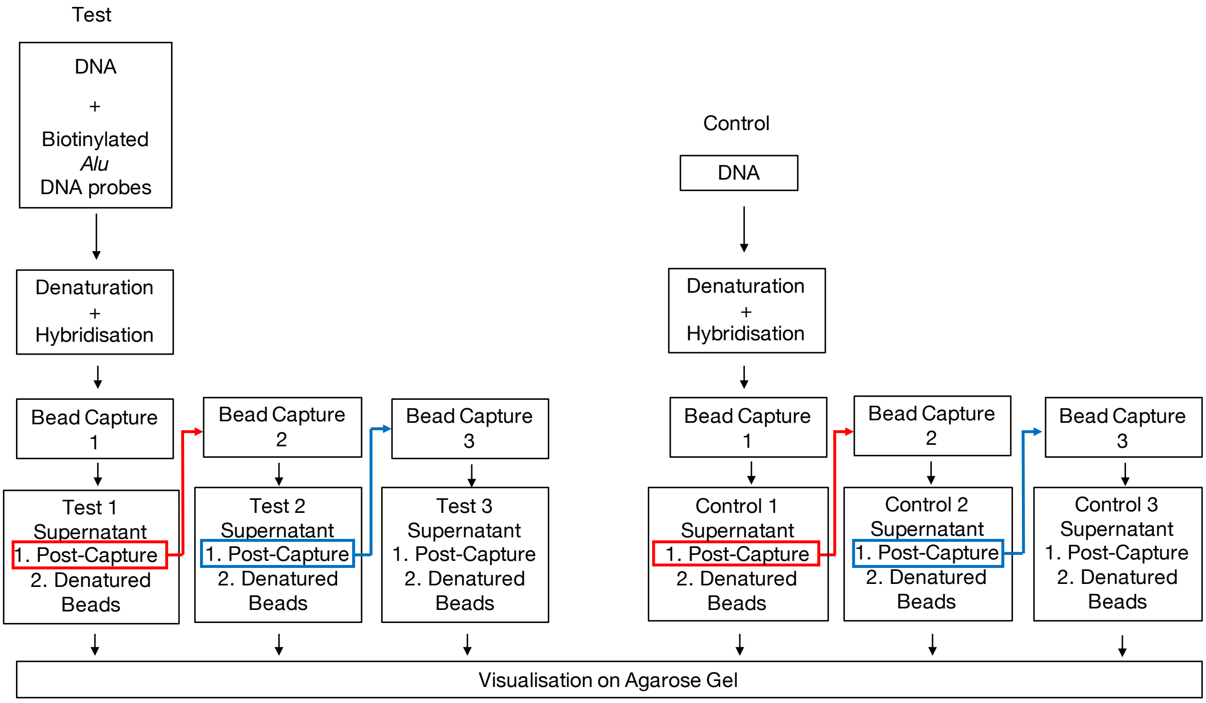


Figure 4.20 Process Flowchart of *Alu* DNA: DNA Hybridisaton Experiment 2.

4.3.2.3.(iii) Results

After three sequential captures, a visible proportion of the biotinylated *Alu* DNA probes remained in the supernatant. There was further evidence of a small amount of human DNA depletion from the first capture (Lane 4: Denatured Beads Test 1) however no further depletion was evident from Capture 2 (Lane 8: Denatured Beads Test 2) and Capture 3 (Lane 12: Denatured Beads Test 3) (**Figure 4.21**). It was concluded that due to lack of biotinylation of all *Alu* DNA probes as well as the potential rapid renaturation dynamics of the probes, a hybridisation approach using biotinylated *Alu* RNA probes would be investigated.

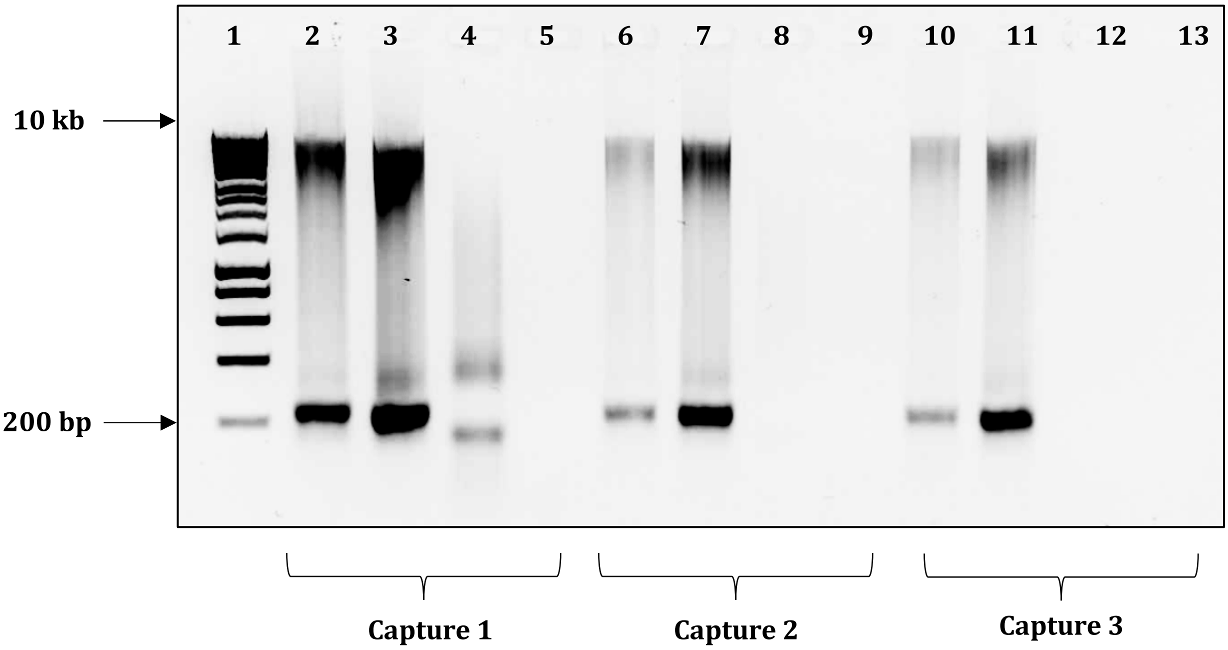


Figure 4.21 *Alu* DNA: DNA Hybridisation Experiment 2. Lane 1 = ladder, 2 = Post-Capture Test 1, 3 = Post-Capture Control 1, 4 = Denatured Beads Test 1, 5 = Denatured Beads Control 1, 6 = Post-Capture Test 2, 7 = Post-Capture Control 2, 8 = Denatured Beads Test 2, 9 = Denatured Beads Control 2, 10 = Post-Capture Test 3, 11 = Post-Capture Control 3, 12 = Denatured Beads Test 3, 13 = Denatured Beads Control 3.

4.3.3 RNA: DNA hybridisation with Biotinylated *Alu* RNA Capture Probes

4.3.3.1 T7-*Alu* DNA Template Preparation and Biotinylated *Alu* RNA Synthesis

4.3.3.1.(i) Methods

A PCR product can be used as template for *in vitro* transcription provided that it contains a double-stranded T7 promoter region upstream of the sequence to be transcribed. A T7-*Alu* PCR was carried out to incorporate a T7 sequence into the forward *Alu* sequence for subsequent RNA synthesis. The sequence of the forward primer was *T7-S1F* (5'- *TAATACGACTCACTATAGGGAGACCATCCTGGCTAACACG* -3'). The sequence of the reverse primer was *A1R* (5'- *AGACGGAGTCTCGCTCTGTC* -3'), as previously described. Primers were sourced from Eurofins. All other PCR and cycling conditions were maintained as described in Section 4.3.2.1.(i). A PCR reaction with the normal *S1F* and *A1R* primers as described in Section 4.3.2.1.(i) was prepared as a control. The components and volumes for the T7-*Alu* PCR are detailed in **Table 4.4.**

Table 4.4 T7-*Alu* PCR - components for one reaction.

Component	Volume µl (X1 reaction)	Final Concentration in Reaction
10X PCR Buffer with $MgCl_2$	10	1X
dGTP (10 mM)	3	300 µM
dATP (10 mM)	3	300 µM
dCTP (10 mM)	3	300 µM
dTTP (10 mM)	3	300 µM
Primer T7-S1F (10 µM)	2	0.2 µM
Primer A1F (10 µM)	2	0.2 µM
Nuclease-free water	66.5	-
Template DNA (100 ng)	1	1 ng/µL
Taq DNA Polymerase (5 units/µl)	0.5	2.5 units

PCR products were purified using the NEB Monarch PCR purification kit as described in Chapter 2, Section 2.7.3 and eluted in a volume of 21 µl. Reactions were run on a 2 % agarose gel at 120 V for 1 hour and visualised as described in Chapter 2, Section 2.8.2. RNA synthesis was carried out using the HiScribe T7 Quick High Yield RNA Synthesis Kit (NEB, Ipswich, MA, USA) and Biotin-16-UTP (Roche). The reaction was prepared as detailed in **Table 4.5**.

Table 4.5 Biotinylated *Alu* RNA Synthesis – components for one reaction.

Component	Volume µl (X1 reaction)	Final Concentration
NTP Buffer Mix	5	5 mM each NTP
Biotin-16-UTP (10 mM)	5	2.5 mM
T7-*Alu* DNA template	8	50 ng/µl
T7 RNA Polymerase Mix	2	-

The reaction was mixed thoroughly, centrifuged briefly and incubated in a thermal cycler at 37 °C for 16 hours. After the incubation period, template DNA was removed by adding 30 µl of nuclease-free water and 2 µl of DNase I (RNase-free) (NEB). The reaction was mixed and incubated at 37 °C for 15 minutes. Synthesised RNA was purified by phenol-chloroform extraction followed by ethanol precipitation according to the following steps: the reaction volume was adjusted to 180 µl by adding nuclease-free water after which a total of 20 µl of 3 M sodium acetate (SIGMA) was added and mixed thoroughly. The 200 µl volume was then added to a heavy phase-lock tube (VWR). Extraction was carried out by adding an equal volume 1:1 of phenol-chloroform to the mixture followed by two extraction with chloroform as described in Chapter 2, Section 2.5. The aqueous phase was collected and transferred to a new tube. RNA was precipitated by adding 2 volumes of 100 % ethanol. The reaction was incubated at –80 °C for 1 hour and the pellet was collected by centrifugation. The supernatant was removed, and the pellet was rinsed with 500 µl of ice cold 70 % ethanol. The RNA was resuspended in 50 µl 1 X Low TE and stored at –80 °C until use. The concentration of the purified RNA was measured using a NanoDrop spectrophotometer (Thermo Fisher Scientific).

4.3.3.1.(ii) Results

A non-biotinylated *Alu* element with a T7 sequence was successfully generated by PCR. **Figure 4.22 (A)** shows the T7-*Alu* element in lane 2 and the normal *Alu* element (no T7 sequence) in lane 3. *Alu* RNA probes were then generated by transcribing the T7 *Alu* forward strand **Figure 4.22 (B)**.

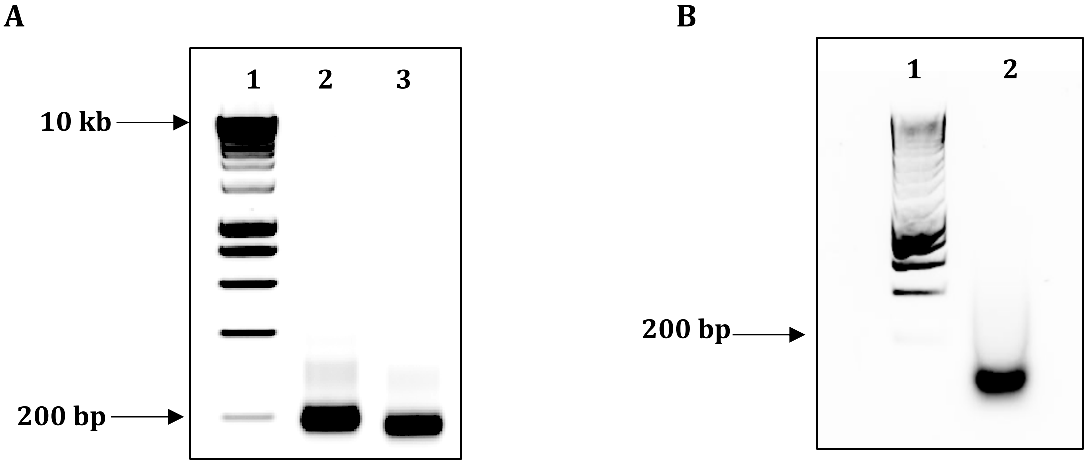


Figure 4.22 T7-*Alu* probe generation. **(A)** T7-*Alu* probe generation. Lane 1 = ladder, 2 = T7-*Alu* DNA element, 3 = normal *Alu* DNA element. **(B)** Lane 1 = ladder, 2 = *Alu* RNA probe.

4.3.3.2 *Alu* RNA: DNA Hybridisation Experiment 1

4.3.3.2.(i) Rationale

A control hybridisation experiment was performed to qualitatively assess if bacterial DNA would bind non-specifically to *Alu* probes or streptavidin-coated beads.

4.3.3.2.(ii) Methods

Bacterial DNA (*L. longbeachae* (see Chapter 2, Section 2.4 for further details) was digested using *EcoRI* as described in Chapter 2, Section 2.9.1 as required. The reaction was purified using the NEB Monarch PCR purification kit as described in Chapter 2, Section 2.7.3 and the concentration was measured by the PicoGreen assay as described in Chapter 2, Section 2.6.1. A total of 500 ng of bacterial genomic DNA was spiked with 0.5 µg FluC plasmid (NEB). Two reactions were prepared in 1.5 ml Eppendorf DNA LoBind tubes, a test hybridisation containing the *Alu* RNA probes and a no probe control. Bead washing for RNA applications (see Chapter 2, Section 2.9.4) and hybridisation reactions (see

Chapter 2, Section 2.9.5) were carried out. Bead capture was carried out as described in Chapter 2, Section 2.9.6. Purifications were carried out as described in Chapter 2, Section 2.7.3 and gel visualisation on a 1.2 % agarose gel (run for 1 hour at 120 V) was carried out as described in Chapter 2, Section 2.8.2. The experimental process is illustrated in **Figure 4.23**.

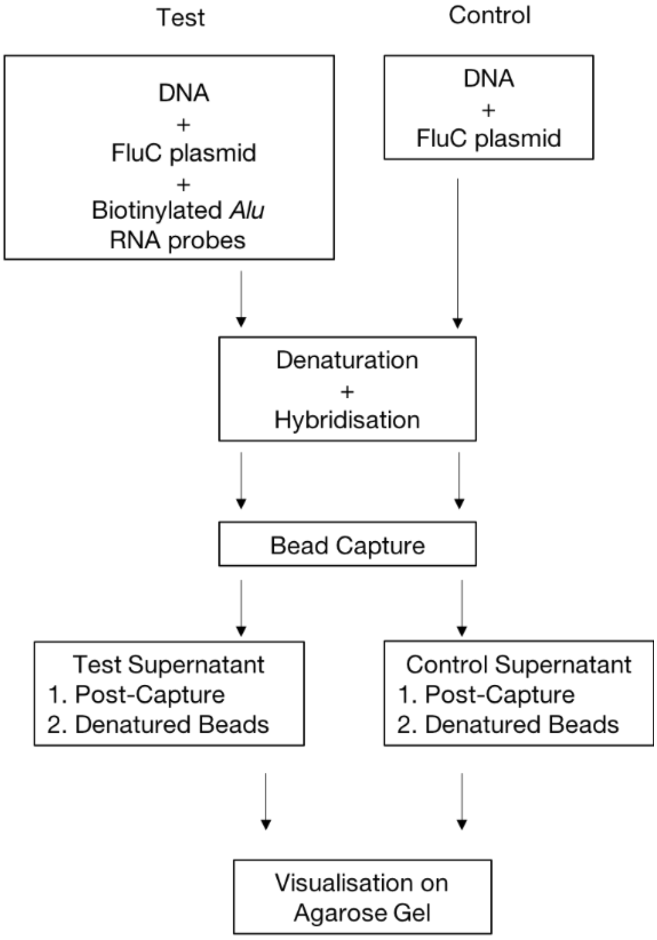

Test
Control
DNA + FluC plasmid + Biotinylated *Alu* RNA probes
DNA + FluC plasmid
Denaturation + Hybridisation
Bead Capture
Test Supernatant 1. Post-Capture 2. Denatured Beads
Control Supernatant 1. Post-Capture 2. Denatured Beads
Visualisation on Agarose Gel

Figure 4.23 Process Flowchart of *Alu* RNA: DNA Hybridisation Experiment 1.

4.3.3.2.(iii) Results

A hybridisation test (with biotinylated *Alu* RNA probes) and a control (no probes) with size selected *L. longbeachae* DNA spiked with a FluC plasmid were carried out to test for non-specific or probe-driven binding of bacterial DNA to the beads. The 'Denatured Bead Test' (lane 4) and 'Denatured Bead Control' (lane 5) did not contain evidence of DNA fragments confirming that there had been not binding of bacterial DNA to the beads (**Figure 4.24**).

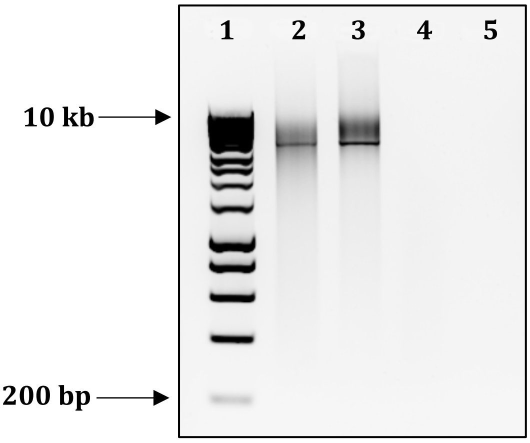


Figure 4.24 *Alu* RNA: DNA Hybridisation Experiment 1. Lane 1 = ladder, 2 = Post-Capture Test, 3 = Post-Capture Control, 4 = Denatured Bead Test, 5 = Denatured Bead Control.

4.3.3.3 *Alu* RNA: DNA Hybridisation Experiment 2

4.3.3.3.(i) Rationale

To qualitatively assess the depletion of human DNA fragments from mock material using biotinylated *Alu* RNA probes and streptavidin-coated magnetic bead capture. Bacterial DNA was included as a control.

4.3.3.3.(ii) Methods

Bacterial genomic DNA (*L. pneumophila* Philadelphia-1) and human genomic DNA were digested using *EcoRI* (see Chapter 2, Section 2.9.1) and purified (see Chapter 2, Section 2.7.3).

The following 3 mock samples were prepared:

1. 100 % bacterial genomic DNA (*L. pneumophila* Philadelphia-1)
2. 50 % bacterial genomic DNA (*L. pneumophila* Philadelphia-1) + 50 % human genomic DNA
3. 100 % human genomic DNA

Three hybridisation reactions were prepared containing 1 µg each of the digested mock samples, 0.87 µg of biotinylated *Alu* RNA probes and 9 µl of nuclease-free water. Bead washing for RNA applications (see Chapter 2, Section 2.9.4), hybridisation (see Chapter 2, Section 2.9.5) bead capture (see Chapter 2, Section 2.9.6), purification (see Chapter 2, Section 2.7.3) and gel visualisation on a 1.2 % agarose gel (run for 1 hour at 120 V) (see Chapter 2, Section 2.8.2) were carried out. The experimental process is shown in **Figure 4.25**.

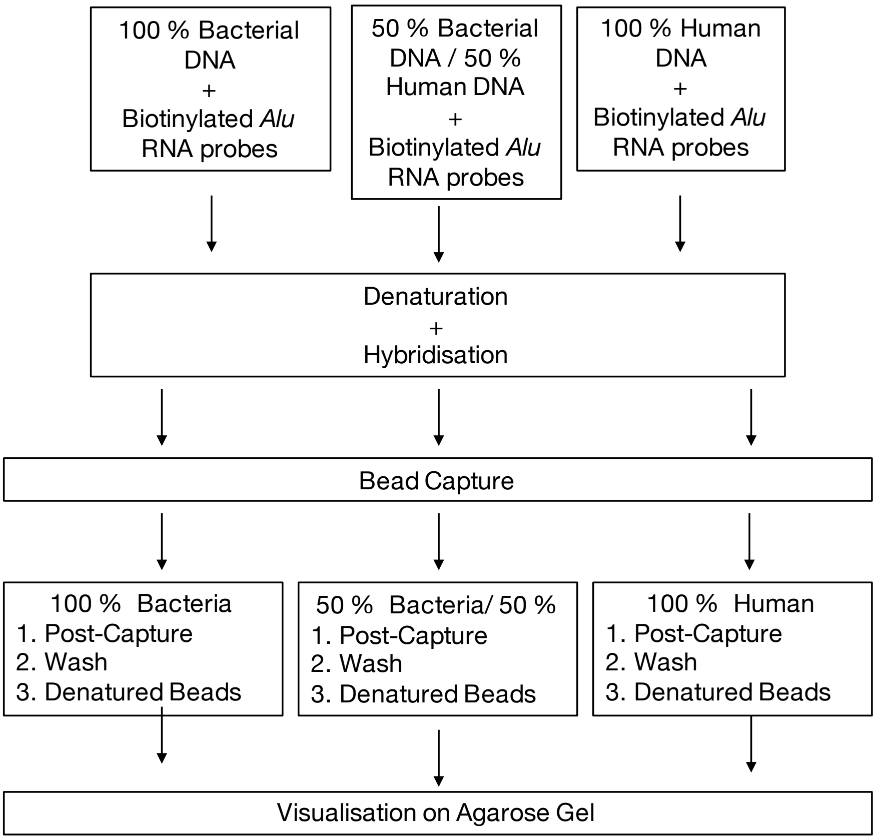

100 % Bacterial DNA + Biotinylated *Alu* RNA probes
50 % Bacterial DNA / 50 % Human DNA + Biotinylated *Alu* RNA probes
100 % Human DNA + Biotinylated *Alu* RNA probes
Denaturation + Hybridisation
Bead Capture
100 % Bacteria 1. Post-Capture 2. Wash 3. Denatured Beads
50 % Bacteria/ 50 % 1. Post-Capture 2. Wash 3. Denatured Beads
100 % Human 1. Post-Capture 2. Wash 3. Denatured Beads
Visualisation on Agarose Gel

Figure 4.25 Process Flowchart of *Alu* RNA: DNA Hybridisation Experiment 2.

4.3.3.3.(iii) Results

Hybridisation experiments were carried out on three tests: 100 % bacterial DNA, a mock sample composed of 50 % bacterial and 50 % human DNA and 100 % human DNA. In **Figure 4.26**, lanes 9 and 10 show evidence of minor depletion of human DNA however there is also some evidence of non-specific binding of bacterial DNA as can be seen in the Bacterial Wash (lane 5) and the Bacteria Denatured Beads (lane 8).

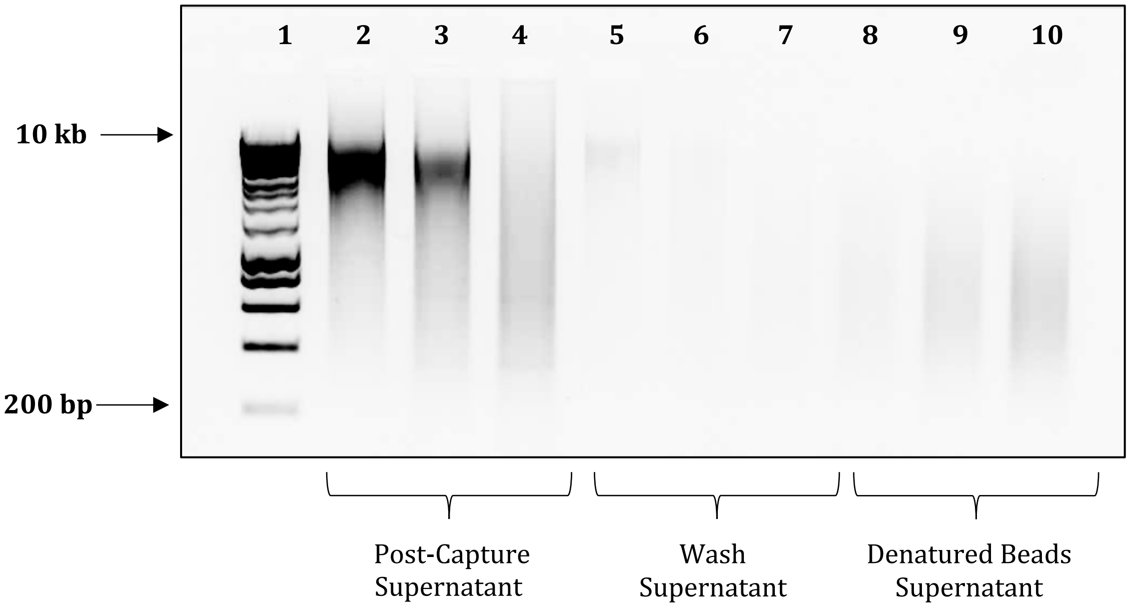


Figure 4.26 *Alu* RNA: DNA Hybridisation Experiment 2. Lane 1 = marker, 2 = Bacteria Post-Capture, 3 = Bacteria/Human Post-Capture, 4 = Human Post-Capture, 5 = Bacteria Wash, 6 = Bacteria/Human Wash, 7 = Human Wash, 8 = Bacteria Denatured Beads, 9 = Bacteria/Human Denatured Beads, 10 = Human Denatured Beads.

4.3.3.4 *Alu* RNA: DNA Hybridisation Experiment 3

4.3.3.4.(i) Rationale

To qualitatively assess the depletion of human DNA fragments from digested human genomic DNA samples using biotinylated *Alu* RNA probes and streptavidin-coated magnetic bead capture at two hybridisation time points: 5 minutes and 1 hour. Bacterial DNA and human DNA without probe were included as controls.

4.3.3.4.(ii) Methods

Bacterial genomic DNA (*L. pneumophila* Philadelphia-1) and human genomic DNA were digested (see Chapter 2, Section 2.9.1) and purified (see Chapter 2, Section 2.7.3).

The following hybridisation tests were prepared with 0.98 µg of biotinylate *Alu* RNA probes:

1. 100 % human DNA: hybridisation carried out at 65 °C for 5 minutes.
2. 100 % human DNA: hybridisation carried out at 65 °C for 1 hour.
3. 100 % human DNA (negative control – no probe added): hybridisation carried out 65 °C for 1 hour.
4. 100 % bacterial DNA (bacterial control with probe): hybridisation carried out at 65 °C for 1 hour.

Hybridisation steps were carried out as described in Chapter 2, Section 2.9.5 with the exception of test 1 where hybridisation was carried out for 5 minutes.

Bead washing (see Chapter 2, Section 2.9.4), bead capture (see Chapter 2, Section 2.9.6), reaction purification (see Chapter 2, Section 2.7.3) and gel visualisation on a 1.2 % agarose gel (run for 1 hour at 120 V) were carried out as described in Chapter 2, Section 2.8.2. The experimental process is illustrated in **Figure 4.27**.

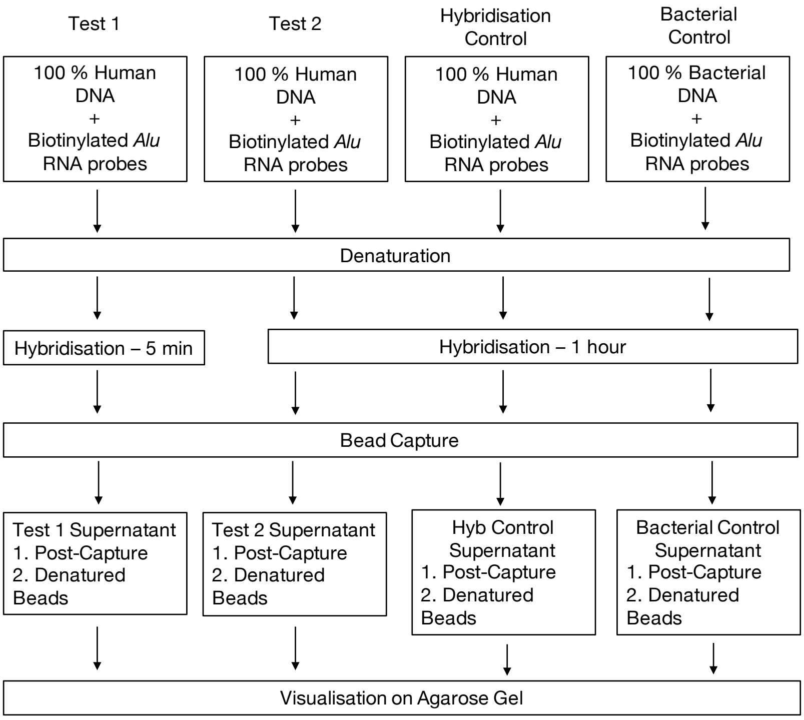

Test 1
Test 2
Hybridisation Control
Bacterial Control
100 % Human DNA + Biotinylated *Alu* RNA probes
100 % Human DNA + Biotinylated *Alu* RNA probes
100 % Human DNA + Biotinylated *Alu* RNA probes
100 % Bacterial DNA + Biotinylated *Alu* RNA probes
Denaturation
Hybridisation – 5 min
Hybridisation – 1 hour
Bead Capture
Test 1 Supernatant 1. Post-Capture 2. Denatured Beads
Test 2 Supernatant 1. Post-Capture 2. Denatured Beads
Hyb Control Supernatant 1. Post-Capture 2. Denatured Beads
Bacterial Control Supernatant 1. Post-Capture 2. Denatured Beads
Visualisation on Agarose Gel

Figure 4.27 Process flowchart for *Alu* RNA: DNA Hybridisation Experiment 3.

4.3.3.4.(iii) Results

Hybridisation experiments were carried out on 100 % human DNA with a hybridisation step of 5 minutes, 100 % human DNA with a hybridisation time of 1 hour, a control hybridisation (no probe) with a hybridisation time of 1 hour and a bacterial control hybridisation (with probe) with a hybridisation time of 1 hour. It was evident from lane 3 (Human DNA Denatured Beads Test 1 [5 minutes]) and lane 5 (Human DNA Denatured Beads Test 2 [1 hour]) of **Figure 4.28** that a small amount of human DNA was depleted after both 5 minutes and 1 hour hybridisation times. There was no evidence of non-specific binding of human DNA to beads (lane 7: Hybridisation Control Denatured Beads) and minor binding of shorter bacterial DNA fragments from the Bacterial DNA Control (lane 9: Bacterial DNA Control Denatured Beads).

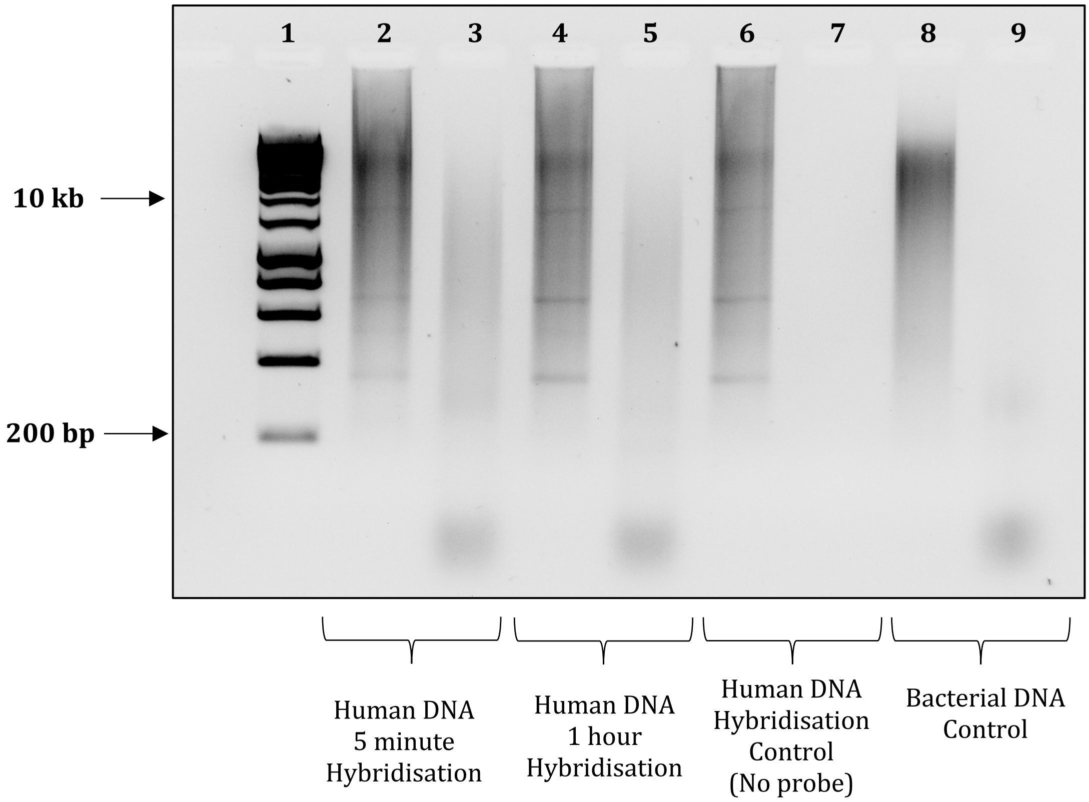

1
2
3
4
5
6
7
8
9
10 kb
200 bp
Human DNA 5 minute Hybridisation
Human DNA 1 hour Hybridisation
Human DNA Hybridisation Control (No probe)
Bacterial DNA Control

Figure 4.28 *Alu* RNA: DNA Hybridisation Experiment 3. Lane 1 = ladder, 2 = Human DNA Post-Capture Test 1 (5 minute hybridisation), 3 = Human DNA Denatured Beads Test 1 (5 minutes), 4 = Human DNA Post-Capture Test 2 (1 hour hybridisation), 5 = Human DNA Denatured Beads Test 2 (1 hour), 6 = Hybridisation Control Post-Capture (no probe), 7 = Hybridisation Control Denatured Beads, 8 = Bacterial DNA Control Post-Capture (with probe) and 9 = Bacterial DNA Control Denatured Beads.

4.3.3.5 *Alu* RNA: DNA Hybridisation Experiment 4 and Sequencing

4.3.3.5.(i) Rationale

Based on the results of *Alu* RNA hybridisation experiments, hybridisation experiments were carried out on mock material and three real sputum extracted nucleic acid samples and the post-capture supernatants were sequenced.

4.3.3.5.(ii) Methods

Bacterial genomic DNA (*L. pneumophila* Philadelphia-1, *L. pneumophila* OLDA, *L. longbeachae*, *S. pneumoniae, H. influenzae* and *V. dispar*) and human genomic DNA was digested as described in Chapter 2, Section 2.9.1. A mock community was assembled the components of which are detailed in **Table 4.6**.

Table 4.6 Mock Community for *Alu* RNA Hybridisation Experiment 4.

Component	Percentage
Human genomic DNA	90 %
L. pneumophila Philadelphia-1	2 %
L. pneumophila OLDA	2 %
L. longbeachae	2 %
S. pneumoniae	2 %
H. influenza	2 %
V. dispar	2 %

Three respiratory samples: 24A, 24Be and S2 (with appropriate ethical permissions) were sourced from prior projects within the Genomic Medicine Section, NHLI. Samples were extracted, precipitated and purified as described in Chapter 2, Section 2.5. Respiratory sample fragment lengths ranged from 4 to 10 kb therefore samples were not digested. Bead washing (see Chapter 2, Section 2.9.4), hybridisation (see Chapter 2, Section 2.9.5) and bead capture (see Chapter 2, Section 2.9.6) were carried out. Hybridisation tests and controls were designed as follows:-

- Mock Community (Control – not depleted)
- Mock Community (Hybridisation Control – no probe)
- Mock Community X 1 hybridisation and capture

- Mock Community X 2 sequential hybridisations and capture
- Mock Community X 3 sequential hybridisations and capture
- Mock Community Illumina library X 1 hybridisation and capture

- 24A (Control – not depleted)
- 24A X 1 hybridisation and capture
- 24A X 2 sequential hybridisations and capture
- 24A X 3 sequential hybridisations and capture

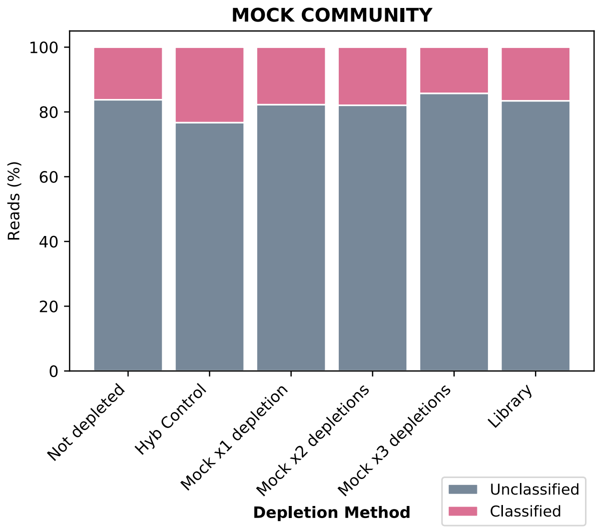
- 24Be (Control – not depleted)
- 24Be X 1 hybridisation and capture
- 24Be X 2 sequential hybridisations and capture
- 24Be X 3 sequential hybridisations and capture

- S2 (Control – not depleted)
- S2 X 1 hybridisation and capture
- S2 X 2 sequential hybridisations and capture
- S2 X 3 sequential hybridisations and capture

AMPure XP purification was carried out as described in Chapter 2, Section 2.7 on all harvested supernatants. Libraries were prepared and sequenced as described in Chapter 2, Section 2.10 and using the JetSeq ™ Library Quantification Lo-ROX Kit (see Chapter 2, Section 2.10.5.2). Sequencing was carried out on an Illumina MiSeq platform. Sequence data was cleaned and quality filtered as described from Section 2.12.2 to Section 2.12.4, Chapter 2. Taxonomic classification to determine the proportion of unclassified and classified reads was carried out using `Centrifuge` (Version 1.0.3) as described in Chapter 2, Section 2.13.1.

4.3.3.5.(iii) Results

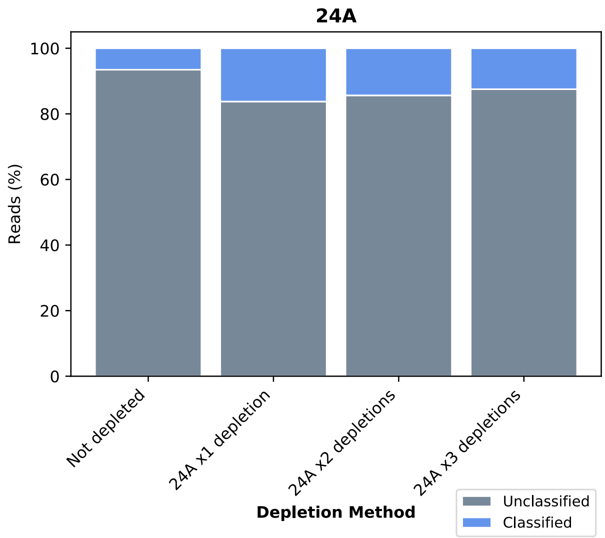
The depletion of human DNA from a mock community and from DNA extracted from three sputum samples using the biotinylated *Alu* RNA probes was investigated. Hybridisation and capture experiments were carried out for a total of 18 tests. **Figures 4.29** to **4.32** detail the proportion of classified (microbial) and unclassified reads, therefore indicating success or failure of the depletion experiments. The original mock community was prepared to contain 90 % human DNA and 10 % microbial DNA. A total of 16.24 % of reads were classified and 83.75 % of reads were unclassified. This non-depleted mock was taken forward as the baseline for comparison of the mock tests and controls. The mock hybridisation control (no probe) contained a classified read proportion of 23.3 % and 76.7 %. This may be indicatory of re-association dynamics whereby in the absence of probe, a limited quantity of human DNA reads renatured when compared to bacterial DNA. The first, second and third sequential mock capture experiments did not show evidence of human DNA depletion. The same was true for the mock metagenomic Illumina library where the unclassified and classified proportions were maintained when compared to the non-depleted mock (**Figure 4.29**).



Mock	Unclassified	Classified
Not depleted	83.75 %	16.24 %
Control (no probe)	76.7 %	23.3 %
1 depletion	82.24 %	17.75 %
2 depletions	82.05 %	17.94 %
3 depletions	85.74 %	14.26 %
Metagenomic Illumina library	83.4 %	16.56 %

Figure 4.29 Mock Community: Pre- and post-capture classified and unclassified reads. (Mock community composition is detailed in **Table 4.6** above).

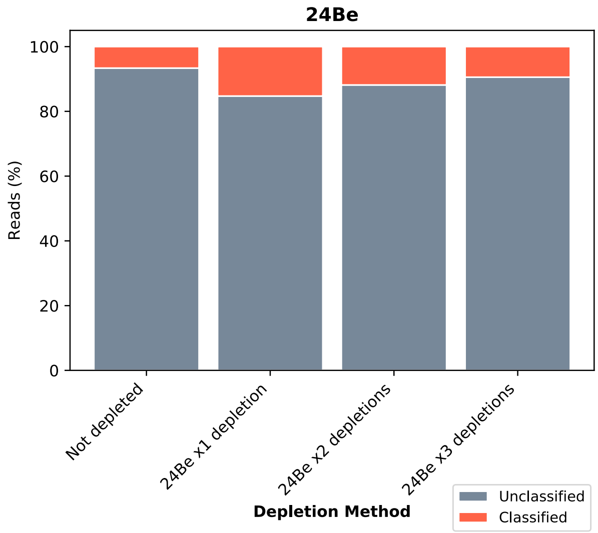
The non-depleted 24A sample was demonstrated to represent a total of 93.45 % unclassified and 6.54 % classified reads. After the first depletion a total of 16.25 % of reads were classified however the number began to decrease for each subsequent depletion (**Figure 4.30**).



24A	Unclassified	Classified
Not depleted	93.45 %	6.54 %
1 depletion	83.75 %	16.25 %
2 depletions	85.63 %	14.37 %
3 depletions	87.51 %	12.48 %

Figure 4.30 Sample 24A: Pre- and post-capture classified and unclassified reads.

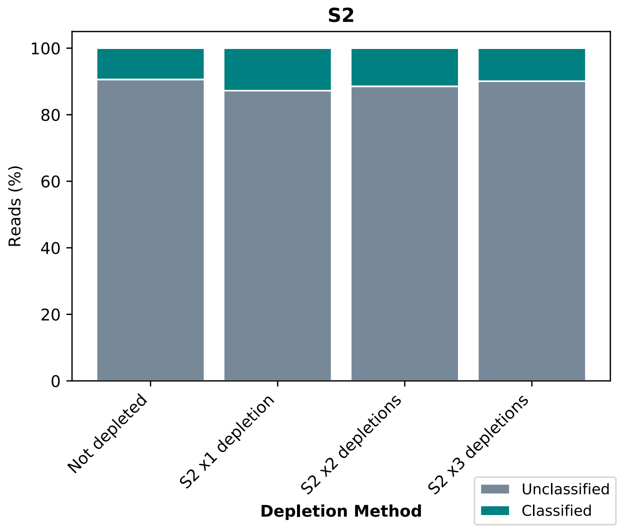
Similarly, the non-depleted 24Be sample was demonstrated to represent a total of 93.35 % unclassified and 6.64 % classified reads. After the first depletion a total of 15.27 % of reads were classifiable however the number began to decrease for each subsequent depletion (**Figure 4.31**).



24Be	Unclassified	Classified
Not depleted	93.35 %	6.64 %
1 depletion	84.72 %	15.27 %
2 depletions	88.15 %	11.85 %
3 depletions	90.55 %	9.44 %

Figure 4.31 Sample 24Be: Pre- and post-capture classified and unclassified reads.

Sample S2 again showed a similar dynamic with 9.40 % of reads classified in the non-depleted sample, 12.78 % of reads classified after the first depletion and the number of classifiable reads for depletion 2 and 3 decreasing thereafter (**Figure 4.32**).



S2	Unclassified	Classified
Not depleted	90.59 %	9.40 %
1 depletion	87.22 %	12.78 %
2 depletions	88.53 %	11.47 %
3 depletions	90.06 %	9.93 %

Figure 4.32 Sample S2: Pre- and post-capture classified and unclassified reads.

4.3.4 Investigating the Depletion of Human DNA from Illumina Metagenomic Libraries

4.3.4.1 Rationale

Owing to the lack of or minor depletion of human DNA from samples sequenced in the previous experiment, it was hypothesised that the size of the genomic DNA fragments was hindering successful capture. Metagenomic Illumina libraries were therefore prepared and capture experiments carried out using the Cot-1 DNA and *Alu* RNA approaches as previously described above in addition to a single cycle *Alu* PCR approach to incorporate biotin.

4.3.4.2 Methods

A total of 7 metagenomic libraries were prepared from an undigested mock community (for mock community composition please see **Table 4.6**), 4 libraries were prepared from sample 24A and 4 libraries were prepared from sample 24Be as described from Section 2.10.1 to 2.10.2 Chapter 2.

Three depletion methods were carried out on the libraries after the post-amplification purification:

1. depletion with biotinylated Cot-1 DNA probes.
2. depletion with biotinylated *Alu* RNA probes.
3. depletion by a single cycle *Alu* PCR (94 °C for 5 minutes, 59 °C for 45 seconds and 68 °C for 1 minute) to incorporate biotin-16-dUTP.

Hybridisation tests and controls were designed as follows:-

- Mock Community: Control – not depleted
- Mock Community: depletion with biotinylated Cot-1
- Mock Community: hybridisation control for Cot-1 reaction – no probe added
- Mock Community: depletion with biotinylated *Alu* RNA probes
- Mock Community: hybridisation control for *Alu* reaction – no probe added
- Mock Community: depletion with single cycle *Alu* PCR
- Mock Community: hybridisation control for single cycle *Alu* – no primers/biotin added

- 24A: Control – not depleted
- 24A: depletion with biotinylated Cot-1
- 24A: depletion with biotinylated *Alu* RNA probes
- 24A: depletion with single cycle *Alu* PCR

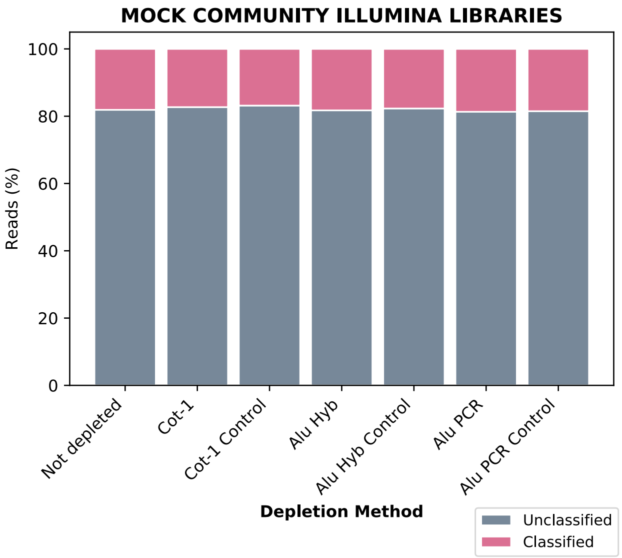
- 24Be: Control – not depleted
- 24Be: depletion with biotinylated Cot-1
- 24Be: depletion with biotinylated *Alu* RNA probes
- 24Be: depletion with single cycle *Alu* PCR

Bead washing (see Chapter 2, Section 2.9.4), hybridisation (see Chapter 2, Section 2.9.5) and bead capture (see Chapter 2, Section 2.9.6) were carried out. Libraries were purified again (see Chapter 2, Section 2.10.2) then pooled and quantified (see Chapter 2, Section 2.10.5.2). The quality and fragment size of the pooled library was analysed (see Chapter 2, Section 2.10.6) and sequenced was carried out (see Chapter 2, Section 2.10.7). Sequence data was cleaned and quality filtered (see Sections 2.12.2 to 2.12.4, Chapter 2).

Taxonomic classification to determine the proportion of unclassified and classified reads, was carried out using `Centrifuge` (Version 1.0.3) as described in Chapter 2, Section 2.13.1.

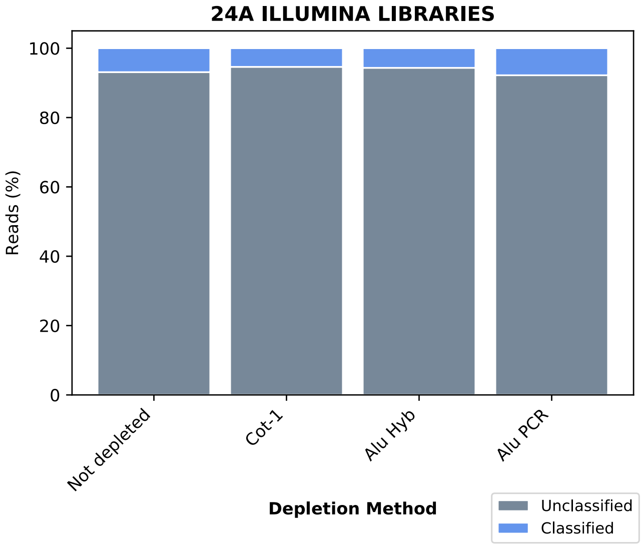
4.3.4.3 Results

The depletion of human DNA from Illumina metagenomic libraries was investigated for a mock community and two sputum samples using biotinylated Cot-1 DNA probes, biotinylated *Alu* RNA probes and a single cycle *Alu* PCR with biotin. Hybridisation and capture experiments were carried out for a total of 15 libraries. The proportion of classified (microbial) and unclassified reads indicated the success or failure of the depletion experiments when compared with the non-depleted control. There was no evidence of depletion from the Illumina libraries for the mock community (**Figure 4.33**), sample 24A (**Figure 4.34**) and sample 24Be (**Figure 4.35**), with unclassified and classified reads from depletion tests representing similar proportions to the Not depleted sample in each case.



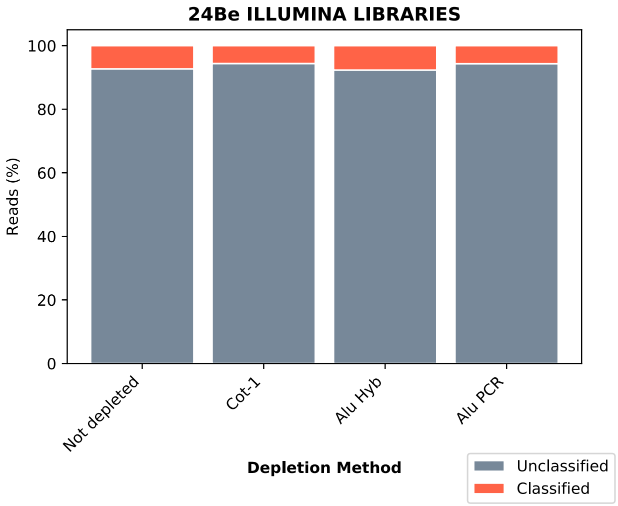
Mock	Unclassified	Classified
Not depleted	81.91 %	18.06 %
Cot-1 depletion	82.7 %	17.30 %
Cot-1 control (no probe)	83.15 %	16.84 %
Alu RNA depletion	81.74 %	18.26 %
Alu RNA control (no probe)	82.3 %	17.7 %
Alu PCR depletion	81.3 %	18.69 %
Alu PCR control (no probe)	81.49 %	18.51 %

Figure 4.33 Mock Community Illumina Libraries: Pre- and post-capture classified and unclassified reads.



24A	Unclassified	Classified
Not depleted	93.09 %	6.9 %
Cot-1 depletion	94.61 %	5.38 %
Alu RNA depletion	94.34 %	5.66 %
Alu PCR depletion	92.2 %	7.8 %

Figure 4.34 Sample 24A Illumina Libraries: Pre- and post-capture classified and unclassified reads.



24Be	Unclassified	Classified
Not depleted	92.68 %	7.32 %
Cot-1 depletion	94.37 %	5.62 %
Alu RNA depletion	92.32 %	7.68 %
Alu PCR depletion	94.31 %	5.69 %

Figure 4.35 Sample 24Be Illumina Libraries: Pre- and post-capture classified and unclassified reads.

4.3.5 Investigating the Depletion of Human DNA from an Oxford Nanopore Library

4.3.5.1 Rationale

To investigate the depletion of human DNA from a sputum sample with fragment lengths of 1.5 to 2 kb using a single cycle *Alu* PCR approach. Owing to the failure of previous experiments, decreasing fragment lengths to below 2 kb could allow bead binding without steric interference and allow *Alu* oligos to incorporate biotin into the forward and reverse strands of a sufficient number of fragments enabling efficient depletion of human DNA from a sequencing library.

4.3.5.2 Methods

4.3.5.2.(i) Step 1 – Shearing, Repair and PCR Adapter Ligation

A single cycle *Alu* PCR was investigated for the depletion of human DNA fragments from an Oxford Nanopore library using a modified version of the protocol for low input genomic DNA (PCR-based) for the 1D SQK-LSK108 kit (Oxford Nanopore Technologies). A total of 4 μg of sample 24A was sheared to between 1.5 – 2 kb fragments with a Covaris microtube using a Covaris M220 Focused Ultrasonicator. The sample was divided into two 2 μg aliquots: 2 μg for a depletion test and 2 μg for a not depleted control. For the test and control sample, FFPE (Formalin-fixed paraffin-embedded) DNA repair was carried out as described in Chapter 2, Section 2.14.2 with end-repair and PCR adapter ligation carried out as described in Chapter 2, Sections 2.14.3 and 2.14.4, respectively. The control (non-depleted) reaction was stored in a 1.5 mL Eppendorf DNA LoBind tube at 4 °C and the test (depleted) reaction was taken forward to Step 2.

4.3.5.2.(ii) Step 2– Single Cycle *Alu* PCR

A single cycle *Alu* PCR reaction was carried out on the test sample to incorporate biotin into the human DNA *Alu* elements. The PCR adapter-ligated test was added to a 0.2 ml PCR tube along with PCR reagents as detailed in **Table 4.7**.

Table 4.7 Single Cycle *Alu* PCR Reaction for biotin incorporation into an ONT Library. Components for one reaction.

Component	Volume µl (X 1 reaction)	Final Concentration in Reaction
10X PCR Buffer with $MgCl_2$	10	1 X
dGTP (10 mM)	3	300 µM
dCTP (10 mM)	3	300 µM
dATP (10 mM)	3	300 µM
dTTP (10 mM)	2.1	210 µM
Biotin-16-dUTP (1 mM)	9	90 µM
Forward Primer (*S1-F*) (10 µM)	3	0.3 µM
Reverse Primer (*A1-F*) (10 µM)	3	0.3 µM
Adapter-ligated DNA	40	23.4 ng/µl
Nuclease-free water	23.4	-
Taq DNA Polymerase	0.5 (5 units/µl)	2.5 units

Thermal cycling was carried out under the following conditions: denaturation at 94 °C for 5 minutes, annealing at 59 °C for 45 seconds and extension at 68 °C for 45 seconds followed by a cool down to 4 °C. The reaction was purified with 100 µl of AMPure XP beads to remove free biotin as described in Chapter 2, Section 2.14.1. Purified DNA was eluted in 31 µl of nuclease-free water.

4.3.5.2.(iii) Step 3– Bead Capture

For bead capture, a 50 µl aliquot of resuspended Streptavidin-coated MyOne beads (Invitrogen) was transferred to a clean 1.5 ml Eppendorf tube. A total of 200 µl of SureSelect Binding Buffer (Agilent) was added to the beads and mixed by pipetting. The tube was incubated on a magnetic stand at room temperature until the beads had pelleted. The supernatant was removed and discarded. This was repeated for a total of 3 washes. Beads were resuspended in 200 µl of SureSelect Binding Buffer and the 30 µl purified test reaction was added. The solution was mixed by pipetting up and down. The tube was incubated at room temperature, mixing vigorously (1,400 rpm) for 30 minutes.

The tube was centrifuged briefly, placed on a magnetic stand and beads allowed to pellet. The supernatant (230 µl) was harvested and stored in a clean Eppendorf. The supernatant was purified using 230 µl of AMPure XP beads with purification carried out as previously described (Chapter 2, Section 2.14.1). The purified DNA was eluted in 31 µl of nuclease-free water and stored in a clean 1.5 ml Eppendorf LoBind tube.

4.3.5.2.(iv) Step 4 – Long PCR Amplification

The concentrations of the test (depleted) sample and control (not depleted) sample were measured using the Qubit BR assay as described in Chapter 2, Section 2.6.2. A PCR amplification step was carried out on the test sample and the control sample as follows:-

DNA Long Amplification: Test Sample

Component	**Volume** µl **(X 1 reaction)**
Template DNA (70 ng)	30
PRM (ONT SQK-LSK108 kit)	20
2X LongAMP *Taq* (NEB)	40
Nuclease-free water	18

DNA Long Amplification: Control Sample

Component	**Volume** µl **(X 1 reaction)**
Template DNA (93.6 ng)	1
PRM (ONT SQK-LSK108 kit)	20
2X LongAMP *Taq* (NEB)	40
Nuclease-free water	47

The following cycling conditions were used for amplification:

Initial denatures at 95 °C for 3 minutes, 18 cycles of 98 °C for 20 seconds, 62 °C for 15 seconds and 65 °C for 3 minutes followed by a final extension of 65 °C for 3 minutes. Reactions were then cooled to 4 °C and then purified using 100 µl AMPure XP beads for each reaction as described previously (Chapter 2, Section 2.14.1). Purified DNA was eluted in 46 µl of nuclease-free water and stored in Eppendorf tubes and concentrations were measured using the Qubit BR assay as described in Chapter 2, Section 2.6.2. As total yield was greater than 500 ng for both reactions no additional PCR reactions were required as sufficient template had been generated for subsequent sequencing.

4.3.5.2.(v) Step 5 – End-repair, Sequencing Adapter Ligation and Sequencing

Post-PCR end-repair was carried out as described in Chapter 2, Section 2.14.5 and purified (Chapter 2, Section 2.14.1). Purified DNA was eluted in 31 µl of nuclease-free water. Sequencing adapter ligation and final library purification was carried out as described in Chapter 2, Section 2.14.5. Sequencing of the depletion test and control was carried out as described in Chapter 2, Section 2.14.6. First the test library was prepared, and the flow cell primed. A 1-hour experiment was initiated using ONT MinKNOW (Version 3.0.0) software on an ONT MinIT device (ONT-MinIT release Version 18.09.1). Once the run had completed, the flow cell was washed with 150 µl of Solution A (ONT Flow Cell wash kit) incubated for 10 minutes followed by 150 µl of Solution B (ONT Flow Cell wash kit). The flow cell was again primed as described in Chapter 2, Section 2.14.6 and 75 µl of the control (not depleted) library was added to the flow cell via the SpotON port. Again, a 1-hour experiment was initiated using the ONT MinKNOW software on an ONT Minit device.

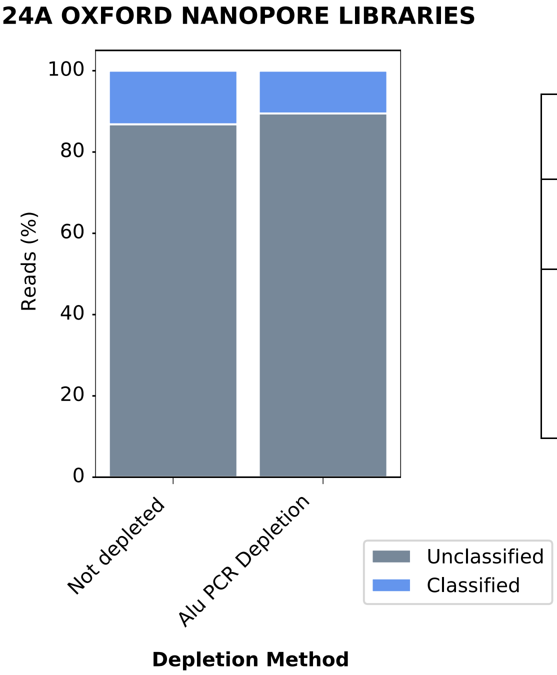
4.3.5.2.(vi) Step 6 - Data Analysis

Test and control data was basecalled with `Guppy` (Version 1.8.5) on the ONT Minit device using standard parameters and a q-score passing filter of 7. See Appendix Section 9.1.19 for the full code.

Only reads that passed a qscore of > 7 were brought forward for further analysis. Base called and quality filtered FASTQ files for each experiment were concatenated into a single file and taxonomic classification using `Centrifuge` (Version 1.0.3) (Kim *et al.,* 2016) was carried out as described in Chapter 2, Section 2.13.1 using the `-U` parameter for unpaired reads.

4.3.5.3 Results

The depletion of human DNA from an Oxford Nanopore metagenomic library was investigated for sample 24A using a single cycle *Alu* PCR with biotin. A not depleted 24A library was also sequenced. The proportion of classified (microbial) and unclassified reads indicated the success or failure of the depletion experiments when compared with the not-depleted control. There was no evidence of depletion from the Oxford Nanopore library for sample 24A (**Figure 4.36**) with the proportion of unclassified reads for the depletion test exceeding that of the not-depleted sample.



24A	Unclassified	Classified
Not depleted	86.81%	13.17%
Alu PCR depletion	89.48%	10.51%

Figure 4.36 Sample 24A Oxford Nanopore Libraries: Pre- and post-capture classified and unclassified reads.

4.4 Discussion

This aim of this chapter was to investigate the development of methods for the depletion of human DNA by targeting repetitive elements in the human genome. This was carried out on mock samples, real samples and sequencing libraries using biotinylated Cot-1 DNA probes, biotinylated *Alu* DNA probes, biotinylated *Alu* RNA probes and a single cycle *Alu* PCR to incorporate biotin. While there was some qualitative evidence of human DNA depletion on the agarose gels after *Alu* RNA hybridisation with digested genomic DNA and capture, there was no evidence of depletion or minor depletion when samples were then sequenced.

Cot-1 DNA is commonly used as a human DNA repeat suppressor in filter, microarray and *in-situ* hybridisation reactions. It was hypothesised in the current study that biotinylated Cot-1 DNA would hybridise with repeat elements of human DNA in a mock sample resulting in capture of human DNA fragments with the intact microbial component only remaining in the post-capture supernatant.

Firstly, fragments less than 100 base pairs were removed from the Cot-1 in order to reduce non-specific binding to the microbial component (the target for sequencing) of the sample from which the human host DNA was being depleted. Biotinylation was carried out using a biotin-16-dUTP as a substrate to replace 33 % of dTTPs. A 16-atom linker arm was chosen as the association between biotin and streptavidin improves with an increase in arm length even though short linker arms are better DNA substrates. After hybridisation, bead capture and quantification of human DNA in test experiments, there was little evidence to suggest that human DNA was being removed from the samples by the biotinylated Cot-1 probes.

Similarly, in the case of *Alu* DNA probes, there was limited evidence of human DNA depletion. The hypothesis here is that DNA probes rapidly renatured after denaturation, reducing the availability of probes for hybridisation with the test DNA. Additionally, a proportion of *Alu* DNA probes remained in the supernatant post-capture indicating unsuccessful biotinylation of all probe material.

The use of *Alu* RNA probes provided greater evidence that a small amount of human DNA was being depleted from the mock and real samples. Single stranded RNA probes are not complementary, being transcribed from one strand only, and therefore will not preferentially hybridise to each other. In addition, RNA: DNA hybrids are more stable than DNA: DNA. This stability is owed to the minor groove of DNA: RNA hybrids which demonstrate more kinetically significant hydration than the DNA: DNA hybrids. This can be attributed to the hydroxyl groups in RNA which demonstrate a hydrophilic lining (Lesnik *et al.,* 1995, Gyi *et al.,* 1998). Upon sequencing test and control samples, whilst approximately a further 10 % of reads were classifiable after the first depletion for samples 24A and 24Be, it is likely that the genomic material began to renature with itself rather than biotinylated RNA probes during the second and third hybridisation reactions. In the case of sample S2, approximately 3 % more reads were classifiable after the first depletion however no further reads were classifiable from the mock sample after the first depletion when compared to the non-depleted mock sample.

Interestingly the hybridisation control for the mock community showed a greater proportion of classifiable reads compared to the non-depleted mock and the sequentially depleted samples. It is uncertain here if this is as a result of a technical issue during the hybridisation procedure or the re-association dynamics of DNA in the absence of a probe. As well as the possibility that the probe did not bind adequately to the test genomic DNA, it is hypothesised that when hybridisation between the probe and test DNA did occur that steric hindrance reduced the ability of biotin to bind to the streptavidin molecules. Binding efficiency may decrease with longer DNA fragments. Owing to this factor, the biotinylated Cot-1 DNA, *Alu* RNA and a single cycle *Alu* PCR with biotin were tested on metagenomic libraries for the mock community and samples 24A and 24Be. There was no evidence of depletion in any test. In the case of the Cot-1 DNA this again may be again due to rapidly reannealing elements. In the case of *Alu* RNA hybridisation it may be due to their spacing in the genome as they are reported to occur approximately every 3 kb on average and the Illumina library fragments were 300 base pairs or less in length.

Sample 24A was sheared to fragment sizes of between 1.5 and 2 kb and Oxford Nanopore libraries were prepared to test the depletion of human DNA using the single cycle *Alu* PCR with biotin. The hypothesis here was that the fragment lengths were short enough so that steric hindrance was less likely to occur during bead capture however were of a sufficient length to allow primers to anneal to full or partial *Alu* elements. In addition, the *Alu* PCR would incorporate biotin into both forward and reverse strands under optimised PCR conditions. There was again lack of evidence for human DNA depletion in this case.

Due to the failure or insufficient depletion of human DNA from mock and real samples using a large number of strategies as detailed above, a targeted capture approach for the enrichment of *Legionella pneumophila* genomes was investigated in the next chapter.

Chapter 5.

A Pilot Study for the Direct Targeted Capture of *Legionella pneumophila* Genomes

5.1 Introduction

The rapid identification of pathogens is essential for the clinical management of patients and source tracking. Whilst first-line diagnostics provide an indication of pathogen presence or absence, the "gold standard" for diagnosis of bacterial infections is culture. In the case of *L. pneumophila*, epidemiological typing, based on a traditional 7-loci scheme (Gaia *et al.,* 2005, Ratzow *et al.,* 2007, Mentasti *et al.,* 2014) is carried out on an isolate or directly from the nucleic acid extract in situations where an isolate does not grow (Coscollá *et al.,* 2009). This is essential for the investigation of clusters and outbreaks and the linkage of epidemiologically related cases. The traditional approach however is currently not proving to be discriminatory enough as a few common sequence types are being identified as being the cause of the majority of *Legionella* cases (Borchardt *et al.,* 2008, Harrison *et al.,* 2009, Tijet *et al.,* 2010, David *et al.,* 2016[b]). Additionally, it is often not possible to obtain allele numbers by direct nested SBT from culture negative cases and assessment of diversity from cultured isolates is biased towards individual colony picks. Next generation sequencing of whole bacterial genomes isolated by culture is now being implemented as a routine methodology in some public health reference laboratories (PHE, 2018). Bacterial whole genomes provide high resolution data over traditional gene typing methods. Due to the proliferation of whole genome sequencing technologies and similarity of costs with sequence-based typing (SBT), a number of whole/core genome MLST schemes have been proposed for *L. pneumophila* (Moran-Gilad *et al.,* 2015, Qin *et al.,* 2016). However, bacterial isolation is not always achievable for a number of reasons, as reviewed in Chapter 1, Section 1.7.3. Ideally, direct metagenomic sequencing from a sample of interest would allow the rapid sequencing of *L. pneumophila* genomes however the pathogen-to-other nucleic acid content is usually extremely low, as addressed in previous studies (Doughty *et al.,* 2014, Pendleton *et al.,* 2017) and data generated in Chapter 3 of this thesis. To overcome these limitations, the use of a targeted hybridisation-based capture approach may allow *L. pneumophila* whole genome reconstruction directly from diagnostic positive clinical or environmental samples.

Agilent Technologies developed, originally for the capture of whole human exomes (Gnirke *et al.,* 2009), an in-solution target capture system (SureSelect™) based on the hybridisation of genomic fragments to biotinylated probes. The system requires the design and synthesis of biotinylated RNA "baits", 120-mer in size that are complementary to the genomic regions of interest. Baits are designed in a tiling fashion to ensure a consistent bait density per genomic region. After nucleic acid extraction from samples of interest, libraries are prepared, indexed and hybridised with the RNA baits. After capture, unbound fragments are removed by a washing step. PCR amplification is carried out to enrich the captured fragments and enriched libraries are then ready to be sequenced on a Next-Generation Sequencing platform.

The Agilent SureSelect™ approach has been successfully applied in previous studies for the capture of difficult-to-grow, slow-growing and non-cultivable bacterial species as summarised in **Table 5.1**. These prior studies demonstrated that the approach significantly reduced time-to-results for antibiotic resistance detection, allowed the application of typing systems to strains from difficult to culture and uncultivable samples and exposed a level of previously unrecognised strain heterogeneity in some cases.

In the case of *L. pneumophila,* the slow growth or lack of growth of the bacteria may hinder the prompt identification of a sequence type. The primary aim of the current chapter therefore was to conduct a pilot study investigating the use of the Agilent SureSelect™ target capture approach for the enrichment of *L. pneumophila* whole genomes directly from clinical and environmental samples containing known and unknown *L. pneumophila* sequence types and other *Legionella* species.

Table 5.1 Studies that have used Agilent SureSelect Targeted Capture to Enrich and Sequence Bacterial Pathogens.

Bacterial Pathogen	Clinical Manifestation of Infection	Sample Types Tested	Probe Design	Why is targeted capture useful here?	Reference
Chlamydia trachomatis	Trachoma	Vaginal & Urinal: 9 isolates and 10 diagnostic positive samples	74 completed reference genomes	Low numbers of pathogens in clinical samples Difficult to grow (obligate intracellular pathogen)	Christiansen *et al.,* 2014
Neisseria meningitidis	Invasive meningococcal disease	Blood & CSF: 10 diagnostic positive samples and 10 matching isolates	77 complete reference genomes and 2,898 drafts	Difficult to grow (early antibiotic administration)	Clark *et al.*, 2017
Mycobacterium tuberculosis	Tuberculosis	Sputum: 24 smear-positive specimens and 24 matching isolates, 10 smear-positive samples that failed to grow culture	1 complete reference genome	Long culture step Extended *in vitro* culturing can introduce genome mutations	Brown *et al.*, 2015
Mycobacterium tuberculosis	Tuberculosis	Sputum: 1 smear-positive sample	1 complete reference genome	Long culture step	Nimmo *et al.*, 2017
Chlamydia trachomatis	Trachoma	Conjunctival swabs: 118 diagnostic positive specimens and 8 isolates	1 complete reference genome	Difficult to grow (obligate intracellular pathogen)	Last *et al.*, 2018
Mycobacterium tuberculosis	Tuberculosis	Sputum: 43 diagnostic positive specimens and 43 matched isolates	1 complete reference genome - Reduced bait set for resistance genes	Long culture step Extended *in vitro* culturing can introduce genome mutations	Doyle *et al.*, 2018
Mycobacterium tuberculosis	Tuberculosis	Sputum: 39 diagnostic positive specimens and 39 matched isolates	1 complete reference genome	Long culture step Viable but non-culturable bacteria Extended *in-vitro* culturing can introduce genome mutations	Nimmo *et al.*, 2018
Treponema pallidum subsp. pallidum	Syphilis	Oropharyngeal, Tongue, Vaginal, Penile, Scrotal CSF, Placental, Blood: 35 diagnostic positive specimens	1 complete reference genome (*in vivo* culture)	Unculturable *in vitro*	Pinto *et al.*, 2016
Haemophilus ducreyi	Chancroid	Cutaneous lesions: 72 diagnostic positive and negative specimens	1 complete reference genome	Difficult to culture	Marks *et al.*, 2018
Chlamydia pecorum	Cattle: sporadic bovine encephalomyelitis Sheep: polyarthritis and conjunctivitis Koala: ocular and urogenital tract diseases	Urogenital Tract, Ocular, Rectal, Joint, Brain: 9 diagnostic positive specimens and 1 isolate	1 complete reference genome	Difficult to culture Extended *in vitro* culturing can introduce genome mutations	Bachmann *et al.*, 2015
Mycobacterium leprae	Hanseniasis (leprosy)	Skin Biopsy: 10 diagnostic positive specimens and 1 reference control	Not specified	Unculturable *in vitro*	Lavania *et al.*, 2018

5.2 Aims and Objectives

1. *L. pneumophila* capture and sequencing:
    a. Design RNA baits based on a database of completed *L. pneumophila* genomes.
    b. Perform targeted capture and sequencing from known *Legionella* positive clinical and environmental specimens, a dilution series with known copy numbers of *L. pneumophila* to determine analytical cut-offs and a mock community containing two *L. pneumophila* sequence types and one other *Legionella* species.

2. Evaluation of the target capture approach:
    a. Determine the number of sequenced reads before and after quality control and *in silico* human DNA read removal.
    b. Examine the sequence coverage across target regions.
    c. Determine the proportion of reads that map to the intended target.
    d. Examine the mean depth of coverage across target regions.
    e. Investigate the taxonomic composition of on- and off-target reads.
    f. Typeability:
        i. Determine if a *L. pneumophila* sequence type (ST) based on the traditional 7-loci scheme can be assigned to captured reads.
        ii. Investigate the typeability of 50 core *L. pneumophila* genes in captured samples based on the newly proposed MLST scheme.

3. Mixed *L. pneumophila* Sequence Type Analysis
    a. Determine if samples contain a mixture of *L. pneumophila* sequence types or if a single strain or closely-related strains can be predicted from capture data.

4. *L. pneumophila* Genome Assembly
    a. Assemble draft *L. pneumophila* genomes.
    b. Investigate genome completeness and levels of contamination.
    c. Carry out a mixed *Legionella* species analysis based on conserved single copy marker genes.

5.3 Methods

5.3.1 Clinical and Environmental Specimens

A total of 10 sputum specimens from 10 individuals and 9 water specimens from 9 environmental sources received at Public Health England (PHE) were obtained for the pilot study. Details regarding the clinical specimens are outlined in **Table 5.2** and the environmental specimens in **Table 5.3**. Ethical approval was granted from the Research Ethics Committee (REC) as described in Chapter 2, Section 2.3 for the sequencing and analysis of the clinical specimens. Sputum samples were previously confirmed positive by the RVPBRU team at PHE for *L. pneumophila* by urinary antigen testing, isolation, qPCR or a combination of these techniques. Water samples were previously confirmed positive by the Food, Water and Environmental group at PHE for the presence of *L. pneumophila* and/or a *Legionella* species by culture and qPCR. *Legionella* abundance was determined semi-quantitatively by an in-house 16S rRNA gene sequencing pipeline (Chapter 2, Section 2.15) and assigned a low or high abundance status. As clinical and environmental samples were sourced from different labs, methods of detection implemented do differ therefore influencing the available data for each group.

Table 5.2 Clinical Specimens for Target Capture Pilot Study.

Sample ID	Specimen Type	Date Received PHE	Lp Urinary Antigen*	Lp Culture*	Lp qPCR (*mip*)*	Serogroup*	Sequence Type*	*Legionella* abundance (16S rRNA gene Analysis)
H1	Sputum	08/09/2016	Positive	Isolated	Lp SG1 detected	1	ST47	Low
H2	Sputum	21/10/2016	Positive	Isolated	Lp SG1 detected	1	ST 37	Low
H3	Sputum	05/01/2016	Positive	Isolated	Lp SG1 detected	1	ST 1694	Low
H4	Sputum	19/09/2016	Positive	Isolated	Lp (non-SG1) detected	6	ST 81	Low
H5	Sputum	23/08/2016	Positive	Not Isolated	Lp SG1 detected	1	Direct SBT yielded partial allele profile (2,0,0,10,9,4,28) consistent with ST 616	Low
H6	Sputum	16/09/2016	Positive	Isolated	Lp SG1 detected	1	ST2287 (6,10,14,10,21,3,9)	High
H7	Sputum	12/08/2016	Positive	Not Isolated	Lp SG1 detected	1	Not Tested	Low
H8	Sputum	11/10/2016	Not Tested	Isolated	Lp SG1 detected	1	ST 445	High
H9	Sputum	10/10/2016	Positive	Isolated	Lp SG1 detected	1	ST 1 from another sample (same patient)	High
H10	Sputum	20/10/2016	Positive	Isolated	Lp SG1 detected	1	ST 616	Low

Lp = *Legionella pneumophila,* SG1 = serogroup 1, ST = sequence type, *mip* = macrophage infectivity potentiator gene, *carried out by RVPBRU team at PHE

Table 5.3 Environmental (Water) Specimens for Target Capture Pilot Study.

Sample ID	Specimen Type	Culture result (CFU/L)*	*Legionella* species identified by qPCR*
E1	Shower	>3,000	*Legionella* spp. & LpSG1
E2	Pool	700	Lp SG1
E3	Alethium Spa Side	280	*Legionella* spp.
E4	Cold Tap	380	*Legionella* spp.
E5	Hot water system	>3,000	Lp SG1
E6	Hot water system	>2,000	Lp SG1
E7	Cold water system	>3,000	Lp SG1
E8	Cold water system	>3,000	Lp SG1
E9	Cold water system	>3,000	Lp SG1

CFU = colony forming units, Lp = *Legionella pneumophila*, SG1 = serogroup 1, spp = species, *carried out by FW&E team at PHE

5.3.2 Phenol-Chloroform Extraction

Clinical and environmental samples were processed and transferred to lysis matrix tubes containing lysis buffer. Samples were then bead beaten and nucleic acid was extracted by the Phenol-Chloroform method. Nucleic acid was precipitated and purified. All processing, extraction and purification steps are described in Chapter 2, Section 2.5. DNA concentration was measured by PicoGreen assay as described in Chapter 2, Section 2.6.1.

5.3.3 Preparation of Dilution Series and Mock Community

The following genomic material was used in the preparation of the dilution series and mock community: human genomic DNA, *L. pneumophila* Philadelphia-1, *L. pneumophila* France 5811, *L. longbeachae*, *S. pneumoniae*, *H. influenzae* and *V. dispar*. The strain designations, sources and ethical considerations regarding the genomic material are detailed in Chapter 2, Section 2.4. The dilution series contained human genomic DNA spiked with a defined copy number of *L. pneumophila* Philadelphia-1 as outlined in **Table 5.4**. The composition of the mock community is described in **Table 5.5**.

Table 5.4 Composition of the Dilution Series.

Dilution ID	Composition	*L. pneumophila* Copy Number
D1	Human genomic DNA *L. pneumophila* Philadelphia-1 (ST36)	1 X 10^6
D2		1 X 10^5
D3		1 X 10^4
D4		1 X 10^3

Table 5.5 Composition of the Mock Community.

Mock Community Components	Composition (%)
Human Genomic DNA	90
L. pneumophila Phil-1 (ST36)	2
L. pneumophila OLDA (ST1)	2
L. longbeachae	2
S. pneumoniae	2
H. influenzae	2
V. dispar	2

5.3.4 Target Capture for *Legionella pneumophila*

Database preparation (see Chapter 2, Section 2.11.1), bait design (Chapter 2, Section 2.11.2), library preparation and hybridisation capture (Chapter 2, Section 2.11.3) and sequencing (Chapter 2, Section 2.11.4) were carried out. Post-capture libraries were sequenced twice, and data from the two runs combined.

5.3.5 Target Capture Data Analysis

5.3.5.1 Data Cleaning and Quality Control

Sequenced reads were demultiplexed into individual libraries as described in Chapter 2, Section 2.12.1.

All data cleaning and quality control steps (adapter trimming for removal of the Illumina Universal adapter sequence, quality filtering, PhiX removal and human DNA removal) were carried out as described from Section 2.12.2 to Section 2.12.5, Chapter 2. Read numbers before and after quality control and after human DNA removal for pre- and post-targeted capture were investigated.

5.3.5.2 Sequence Alignment with *Legionella* Reference Genomes

Sequenced reads from the captured data were aligned against a completed reference sequence of the same sequence type. If no known ST-specific reference genome was available, samples were analysed against a database of completed *Legionella* genomes using `KmerID` (Version 0.1) (Schaefer, 2014) for identification of the closest related genome (see Appendix Section 9.5 for full list of genomes). Sequence alignment was carried out using `Bowtie2` (Version 2.3.2) (Langmead *et al.,* 2012) with default sensitivity parameters and the `--no-unal` option to suppress SAM records for reads that failed to align. The SAM file was converted to BAM and sorted using `picard` (Version 2.12.1) (Picard Toolkit, 2019). Duplicates were marked and removed using `picard` (Version 2.12.1) (Picard Toolkit, 2019). Metrics for mean depth of coverage, percentage of reads mapping to the reference sequence and total percentage of reference genome covered were generated for samples before and after duplicate removal using the `pileup` script from `BBTools` (Version 37.38) (Bushnell, 2014). See Appendix Section 9.1.12 for alignment analysis code.

5.3.5.3 *In silico* Sequence-Based Typing for *L. pneumophila*

In silico sequence-based typing (SBT) analysis based on the traditional ESGLI *L. pneumophila* scheme (Gaia *et al.,* 2005, Ratzow *et al.,* 2007, Mentasti *et al.,* 2014) was carried out using the ESGLI database (http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php) and `SRST2` (Inouye *et al.,* 2014) on the captured data to determine if a partial or full sequence type could be generated. The analysis method is described in Chapter 2, Section 2.13.2.

5.3.5.4 Identification of 50-Core Genes for MLST

An investigation of gene presence/absence of 50 core genes pertinent to the multi-locus sequence-based typing scheme as defined by David *et al.,* 2016(b) was carried out. Captured data was aligned to the coding sequences of the *L. pneumophila* Philadelphia-1 reference genome. The coding sequence file for *L. pneumophila* Philadelphia-1 (https://www.ncbi.nlm.nih.gov/genome/416?genome_assembly_id=300116) was used as a gene database. Sample reads were mapped to the gene database using `SRST2` (Version 0.2.0) (Inouye *et al.,* 2014). The presence or absence of the 50 genes was reported. See Appendix Section 9.1.13 for the full code.

5.3.5.5 Taxonomic Classification

Taxonomic classification of captured data was carried out using `Centrifuge` (Version 1.0.3) (Kim *et al.,* 2016) and species assignments were extracted from the output as described in Chapter 2, Section 2.13.1.

5.3.6 Analysis for Mixed *L. pneumophila* Sequence Types: Strain Estimation Analysis from Reads

Strain-level analysis using the `StrainEst` program (Version 1.2.2) (Albanese *et al.,* 2017) was carried out to determine if a *L. pneumophila* strain could be predicted from capture data based on core genome single nucleotide polymorphisms (SNPs) by alignment to a database of *L. pneumophila* genomes (see Appendix Section 9.3 for the full list of genomes). Preparation of database, SNP matrix and strain estimation methods are described in Chapter 2, Section 2.13.3. In cases where a strain was not identified at an identity threshold of 99 %, the identity threshold was decreased incrementally (98 %, 97

%, 96 %) until a sequence type was predicted. If a strain was not predicted by a 95 % identity threshold, the strain was reported as undetermined. The minimum depth of coverage threshold was also adjusted based on reported maximum/minimum SNP depth.

5.3.7 Genome Assembly and Analysis

Metagenome assembly was carried out using `metaSPAdes` (Version 3.10.1) (Nurk *et al.,* 2017) without error correction (to avoid losing information from potentially closely related strains) and integrating assemblies spanning from a k-mer size of 27 to 127 base pairs. See Appendix Section 9.1.14 for the full code.

Taxonomic classification of assemblies was carried out using `Centrifuge` (Version 1.0.3) (Kim *et al.,* 2016) as described in Chapter 2, Section 2.13.1. Assemblies were decontaminated by extracting contigs belonging to the order Legionellalaes only using a custom shell script (see Appendix Section 9.1.15). Decontaminated *de novo* assemblies were investigated using `CheckM` (Version 1.0.8) (Parks *et al.,* 2015) with default parameters for genome completeness and evidence of further contamination.

5.3.8 Mixed *Legionella* Species Analysis

During `CheckM` (Version 1.0.8) (Parks *et al.,* 2015) strain heterogeneity analysis on decontaminated *de novo* assemblies, a file containing amino acid sequence alignments of single copy marker genes present more than once (with >= 90 % amino acid identity) in each sample assembly was generated. A species (or closest related species) was assigned to each amino acid sequence using the `phmmer` algorithm from `HMMER` (Version 3.1b2) (Finn *et al.,* 2011) and the UniProtKB database (The UniProt Consortium, 2019). Sequences were then manually inspected to ensure multiple copy reporting was not a result of errors during assembly. Sequence alignments presenting as such were eliminated from the analysis.

Libraries of conserved single-copy marker genes reported as multi-copy in the samples were downloaded for the Legionellales order from the Pfam database (El-Gelbali *et al.,* 2019) (**Table 5.6**). Sequences for Berkiella, Rickettsiella and Coxiella species other *Coxiella burnetii* (which was used as an outgroup during phylogenetic analysis) were removed from the Legionellales Pfam files. *Legionella drozanskii* and *Legionella anisa* were not present in the Pfam database therefore their sequences, if present, for the protein families outlined in **Table 5.6** were retrieved from the RefSeq non-redundant protein database (https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/).

For each sample, multi-copy amino acid sequences were aligned with respective Pfam marker gene files (**Table 5.6**) and trimmed using MEGA7 (Kumar *et al.,* 2016). Some sequences were eliminated at this point due to insufficient coverage. Then, partial amino acid sequences for each sample were concatenated together based on HMMER assignment.

Table 5.6 Protein Families and Pfam Accession Numbers.

Protein Family	Pfam Accession Number
Ribosomal Protein L10	PF00466
Ribosomal Protein L2 C-terminal domain	PF03947
Ribosomal Protein L21	PF00829
Ribosomal Protein L27	PF01016
GTP1/OBG	PF01018
NADH dehydrogenase	PF00507
Ribosomal Protein L7/L12	PF0542
Ribosomal Protein S8	PF00410
Ribosomal Protein L25	PF01386
RimP N-terminal domain	PF02576
NusA N-terminal domain	PF08529
SecG	PF03840
Ribosomal Protein S19	PF00203
Ribosomal Protein L22/L17	PF00237

Phylogenetic analysis was carried out on the concatenated partial protein sequences for each sample using RaxML (Version 8.0.0) (Stamatakis *et al.,* 2014). A maximum-likelihood search was performed with 1,000 bootstrap inferences. The best scoring maximum-likelihood tree with bootstrap support values was written to file. See Appendix Section 9.1.16 for full code.

Due to the fragmented nature of the assemblies and incompleteness of genomic data, as well as the shallow depth of sequencing, not all species identified within samples E7 and E8 (3 species) had the same marker genes identified (e.g. Species1 and Species2 [but not Species3] each had copies of PF0466 or Species 2 and Species3 [but not Species 1] each had copies of PF03947). In these cases, two phylogenetic trees were generated. The trees were visualised using FigTree (Version 1.4.4) (Rambaut, 2008) and exported in SVG format. The SVG file was imported into the Gravit Design editor (Version 2019-1.5) (https://www.designer.io/) for annotation.

5.4 Results

5.4.1 Evaluation of the Target Capture Approach

5.4.1.1 Sequenced Reads and Proportion of Reads Mapped to Human Reference Genome

Read pair numbers before and after quality control and proportions of reads mapping to the human reference genome for the post-capture libraries are detailed in **Table 5.7**. The total number of read pairs sequenced ranged from 248,779 to 5,589,723. After quality control, the proportion of reads aligning to the human reference genome for the dilutions represented 1 % for D1 (D1 composition Human Genomic DNA and 10^6 *L. pneumophila* copies), 1 % for D2 (Human Genomic DNA plus 10^5 *Legionella* copies), 5 % for D3 (Human Genomic DNA plus 10^4 *Legionella* copies) and 92 % for D4 (Human Genomic DNA plus 10^3 *Legionella* copies). For the Mock sample (for composition see **Table 5.5**) post-capture 0.6 % of the sequence reads were found to map to the human reference genome. For the clinical samples (**Table 5.2**) the proportion of reads mapping to the human reference varied from 0.5 % to 97.3 %. As a control, environmental samples (**Table 5.3**) also underwent screening for human reads and the proportion of reads mapping to the human reference varied from 0.21 % to 0.34 %.

Table 5.7 Reads Before and After Sequencing QC and Reads Post QC Mapping to the Human Reference Genome.

	Sample	**Read Pairs Before QC**	**Read Pairs After QC**	**Read Pairs Mapped to Human Reference (%)**
Dilution Series	**D1**	4,100,342	3,818,512	38,971 (1 %)
	D2	2,216,876	2,056,091	22,306 (1 %)
	D3	2,366,006	2,201,835	2,094,757 (5 %)
	D4	2,414,384	2,243,825	2,015,873 (92 %)
Mock	**Mock**	5,589,723	5,198,683	32,910 (0.6 %)
Clinical	**H1**	621,002	549,562	534,945 (97.3 %)
	H2	2,282,283	2,080,769	9,610 (0.5 %)
	H3	1,566,010	1,397,353	854,519 (61.2 %)
	H4	248,779	224,399	90,803 (40.5 %)
	H5	1,609,295	1,394,421	1,304,263 (93.5 %)
	H6	807,079	748,172	12,418 (1.7 %)
	H7	866,070	768,607	760,826 (99 %)
	H8	1,630,845	1,491,522	475,043 (31.8 %)
	H9	2,694,977	2,483,287	136,537 (5.5 %)
	H10	2,124,350	1,888,941	808,634 (42.8 %)
Environmental	**E1**	2,478,751	2,215,622	6,302 (0.3 %)
	E2	1,684,008	1,518,777	3,882 (0.26 %)
	E3	1,883,598	1,518,777	3,882 (0.26 %)
	E4	1,330,629	1,192,361	2,884 (0.24 %)
	E5	2,228,002	1,968,387	4,186 (0.21 %)
	E6	1,941,943	1,761,015	5,901 (0.34 %)
	E7	2,400,882	2,155,014	6,079 (0.28 %)
	E8	2,565,409	2,300,344	4,976 (0.22 %)
	E9	3,465,576	3,210,174	9,101 (0.28 %)

5.4.1.2 Genome Coverage, Proportion of On-Target Reads and Mean Depth of Coverage

Reads from post-capture libraries were mapped to a closely related complete *Legionella* reference genome. For the dilution series and mocks this was *L. pneumophila* as this was the strain used to prepare the test samples (**Tables 5.4** and **5.5**). The percentage of reference bases covered, the percentage of on-target reads and the mean depth of coverage before and after read duplicate removal were investigated.

For the dilution series samples, the proportion of the reference genome covered varied from 99.9 % for D1 to 94.5 % for D4. With regard to on-target reads, 99 % of captured reads aligned to the *Legionella* reference sequence. For D3 (10^4 *Legionella* copies), 95 % of reads aligned to the reference and D4 (10^3 *Legionella* copies), 8 % of reads aligned to the reference sequence. The mean depth of coverage varied from 370 times for D1 to 10 times for D4, after duplicate removal. Read duplication levels varied from 1 % for D1 to 45 % for D4. The mock sample reads covered 99.9 % of the reference genome, 87.5 % of reads mapped to the *Legionella* reference genome with a mean depth of coverage of 447 times, after duplicate removal. Read duplication levels were 1.3 % (**Table 5.8**). Based on the mock and dilution series this highlights capture is robust to the level of 10^4 copies total input of *L. pneumophila.*

Table 5.8 Target Capture Statistics: Dilutions and Mock Community.

Sample	Closely related *Legionella* Reference Genome*	Reference Genome covered (%)	Reads mapped to Reference (Number & %)*	Mean Depth of Coverage (SD)	Reads Remaining After Duplicate Removal	Mean Depth of Coverage After Duplicate Removal (SD)*	Read Duplication Levels (%)
D1	Philadelphia1	99.9	3,757,304 (+) 3,757,425 (-) (99 %)	376 (667.26)	3,396,399 (+) 3,688,496 (-)	370 (589.57)	1
D2	Philadelphia1	99.9	2,021,209 (+) 2,021,334 (-) (99 %)	198 (403.88)	1,966,858 (+) 1,966,853 (-)	193 (360.28)	2
D3	Philadelphia1	99.9	2,081,567 (+) 2,081,775 (-) (95 %)	205 (500.74)	1,582,582 (+) 1,582,662 (-)	158 (270.27)	24
D4	Philadelphia1	94.2	181,613 (+) 181,634 (-) (8 %)	18 (43.17)	100,677 (+) 100,682 (-)	10 (26.6)	45
Mock	Philadelphia1	99.9	4,552,703 (+) 4,553,186 (-) (87.6 %)	452 (539.28)	4,492,958 (+) 4,492,833 (-)	447 (495.55)	1.3

* (+) refers to the forward strand and (-) refers to the reverse strand

For clinical samples, the proportion of the *Legionella* reference genome covered varied from 3.4 % to 95.7 %. The proportion of reads mapping to the *Legionella* reference genome varied from 0.3 % to 87 %. Mean depth of coverage varied from 0.11 times to 164.47 times, after duplicate removal. Read duplication levels varied from 0.97 % to 57.7 % (**Table 5.9**).

Table 5.9 Target Capture Statistics: Clinical Samples

Sample	Closely related *Legionella* Reference*	Reference Genome covered (%)	Reads mapped to reference (Number & %)	Mean Depth of Coverage (SD)	Reads Remaining After Duplicate Removal	Mean Depth of Coverage After Duplicate Removal (SD)	Read Duplication Levels (%)
H1	Lorraine (NC_018139)	9.6	8,478 (+) 8,507 (-) (1.5 %)	0.71 (6.2)	5,208 (+) 5,239 (-)	0.45 (4.4)	38
H2	ST37 (NZ_LT632616)	5.3	1,091,691 (+) 1,087,410 (-) (52 %)	91.97 (1964.4)	575,061 (+) 575,079 (-)	51.66 (1314.8)	47
H3	ST42 (NZ_LT632617)	84.6	320,191 (+) 319,741 (-) (23 %)	26.86 (34.7)	168,295 (+) 167,822 (-)	14.53 (20.2)	47
H4	Leiden-1 (ERS1080593)	27.7	28,221 (+) 28,336 (-) (13 %)	2.27 (20.9)	15,008 (+) 15,000 (-)	1.23 (10.9)	47
H5	Leiden-1 (ERS1080593)	32.6	38,571 (+) 38,442 (-) (3 %)	2.69 (9.6)	21,785 (+) 21,736 (-)	1.54 (6.2)	43
H6	Corby (NC_009494)	89.8	647,855 (+) 647,763 (-) (87 %)	61.65 (47.6)	641,536 (+) 641,488 (-)	61.09 (47)	0.97
H7	FFI329 (NZ_CP016874)	3.4	2,135 (+) 2,130 (-) (0.3 %)	0.17 (1.2)	1,291 (+) 1,286 (-)	0.11 (0.7)	39.5
H8	ST23 (NZ_LT632615)	92.8	831,624 (+) 829,431 (-) (56 %)	82.67 (109.3)	425,693 (+) 424,829 (-)	43.82 (53.8)	48.7
H9	OLDA (NZ_CP016030)	95.7	2,150,120 (+) 2,150,015 (-) (86.5 %)	178.48 (751.9)	1,961,655 (+) 1,961,048 (-)	164.47 (604.1)	8.7
H10	Leiden-1 (ERS1080593)	19.9	164,530 (+) 164,640 (-) (8.7 %)	10.37 (356.7)	69,625 (+) 69,586 (-)	4.69 (153.8)	57.7

For environmental samples, the proportion of the reference genome covered varied from 3.4 % to 95.7 %. The proportion of reads mapping to the *Legionella* reference genome varied from 14 % to 88.7 %. Mean depth of coverage varied from 5.4 times to 232 times, after duplicate removal and read duplication levels varied from 12.9 % to 57.7 %. (**Table 5.10**).

Table 5.10 Target Capture Statistics: Environmental Samples.

Sample	Closely related *Legionella* Reference*	Reference Genome covered (%)	Reads mapped to reference (Number & %)	Mean Depth of Coverage (SD)	Reads Remaining After Duplicate Removal	Mean Depth of Coverage After Duplicate Removal (SD)	Read Duplication Levels (%)
E1	*L. pneumophila* OLDA (NZ_CP016030)	72.7	897,987 (+) 894,436 (-) (40.5 %)	69.1 (1286.7)	482,556 (+) 482,539 (-)	38.42 (699)	46
E2	*L. pneumophila* OLDA (NZ_CP016030)	76.3	358,572 (+) 358,944 (-) (24 %)	32.95 (360.2)	151,879 (+) 151,796 (-)	14.29 (168.6)	57.7
E3	*L_taurinensis* (ERS1324129L)	0.45	523,532 (+) 530,401 (-) (31 %)	50.24 (1273.7)	302,800 (+) 302,817 (-)	29.79 (842.2)	42.5
E4	*L_anisa* (NBTX01000001)	2.2	162,239 (+) 164,652 (-) (14 %)	10.69 (370.6)	79,884 (+) 79,862 (-)	5.4 (187.1)	51
E5	*L. pneumophila* Corby (NC_009494)	91.9	461,427 (+) 462,335 (-) (23.4 %)	36.13 (371.4)	300,274 (+) 300,288 (-)	23.93 (228.9)	35
E6	*L. pneumophila* Corby (NC_009494)	95.9	638,976 (+) 640,406 (-) (36 %)	56.77 (301.5)	455,656 (+) 455,603 (-)	41.1 (207.4)	28.7
E7	*L. pneumophila* Dallas-1E (subsp.fraseri) (NZ_CP017458)	54.9	808,466 (+) 809,098 (-) (37.5 %)	67.25 (971.6)	426,207 (+) 426,327 (-)	36.1 (521.3)	47.3
E8	*L. pneumophila* Dallas-1E (subsp.fraseri) (NZ_CP017458)	95	839,226 (+) 841,186 (-) (36.5 %)	72.13 (884.4)	551,066 (+) 551,051 (-)	49.24 (531.4)	34
E9	*L. pneumophila* OLDA (NZ_CP016030)	98.6	2,847,124 (+) 2,847,058 (-) (88.7 %)	265.46 (510.9)	2,479,917 (+) 2,479,781 (-)	232.08 (371)	12.9

5.4.1.3 Taxonomic Classification of On- and Off-Target Reads

Taxonomic classification was carried out to investigate the overall composition of captured reads and to consolidate the proportions of captured *L. pneumophila*, other *Legionella* species, other bacterial species and unclassified captured reads with human DNA reads. **Figure 5.1** shows the proportions of classified captured reads for the dilutions and mock sample. For the mock sample, 97.72 % of reads were classified as belonging to the genus *Legionella*. Of this, 96.24 % of reads were classified as *L. pneumophila* and 1.48 % were classified as belonging to other *Legionella* species. Since the *L. longbeachae* species was included in this sample, it was found that 23,547 reads were classified uniquely as belonging to *L. longbeachae*, indicating capture of the species (see Appendix Section 9.6, **Figure 5**). Dilution samples D1 to D4 contained *L. pneumophila* only and few reads were classified as other *Legionella* species due, most likely, to minor misclassification by the classifier which is expected or less possible minor contamination during sample processing/library preparation. For full bacterial classification information of captured dilutions and the mock community sample, please see Appendix Section 9.6 for Sankey diagrams.

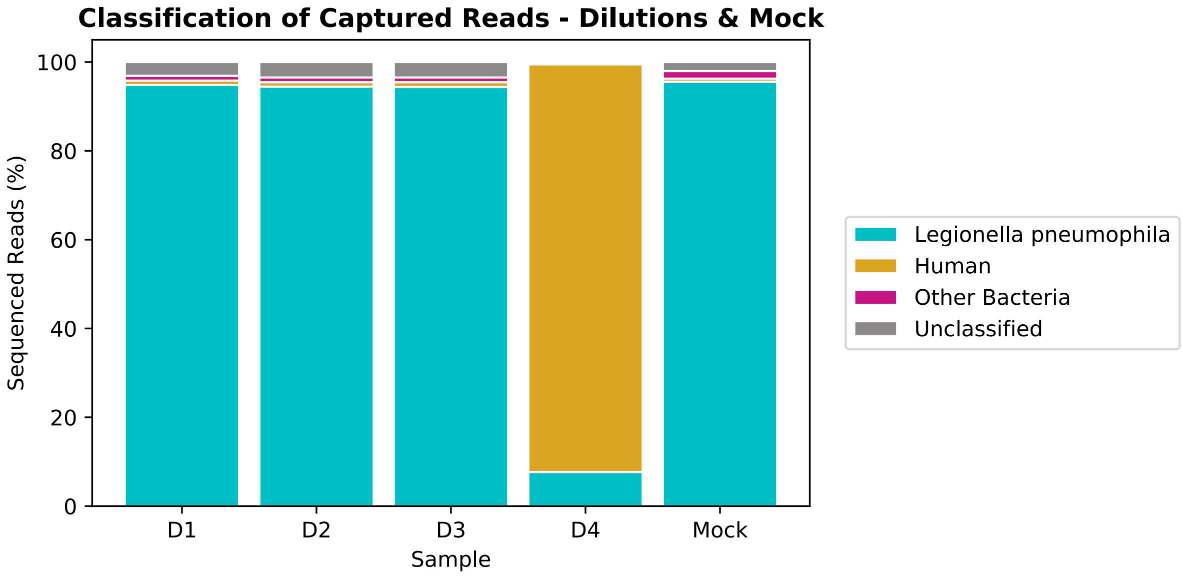


Figure 5.1 Taxonomic Classification of the Captured Reads for the Dilution samples and the Mock sample. Dilution samples: from 10^6 *Legionella* copies [D1] to 10^3 *Legionella* copies [D4]. Dilution samples D1 to D4 contained *L. pneumophila* only and few reads were classified as other *Legionella* species. For the mock sample, 97.72 % of reads were classified as belonging to the genus *Legionella*. Of this, 96.24 % of reads were classified as *L. pneumophila* and 1.48 % were classified as belonging to other *Legionella* species

Figure 5.2 shows the proportions of classified and unclassified captured reads for the clinical samples. The proportions of reads classified as *L. pneumophila* were 1.23 % for H1, 0.09 % for H2, 32.2 % for H3, 14.2 % for H4, 3.8 % for H5, 97 % for H6, 0.28 % for H7, 62.4 % for H8, 92.6 % for H9 and 0.73 % for H10. Non-specific capture of other bacteria and human DNA occurred in low *Legionella* abundance specimens. High abundance *Legionella* specimens had less non-specific capture, as expected based on results from the dilution tests and information obtained from prior qPCR analysis by PHE (see **Table 5.2**) and 16S rRNA gene sequencing. Sankey diagrams showing full bacterial classification data for all clinical samples can be viewed in Appendix Section 9.6.

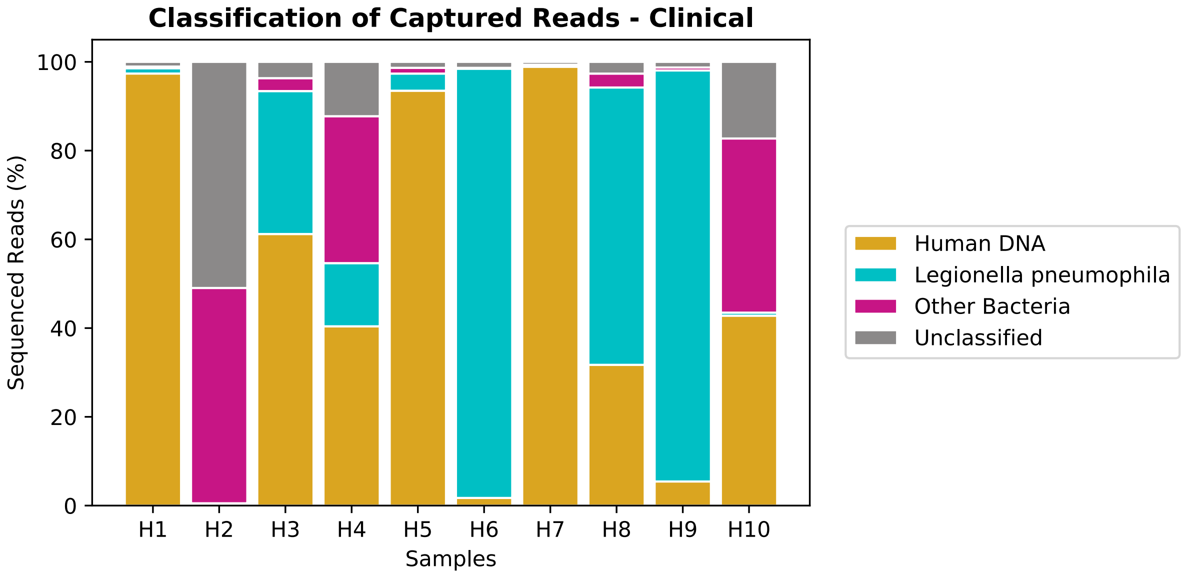


Figure 5.2 Taxonomic Classification of the Captured Reads from the Clinical Samples. The proportions of reads classified as *L. pneumophila* were 1.23 % for H1, 0.09 % for H2, 32.2 % for H3, 14.2 % for H4, 3.8 % for H5, 97 % for H6, 0.28 % for H7, 62.4 % for H8, 92.6 % for H9 and 0.73 % for H10. Non-specific capture of other bacteria and human DNA occurred in low *Legionella* abundance specimens.

Figure 5.3 shows the proportion of classified and unclassified reads for the environmental samples. The proportion of reads classified as *L. pneumophila* were 9.21 % for E1, 27.16 % for E2, 0.02 % for E3, 0.37 % for E4, 15.49 % for E5, 30.19 % for E6, 18.77 % for E7, 20.34 % for E8 and 90.31 % for E9. The results of taxonomic classification also indicated that in addition to *L. pneumophila*, a proportion of reads were assigned to other *Legionella* species in some samples: 10.76 % for E1, 8.46 % for E2, 15.04 % for E7, 1.94 % for E8 and 3.45 % for E9. Whilst E3 and E4 had 0.01 % and 0.54 % reads assigned to other *Legionella* species, the number of total *Legionella* reads was very low therefore causing difficulty in distinguishing between reads from *L. pneumophila* and other *Legionella* species. Sankey diagrams showing full bacterial classification data for all clinical samples can be viewed in Appendix Section 9.6.

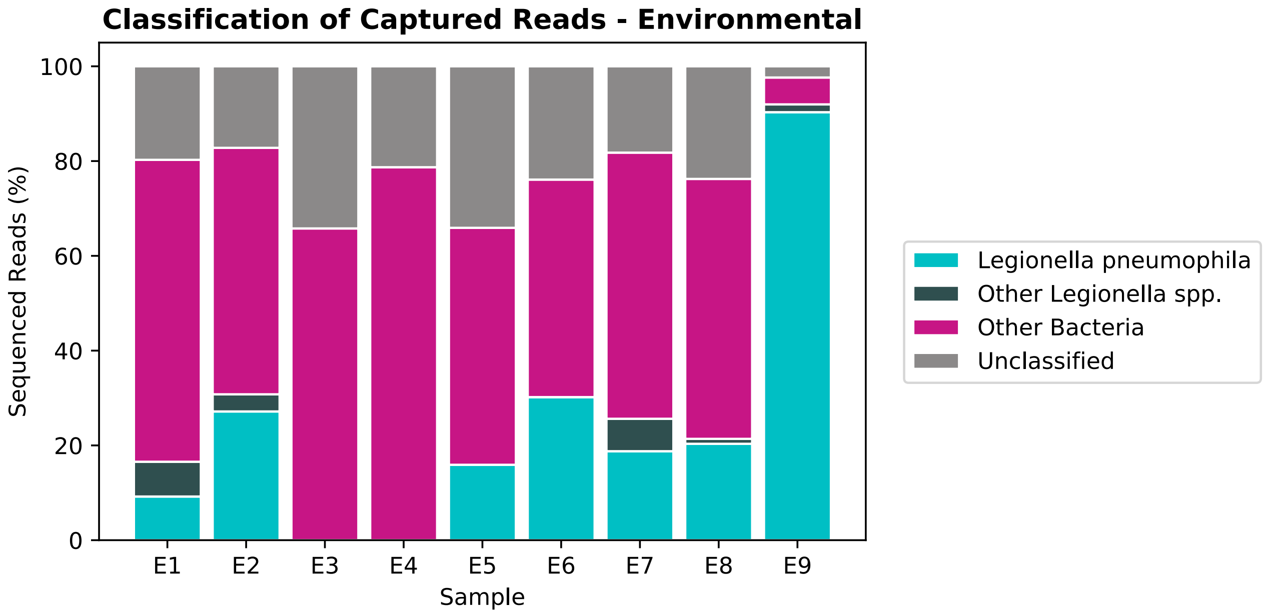


Figure 5.3 Taxonomic Classification of the Captured Reads from the Environmental Samples. The proportion of reads classified as *L. pneumophila* were 9.21 % for E1, 27.16 % for E2, 0.02 % for E3, 0.37 % for E4, 15.49 % for E5, 30.19 % for E6, 18.77 % for E7, 20.34 % for E8 and 90.31 % for E9. Taxonomic classification results also indicated that in addition to *L. pneumophila*, a proportion of reads were assigned to other *Legionella* species in a number of samples: 10.76 % for E1, 8.46 % for E2, 15.04 % for E7, 1.94 % for E8 and 3.45 % for E9.

5.4.1.4 Typeability of Target Capture Samples

5.4.1.4.(i) *L. pneumophila In silico* Sequence-Based Typing (Traditional Scheme)

An *in silico* sequence-based typing analysis, analogous to MLST, based on the traditional ESGLI 7-loci scheme, was carried out on the QC'ed sequencing reads of each sample. The aim was to determine if a *L. pneumophila* sequence type could be determined from the capture data. The allele numbers derived from the capture data were then compared to the SBT result from PHE which had been performed either on actual isolates obtained from samples or by a direct nested approach on the sample without culture.

Table 5.11 shows the sequence type results from the target capture data for dilutions, mock, clinical and environmental samples. Allele numbers marked in blue represent alleles determined with certainty. Allele numbers marked in yellow represent alleles determined with uncertainty. Uncertainty signifies low depth coverage of bases, missing bases or sequence truncation. Alleles marked in grey indicate that > 90 % of the allele reference was not covered, therefore an allele number could not be determined. For the dilutions, full sequence type was determined in samples containing 10^6 (D1), 10^5 (D2) and 10^4 *Legionella* copies (D3). However, in the sample containing 10^3 *Legionella* copies (D4), certainty fell for two alleles (*mip* and *neuA/h*) and one allele was not determined (*pilE*). For the mock sample, a ST36 was determined despite the sample containing genomic information from both ST36 and ST1 material.

In the case of the clinical samples H1, H2, H7 and H10, no allele number was determined. Samples H6, H8 and H9 had full sequence type determined and alleles matched those previously reported by PHE. Sample H3 had a full sequence type determined however the *neuA/h* allele was determined with uncertainty and did not match that reported by PHE. H4 had 3 alleles (*flaA*, *asd* and *mip*) determined with uncertainty however they matched the allele numbers reported by PHE. Sample H5 had 1 allele (*mompS*) determined with certainty and matching that reported by PHE.

Sequence type analysis of environmental samples revealed sequence types not reported in the ESGLI database from 3 specimens: E5, E6 and E8. Sample E9 had full sequence type information for ST1. Samples E1 and E2 had partial sequence types consistent with ST1. E7 had two allele numbers determined: *flaA* with certainty and a *neuA/h* with uncertainty. Since there was no available sequence type information for the environmental specimens from PHE, a comparison between results obtained was not possible.

Table 5.11 *In silico* Analyses for Traditional *L. pneumophila* Sequence-Based Typing.

	Sample	***flaA***	***pilE***	***asd***	***mip***	***mompS***	***proA***	***neuA/h***	**Target Capture ST**	**Actual ST (Allele Numbers)**
	D1 (10^6)	**3**	**4**	**1**	**1**	**14**	**9**	**1**	ST36	ST36
	D2 (10^5)	**3**	**4**	**1**	**1**	**14**	**9**	**1**	ST36	ST36
	D3 (10^4)	**3**	**4**	**1**	**1**	**14**	**9**	**1**	ST36	ST36
	D4 (10^3)	**3**		**1**	**1**	**14**	**9**	**1**	ST36	ST36
	Mock (Mix)	**3**	**4**	**1**	**1**	**14**	**9**	**1**	ST36	ST36 & ST1
	H1								No Info	ST47
	H2								No Info	ST37
	H3	**12**	**8**	**11**	**21**	**40**	**12**	**23**	?	ST1694 (*12,8,11,21,40,12,9*)
	H4	**2**		**3**	**28**				?	ST81 (*2,10,3,28,9,4,9*)
	H5					**9**			?	Consistent with ST616 (*2,0,0,10,9,4,28*)
	H6	**6**	**10**	**14**	**10**	**21**	**3**	**9**	ST2287	ST2287
	H7								No info	Not Tested
	H8	**2**	**3**	**18**	**13**	**2**	**1**	**6**	ST445	ST445
	H9	**1**	**4**	**3**	**1**	**1**	**1**	**1**	ST1	ST1
	H10								No Info	ST616
	E1		**4**		**1**	**1**		**1**	?	Unknown
	E2	**1**	**4**	**3**		**1**	**1**		?	Unknown
	E3								No info	Unknown
	E4								No info	Unknown
	E5	**6**	**10**	**15**	**12**	**12**	**4**	**11**	New ST	Unknown
	E6	**6**	**10**	**15**	**12**	**12**	**4**	**11**	New ST	Unknown
	E7	**11**						**13**	?	Unknown
	E8	**11**	**14**	**16**	**16**	**7**	**13**	**2**	New ST	Unknown
	E9	**1**	**4**	**3**	**1**	**1**	**1**	**1**	ST1	Unknown

🟦 Determined

🟧 Determined with uncertainty

⬜ Undetermined

5.4.1.4.(ii) *L. pneumophila* 50-Gene MLST – Presence/Absence Analysis

A gene presence/absence analysis was carried to determine the typeability of genes from the target capture data based on the extended 50-gene MLST scheme proposed by David *et al.*, 2016(b). Samples were mapped to a gene database based on the Philadelphia-1 reference sequence and the presence of genes with >= 90 % coverage only are reported. For dilutions D1 to D3 and the mock sample, all 50 genes were typeable. In the case of D4, 44 of the 50 genes were typeable. Typeability for the clinical samples varied from 0 to 49 of the 50 genes. For the environmental samples, typeability varied from 0 to 50 of the 50 genes (**Figure 5.4**). The list of 50 genes and their presence/absence in samples can be viewed in Appendix Section 9.7.

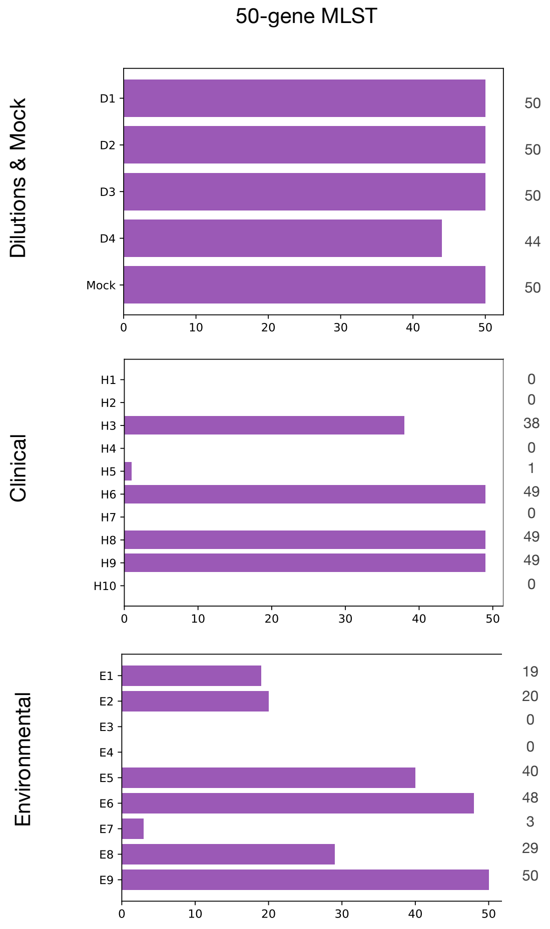


Figure 5.4 Quality Analysis of Target Capture Data based on the typeability of 50 core genes of the extended MLST scheme. For dilutions D1 to D3 and the mock sample, all 50 genes were typeable. In the case of D4, 44 of the 50 genes were typeable. Typeability for the clinical samples varied from 0 to 49 of the 50 genes. For the environmental samples, typeability varied from 0 to 50 of the 50 genes.

5.4.2 Mixed *L. pneumophila* Sequence Types: Strain Estimation Analysis

The original purpose of the StrainEst analysis was to investigate samples for mixtures of *L. pneumophila* sequence types, as validated in Chapter 3, Section 3.4.4.2. In this current chapter, true *L. pneumophila* sequence type mixtures were not identified in any of the clinical or environmental samples studied. This was concluded under the assumption that a mixed infection could only be confirmed where SNP depth was >= 10x and the minimum threshold for identification was 99 %. However, upon adjusting depth and identity thresholds, it became evident that even with sparse *Legionella* data in some samples, a strain could be predicted. *Legionella* reads from samples H1, H2, H4, H5, H7 and H10 were of relatively low abundance. A large proportion of reads had been identified as being off-target (prior analysis Section 5.4.1.3) with sequence type (ST) determination not possible either (prior analysis Section 5.4.1.4.(i)). When a matching ST was present in the StrainEst database, as was the case with samples H1, H2, H5 and H10, a corresponding ST was successfully predicted as the only strain.

In the case of H3, H4, H6 and H8 where strain representatives of ST1694, ST81, ST228, ST445 were not present in the database, a number of strains were inferred with an identity of less than 99 %. These inferred strains shared allele numbers with the actual sequence type. H7 was not previously tested at PHE, therefore it was not possible to confirm if it matched the predicted ST. ST1 was predicted in each of samples H9, E1, E2 and E9 with an identity of 99 %. A number of strains were inferred for samples E5, E6, E7 and E8 with an identity of < 99 %. In all cases, where allele information was known from *in silico* SBT, allele combinations from strain predictions were shared with the actual ST. Strain predictions for the clinical samples are outlined in **Table 5.12** whilst strain predictions for the environmental samples are outlined in **Table 5.13**.

Table 5.12 StrainEst Analysis of Clinical Samples.

Sample Name	Known ST (Alleles)	StrainEst Results			
		Identity (%)	**Strain**	**ST (Alleles)**	**SNP D.O.C (Min/Max)**
H1	ST47 (*5,10,22,15,6,2,6*)	98	GCF_900049305	ST47	2 / 6
H2	ST37 (*3,4,1,1,14,9,11*)	97	GCF_900062465 GCF_900073025	ST37 ST37	1 / 2
H3	ST1694* (*12,8,11,21,40,12,9*)	96	GCF_900063055 GCF_003004295	ST42 (*4,7,11,3,11,12,9*) ST44 (*4,8,11,10,10,12,2*)	4 / 39

H4	ST81*
	(*2,10,3,28,9,4,9*)
		97	GCF_000586095
			GCF_900060725
			GCF_000699225
				ST1362 (*2,10,3,28,9,4,207*)
				ST2122(*2,10,3,10,9,4,9*)
				ST1323(*6,10,3,28,9,4,207*)
					1 / 8

H5	Consistent with ST616 (*2,10,3,10,9,4,28*)	99	GCF_000823485	ST616	2 / 10
H6	ST2287* (*6,10,14,10,21,3,9*)	97	GCF_000823425 GCF_002002625 GCF_900053335 GCF_001583565	ST? (*6,10,14,28,4/9,3,207*) ST? (*6,10,15,3,21,14,9*) ST2 (*6,10,19,3,19,4,9*) ST1119 (*2,10,14,10,21,4,3*)	16 / 70
H7	Not Tested	97	GCF_001766275	ST? (*12,9,26,5,3/26,17,15*)	2 / 6
H8	ST445* (*2,3,18,13,2,1,6*)	98	GCF_900063795 GCF_900052905	ST? (*2,10,18,10,63,1,9*) ST23 (*2,3,9,10,2,1,6*)	19 / 90
H9	ST1 (*1,4,3,1,1,1,1*)	99	GCF_000953915 GCF_001601485 GCF_001601245 GCF_900053675	ST1 ST1 ST1 ST1	23 / 143
H10	ST616 (*2,10,3,10,9,4,28*)	99	GCF_000823485	ST616	2 / 7

*= sequence type not in database. D.O.C = depth of coverage. Multiple STs are predicted if the ST is not present in the database and/or the threshold is set below 99 %.

Table 5.13 StrainEst Analysis of Environmental Samples.

Sample Name	Known ST (Alleles)	StrainEst Results			
		Identity (%)	**Strain**	**ST (Alleles)**	**SNP D.O.C (Min/Max)**
E1	Consistent with ST1 (*0,4,0,1,1,0,1*)	99	GCF_000953915 GCF_001601245	ST1 ST1	2 / 14
E2	Consistent with ST1 (*1 0,4,0,1,1,0,1*)	99	GCF_000953915	ST1	2 / 21
E3	Unknown	No Result			
E4	Unknown	No Result			
E5	ST? (*6,10,15,12,12,4,11*)	98	GCF_000092625 GCF_900059935	ST578 (*6,10,15,13,9,14,6*) ST? (*6,10,15,24,17/98,14,6*)	4 / 29
E6	ST? (*6,10,15,12,12,4,11*)	97	GCF_000092625 GCF_900059935	ST578 (*6,10,15,13,9,14,6*) ST? (*6,10,15,24,17/98,14,6*)	8 / 49
E7	ST? (*11,0,0,0,0,0,0*)	98	GCF_002934205	ST? (*11,14,16,16,7/15,13,2*)	2 / 12
E8	ST? (*11,14,16,16,7,13,2*)	98	GCF_002934205 GCF_001582295 GCF_001549915 GCF_003004255	ST? (*11,14,16,16,7/15,13,2*) ST154 (*11,14,16,16,15,13,2*) ST? (*11,14,16,10,7,13,2*) ST? (*11,14,16,16,7/15,13,2*)	5 / 47
E9	ST1 (*1,4,3,1,1,1,1*)	99	GCF_000953915 GCF_001601245	ST1 ST1	49 / 243

DOC = depth of coverage. Multiple STs are predicted if the ST is not present in the database and/or the threshold is set below 99 %.

5.4.3 Genome Assembly

Genomes were assembled and assemblies were decontaminated by removing contigs belonging to taxonomic orders other than the order Legionellales. A quality analysis was carried out to determine genome length, maximum contig length, N50 (the minimum contig length needed to cover 50 % of the genome. In this way the sum of the lengths of all the cotigs of N50 size or longer are greater than or equal to 50 % of the total genome sequence), GC content, genome completeness, residual contamination and the number of predicted genes. Genome draft quality (high, medium or low) was assigned based on criteria defined by Bowers *et al.*, 2017. Genome length for the dilution series samples varied from 3,416,931 to 2,975,552 bases with maximum contig length varying from 474,275 to 10,180. The reported GC content for all dilution genomes was 38 %, completeness was reported at 100 % for D1, D2 and D3 and 93 % for D4. Contamination levels were reported as 0.191 % for D1, D2 and D3 and 4 % for D4. These genomes were classified as high-quality drafts (>= 90 % completeness and < 5 % contamination). The mock sample had a genome length of 5,729,401 with maximum contig length of 95,145 and a GC content of 38.11 %. Completeness was reported as 96.55 % and contamination levels as 55.27 % (**Table 5.14**). The high contamination levels here were due to strain heterogeneity since the sample contained two *L. pneumophila* sequence types (full composition of Mock Sample see **Table 5.5**).

Table 5.14 Assembly Statistics for Dilution Series Samples and Mock Sample.

Assembled Genome	Genome Length (bases)	Max Contig Length (bases)	N50	GC Content	Completeness	Contamination	No. of predicted Genes	Draft Quality
D1	3,413,278	474,275	249,787	38 %	100 %	0.191 %	3,075	High
D2	3,416,931	362,169	243,800	38 %	100 %	0.191 %	3,084	High
D3	3,421,971	362,590	249,127	38 %	100 %	0.191 %	3,098	High
D4	2,975,552	10,180	1,761	38 %	93 %	4 %	4,091	High
Mock	5,729,401	95,145	7,977	38.11 %	96.55 %	55.27	6,116	Mixed

High quality genomes were assembled for clinical samples H6, H8 and H9, with each reported as being 100 % complete and with minimal levels of contamination (0.19 %, 0.19 % and 2.22 %, respectively). A medium quality draft was assembled for sample H3 with 66.76 % completeness and 9.77 % contamination. Low quality genomes could not be assembled for samples H1, H2, H4, H5, H7 and H10 due to lack of captured coding regions (**Table 5.15**).

Table 5.15 Assembly Statistics for Clinical Samples.

Assembled Genome	Genome Length (bases)	Max Contig Length (bases)	N50	GC Content	Completeness	Contamination	No. predicted Genes	Draft Quality
H1	9,831	1,291	485	0.42	0.0 %	0.0 %	23	-
H2	1,893	554	459	0.52	0.0 %	0.0 %	4	-
H3	2,173,267	5,276	697	0.39	66.76 %	9.77 %	4,127	Medium
H4	172,680	1,080	489	0.41	2.74 %	0.0 %	412	-
H5	57,354	1,299	498	0.44	0.0 %	0.0 %	123	-
H6	3,298,638	121,016	46,521	0.38	100 %	0.19 %	2,978	High
H7	2,813	649	493	0.48	0.0 %	0.0 %	7	-
H8	3,388,205	63,201	16,496	0.38	100 %	2.22 %	3,355	High
H9	3,351,794	53,336	10,077	0.38	100 %	0.191 %	3,294	High
H10	43,591	1,223	511	0.42	4.17 %	0.0 %	89	-

High quality genomes were assembled from environmental samples E5, E6 and E8. (**Table 5.16**) Sample E9 could not be classified as a high-quality draft owing to the levels of contamination observed. E3 and E4 did not yield a draft genome. Low quality draft genomes were assembled from samples E1 and E2 (41.76 % and 76.16 % completeness, respectively). Upon closer inspection of contamination analytics for environmental samples, it was observed that a significant degree of contamination reported was due the presence of strain heterogeneity in the sample rather than contamination from divergent taxa. Owing to this and the previous observation of reads classified as *Legionella* species other than *L. pneumophila* in (see Section 5.4.1.3), the presence of mixed *Legionella* species was further explored (next Section 5.4.4).

Table 5.16 Assembly Statistics for Environmental Samples.

Assembled Genome	Genome Length (bases)	Max Contig Length (bases)	N50	GC Content	Completeness	Contamination	No. predicted Genes	Draft Quality
E1	1,567,970	21,296	692	0.40	41.76 %	3.1 %	3,013	Low
E2	2,126,983	65,249	902	0.39	76.16 %	14.8 %	3,694	Low
E3	1,505	648	429	0.455	0.0 %	0.0 %	3	-
E4	70,158	4,867	694	0.429	0.0 %	0.0 %	124	-
E5	3,194,627	15,472	2,477	0.395	95.11 %	2.72 %	4,013	High
E6	3,583,038	61,148	9,592	0.394	98.85 %	1.39 %	3,664	High
E7	986,197	65,060	687	0.405	22.49 %	2.42 %	1,892	Low
E8	3,444,312	29,850	4,018	0.388	97.8 %	2.67 %	3,921	High
E9	4,035,312	223,629	90,798	0.386	100 %	21.3 %	4,004	High

5.4.4 Mixed *Legionella* Species Analysis

Next all assembled genomes (Section 5.4.3) were analysed for the presence of multiple copies of single copy genes (with amino acid sequence identities of >= 90 %). Whilst some multiple copies were reported for the clinical samples, after careful manual inspection it was clear that these were the result of either mis-assembly or the presence of a previously unrecognised contaminants from more divergent taxa. In the case of environmental samples E1, E2, E7, E8 and E9 however, there was evidence for the presence of single copy genes for multiple *Legionella* species. Phylogenetic analyses of partial single copy amino acid sequences were therefore carried out.

In the case of E1, there was evidence for the presence of two *Legionella* species. E1_Species1 clustered with the previously determined *L. pneumophila* and E1_Species2 clustered with *L. drozaskii* (**Figure 5.5**). For E2, the single copy amino acid sequences clustered closely with *L. pneumophila* (E2_Species1) and *L. anisa* (E2_Species2) (**Figure**

5.6). There was evidence for the presence of three *Legionella* species in E7 (**Figure 5.7 A** and **B**). E7 amino acid sequences clustered with *L. geestiana* (E7_Species1), *L. anisa* (E7_Species2) and *L. pneumophila* (E7_Species3). Similarly, there was evidence of three *Legionella* species in E8 (**Figures 5.8 A** and **B**), with amino acid sequences clustering with *L. pneumophila* (E8_Species1) and *L. geestiana* (E8_Species3). E8_Species2 did not cluster specifically with any one *Legionella* species present in the analysis. In the case of sample E9, there was evidence for the presence of three *Legionella* species (**Figure 5.9**), with amino acid sequences clustering with *L. anisa* (E9_Species2) and *L. pneumophila* (E9_Species3). E9_Species1 did not cluster specifically with any one *Legionella* species, however in this analysis *L. shakespearei* was the closest related genome. The lack of clustering of Species 1 and Species 2 from E9 and E8 is likely due to the lack of inclusion of protein sequences for all known *Legionella* species. Resolution of phylogenetic trees (bootstrap reports for likelihood) in some cases was quite low due to the partial profiles of the single copy sequences.

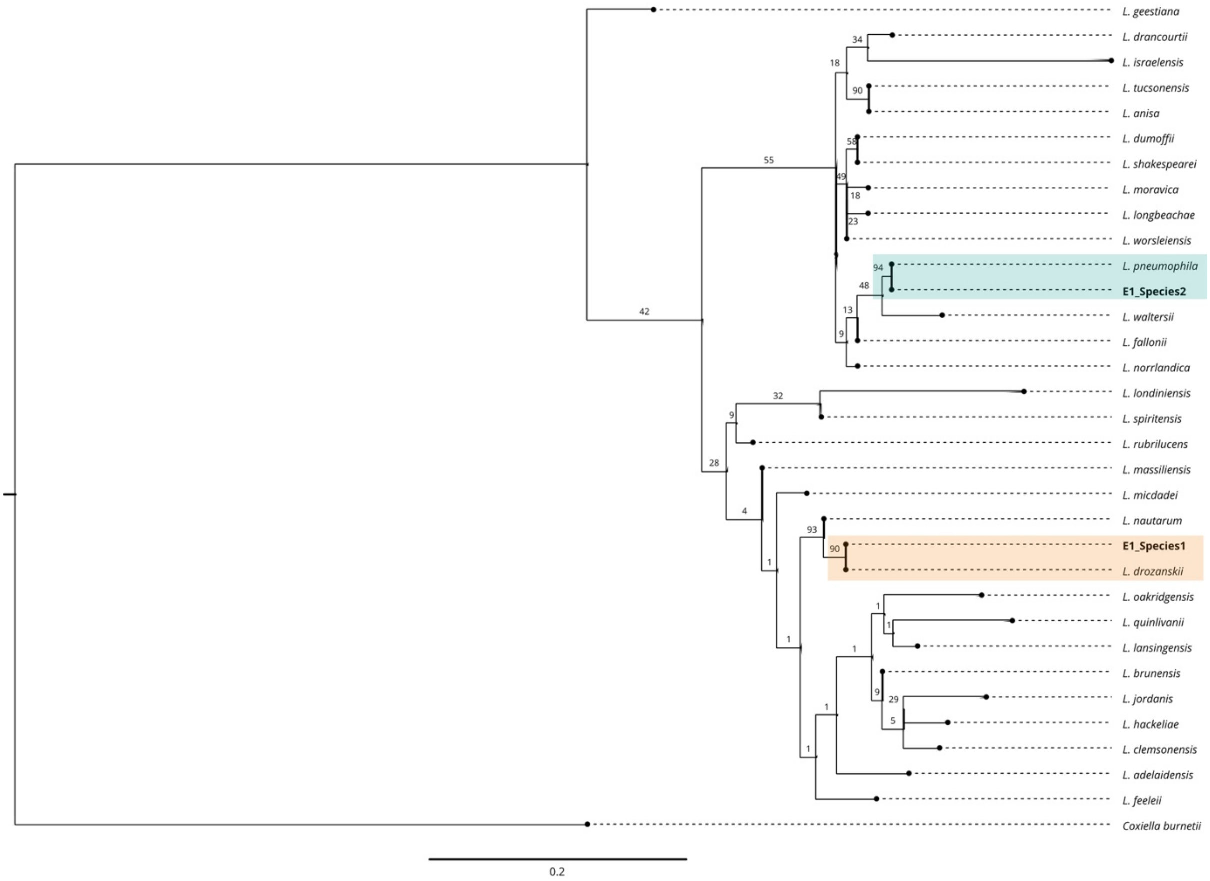


Figure 5.5 Phylogenetic contextualisation of *Legionella* species identified in sample **E1.** Based on concatenated partial amino acid sequences from two single copy genes: ribosomal protein L10 (PF00466) and ribosomal protein L2, C-terminal domain (PF03947). The genes were identified in multiple copies using `CheckM` and the Pfam database.

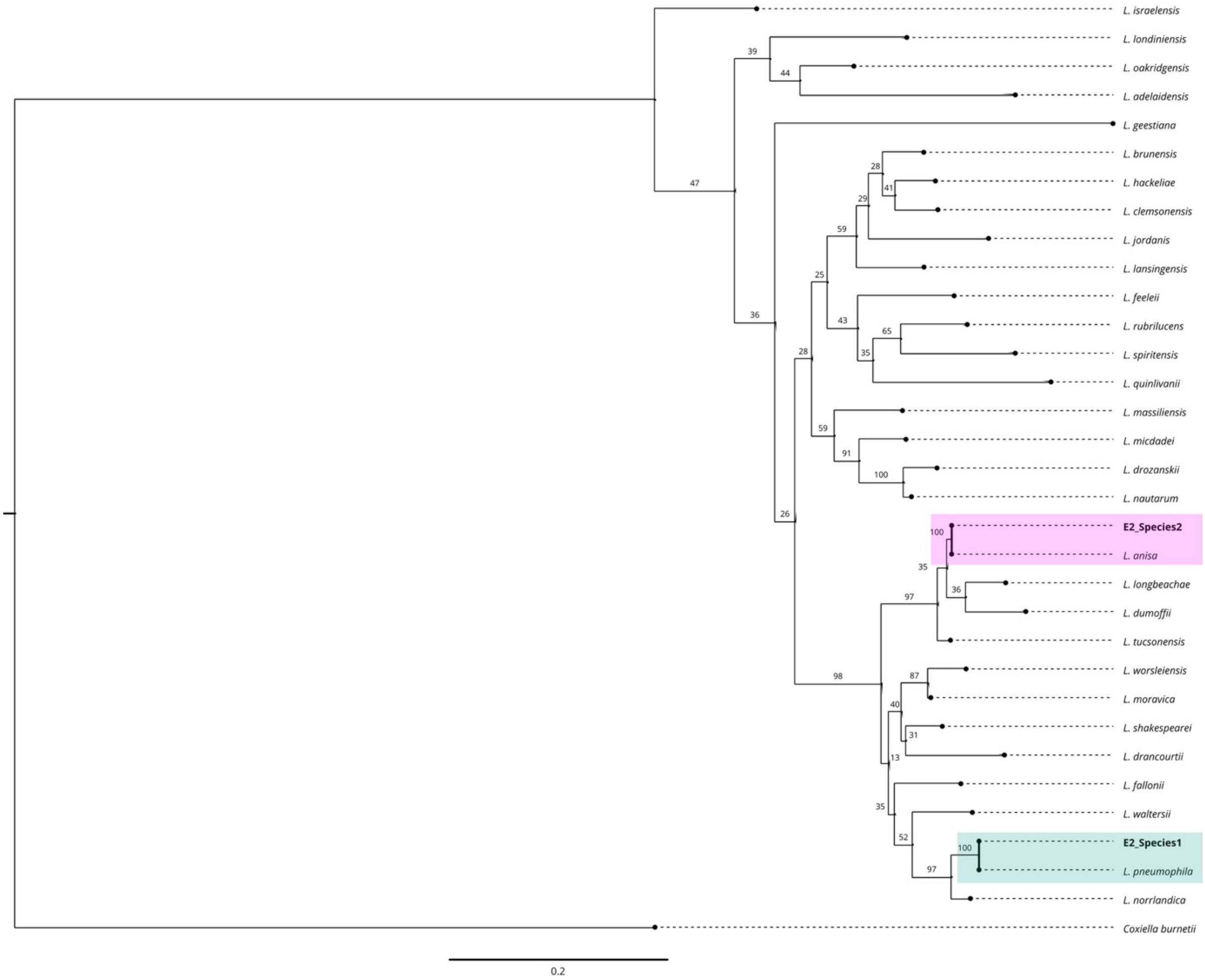


Figure 5.6 Phylogenetic contextualisation of *Legionella* species identified in sample **E2.** Based on concatenated partial amino acid sequences from four single copy genes: ribosomal protein L21 (PF00829), ribosomal protein L27 (PF01016), GTP1/OBG (PF01018) and bacterial trigger factor protein (PF05697). The genes were identified and reported in multiple copies using `CheckM` and the Pfam database.

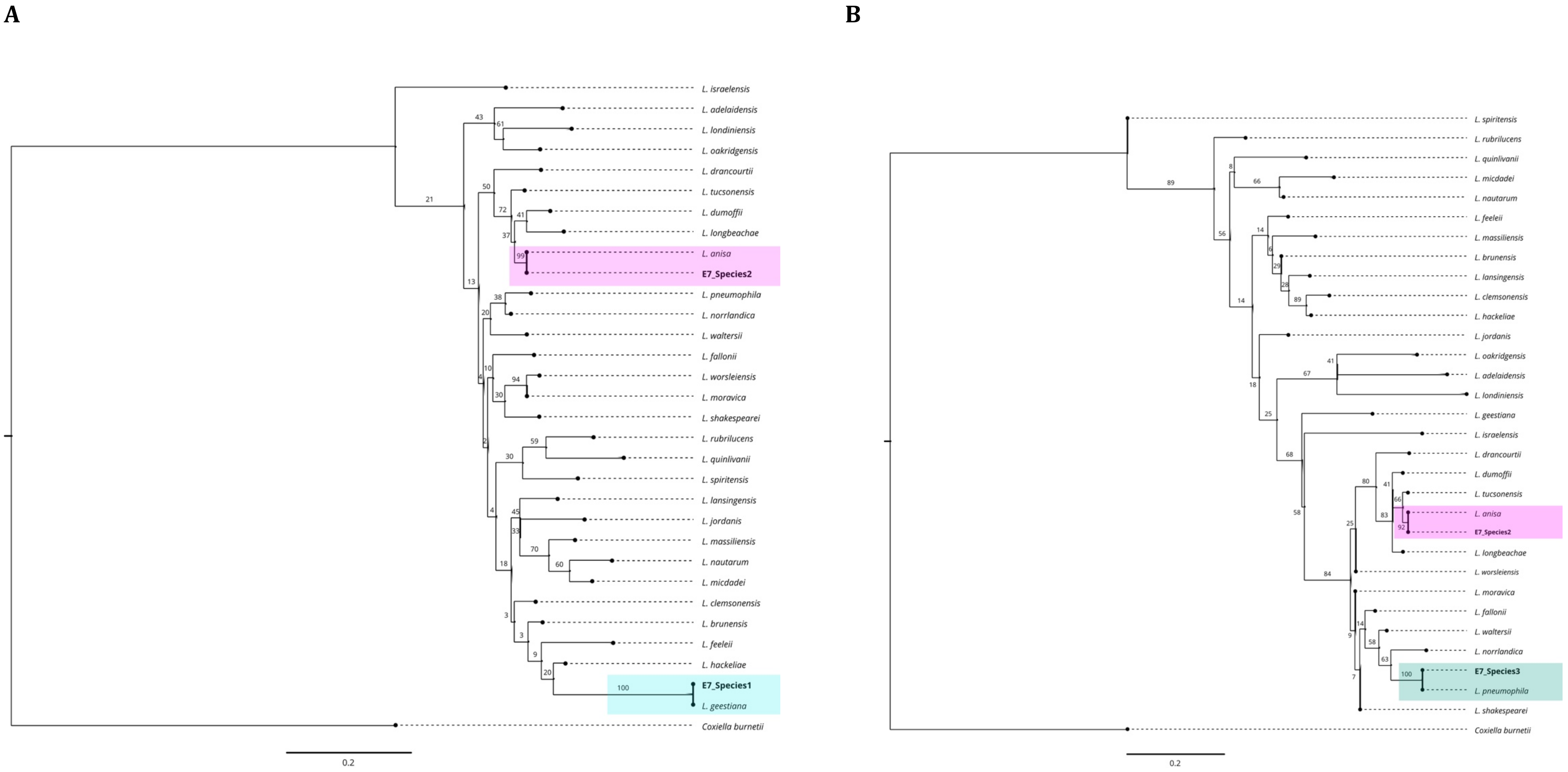


Figure 5.7 Phylogenetic contextualisation of *Legionella* species identified in sample **E7.** Based on concatenated partial amino acid sequences from (**A**) three single copy genes for Species1 and Species2: bacterial trigger factor protein (PF05697), NADH dehydrogenase (PF00507) and ribosomal protein L7/L12 (PF0542) and (**B**) one partial single copy gene for Species 2 and Species3: ribosomal protein S8 (PF00410). The genes were identified and reported in multiple copies using `CheckM` and the Pfam database.

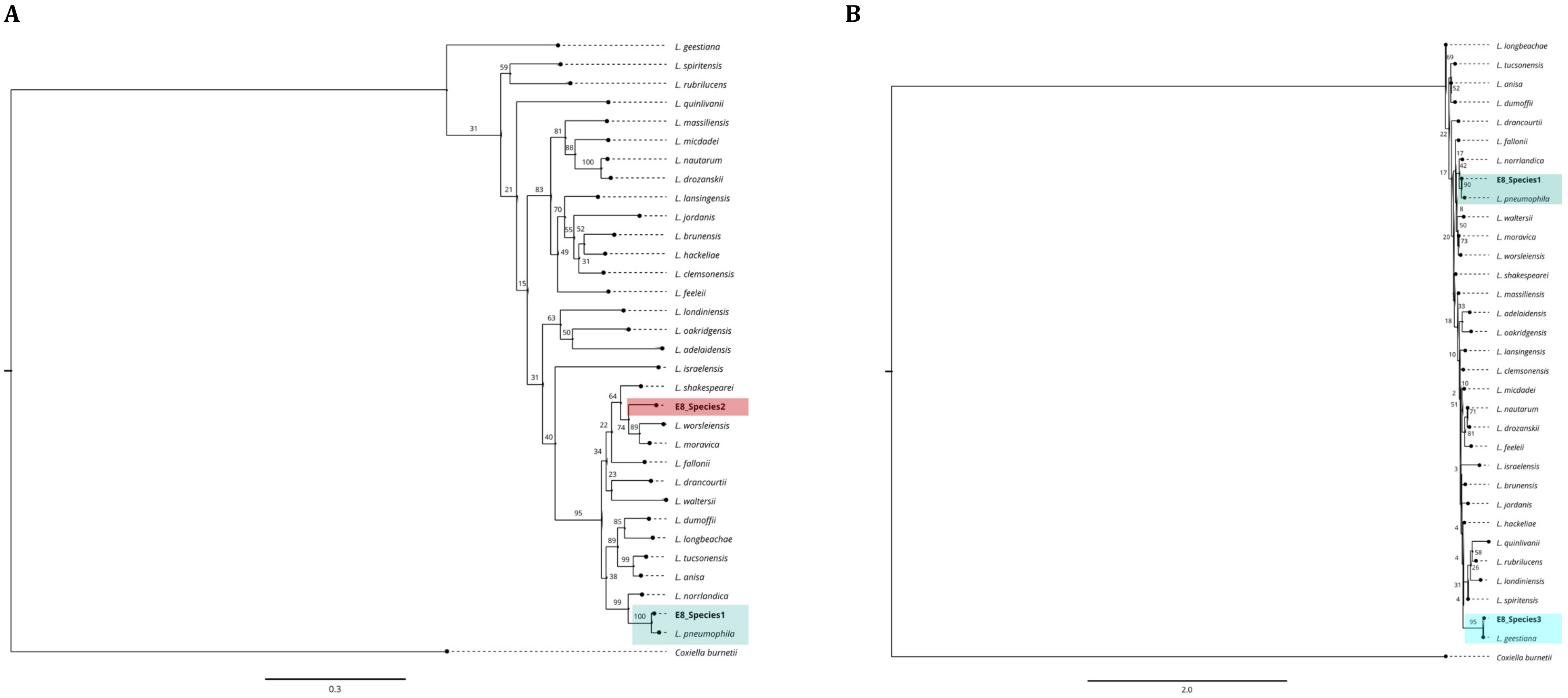


Figure 5.8 Phylogenetic contextualisation of *Legionella* species identified in sample **E8.** Based on the concatenated partial amino acid sequences from (**A**) five partial single copy genes for Species 1 and Species 2: bacterial trigger factor protein (PF05697), ribosomal protein L10 (PF00466), ribosomal protein L25 (PF01386), RimP N-terminal domain (PF02576) and NusA N-terminal domain (PF08529) and (**B**) two partial single copy genes for Species 1 and Species 3: bacterial trigger factor protein (PF05697) and SecG (PF03840). The genes were identified and reported in multiple copies using `CheckM` and the Pfam database.

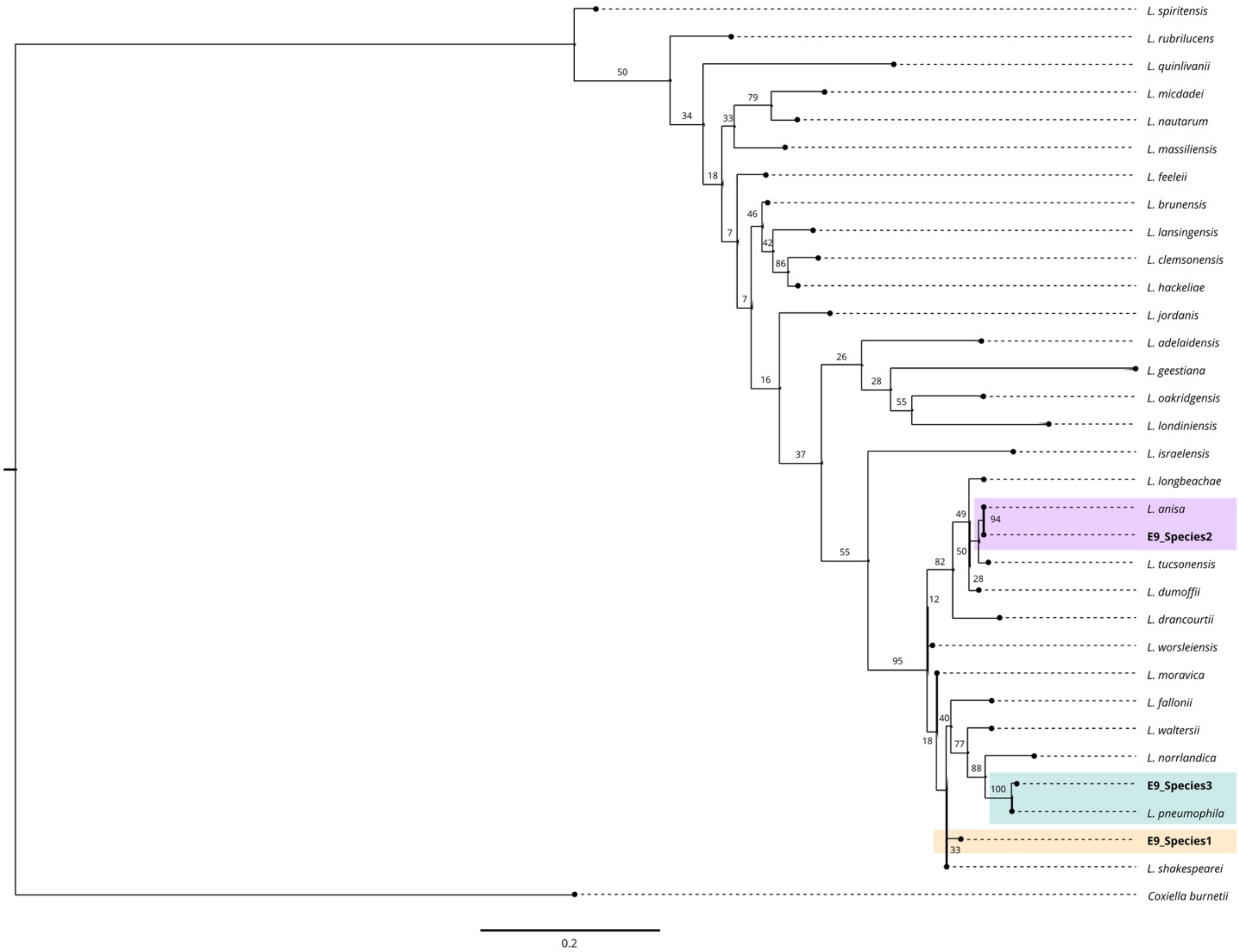


Figure 5.9 Phylogenetic contextualisation of *Legionella* species identified in sample **E9.** Based on concatenated partial amino acid sequences from three single copy genes: ribosomal protein S19 (PF00203), ribosomal protein L22/L17 (PF00237) and ribosomal protein S8 (PF00410). The genes were identified and reported in multiple copies using `CheckM` and the Pfam database.

5.5 Discussion

The aim of the current pilot study was to directly capture and sequence *L. pneumophila* genomes from a test panel of 10 clinical and 9 environmental samples. Samples (Dilution series and Mock community) with known copy numbers and a mixture of sequence types were also included to determine analytical cut-offs. The objectives were to evaluate the reference base coverage, proportion of on-target reads, mean depth of coverage and *L. pneumophila* typeability using the capture approach. Secondary aims were to evaluate samples for a mixture of *L. pneumophila* sequence types or *Legionella* species.

Out of the 19 clinical and environmental samples, good reference base coverage, proportion of on-target reads and mean depth of coverage was achieved for 3 clinical samples (H6, H8 and H9) and 4 environmental samples (E5, E6, E8 and E9). Poor genome coverage and depth of coverage from a number of other test specimens was likely due to the presence of few *Legionella* genome copies. This is exemplified by the dilution tests which contained 10^6, 10^5, 10^4 and 10^3 *L. pneumophila* genome copies. A cut-off for reliable allele number determination was established between Dilution Test 3 (10^4 copies) and Dilution Test 4 (10^3 copies). While there was relatively good base coverage for D4 capture data (greater than 90 % of the reference genome bases were covered), the mean depth of coverage fell significantly when compared to D3. This reduced coverage depth could compromise downstream analyses such as variant calling which requires a high depth of coverage to achieve reliable results. Similarly, in the case of low *Legionella* abundance specimens, there was evidence of increased read duplication levels. The high duplication levels were not attributable to the presence of optical duplicates therefore it is probable that duplicates arose as a result of PCR and enrichment.

Similarly, low abundance *Legionella* specimens had a smaller proportion of reads aligning to the intended target sequence. Taxonomic classification and human DNA screening were carried out to determine the proportions and composition of off-target reads in the sequenced data. The presence of off-target reads in the clinical data was primarily attributable to human DNA. Since the majority of reads from high abundance specimens were on-target, the non-specific capture is hypothesised here as being due to the low abundance of *Legionella* in the specimen rather than complementarity with other microorganisms or human DNA sequences.

Out of the 19 clinical and environmental samples, a full *L. pneumophila* sequence type was determined from 8 samples and a partial sequence type from 5 samples, demonstrating for the first time, the application of SBT to metagenomic data from LD cases. Clinical sample sequence types matched those previously determined by PHE however sample S3 had one allelic discrepancy between the sequence type determined from the capture data and the sequence type from culture reported by PHE. This was due to low sequence coverage or poor quality/missing bases from the capture data allele sequence. Two previously unreported sequence types were found in environmental specimens E5 and E6 (alleles *6, 10, 15, 12, 12, 4, 11*) and E8 (alleles *11, 14, 16, 16, 7, 13, 2*) which is consistent with *L. pneumophila subspecies fraseri* sequence type pattern.

The traditional 7-loci SBT scheme (Gaia *et al.,* 2005, Ratzow *et al.,* 2007, Mentasti *et al.,* 2014) is important in the epidemiological typing of Legionnaires disease cases and environmental sources worldwide. Additionally, the practicality of the approach has resulted in rapid dissemination of data via the SBT web server (http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php). Currently (19th July 2019), the SBT server hosts 12,935 sample records composed of 2,791 *L. pneumophila* sequence types. A significant proportion of Legionnaires disease cases are however caused by a limited number of *L. pneumophila* sequence types (Borchardt *et al.,* 2008, Harrison *et al.,* 2009, Tijet *et al.,* 2010, David *et al.,* 2016[a]). Owing to this, the traditional SBT approach is not of high enough resolution to discriminate between different outbreaks caused by the same sequence type and highly reliant on the pre-culture of *Legionella* in most cases. As a measure of genome quality, samples were therefore mapped to a *L. pneumophila* reference genome and the 50 core genes extracted to determine how many were sequenced with > 90 % base coverage. All 50 genes were sequenced from the mock sample and dilution series samples D1, D2, D3. Dilution sample D4 (containing 10^3 genome copies) had 44 of the 50 complete genes sequenced. For high abundance *Legionella* clinical specimens (H6, H8 and H9), 49 out of 50 complete genes were sequenced [demonstrating for the first time the 50-Core Gene MLST schema may be applicable without culture in some specimens]. Interestingly, for the environmental sample E8, where all 7 alleles were typed, only 29 out of

50 complete genes were sequenced.

The *L. pneumophila* genome is known to demonstrate high plasticity and recombination events are frequent (Gomez-Valero *et al.,* 2011, Sanchez-Buso *et al.,* 2014, David *et al.,* 2017[b]). This has the potential to reduce the binding affinity of some baits to the whole genome. In this study, a variety of different sequence types (ST36, ST1, ST2287, ST445, a partial profile from ST81 and two unreported sequence types) were captured from the test samples. These included sequence types that were not represented in the genome database for bait design. However due to the number of sequence types currently reported for the *L. pneumophila* species, future studies would be needed to validate the capture of different complexes.

There was no evidence of mixed *L. pneumophila* sequence types in the clinical and environmental samples using the `StrainEst` approach (Albanese *et al.,* 2017). The strain estimation analysis was originally validated to investigate mixtures of *L. pneumophila* sequence types. However, an alternative use-case was studied: determining if a strain or a closely-related strain could be predicted from limited *Legionella* capture data. Such an approach could potentially be applied for example, to the rapid screening of samples to rule out an environmental source during an investigation, but not as a diagnostic tool due to the low resolution of the data. A further disadvantage of this approach is that a database of *L. pneumophila* genomes must be regularly updated to address new sequence types.

In relation to genome assembly from the data generated, a total of 6 high quality, 1 medium quality and 3 low quality draft genomes were successfully constructed. The `CheckM` program reports contamination resulting from the presence of genomic fragments from multiple closely related strains or genomic fragments from more divergent taxa. It does this by examining the amino acid identity between conserved single copy marker genes present in multiple copies in the assembly. If the marker genes present in multiple copies have an amino acid identity of >= 90 %, it is likely that they are representative of strain heterogeneity within the assembly.

There was good evidence for the presence of other *Legionella* species in addition to *L. pneumophila* in the environmental samples E1, E2, E7, E8 and E9. Partial single copy amino acid sequences were concatenated to gain a better understanding of their phylogenetic localisation. Clustering of amino acid sequences was observed for *Legionella* species other than *L. pneumophila*. It is only possible to conclude from this analysis that other *Legionella* species were present and that they clustered closely with certain *Legionella* species. Due to the low resolution of the approach, the identity of these species could not be confirmed based on this analysis alone. It is not uncommon to find multiple *Legionella* species in environmental sources. They are ubiquitous in freshwater and survive in water systems if control strategies are not maintained (as reviewed in Chapter 1, Section 1.4).

The results of this pilot study were promising as they demonstrated, to my knowledge that for the first time, whole or partial *L. pneumophila* genomes could be recovered directly from both clinical and environmental samples which could benefit or support the timely analysis of clusters. This paves the way for future studies that improve detection and discrimination of infection without culture in clinical and environmental pathogens. The in-solution targeted hybridisation capture approach is advantageous in that it does not require culturing or a large quantity of starting DNA (<10 – 200 ng). Additionally, target capture is well suited to multiplexing and the probe design is scalable. For example, it is possible to design baits only for targets of interest such as 100s of core genes relevant to a typing scheme. Additionally, the approach does not require a new analytical pipeline and can take advantage of bioinformatic systems already established for whole genome sequences from culture.

Disadvantages of the hybridisation capture approach are that it is less specific than other approaches. Therefore, the enriched portion after capture may include DNA that does not come from the region of interest. Ideally, the system requires a minimum of 10,000 *L. pneumophila* genome copies for complete or near-complete genome recovery. For this purpose, sampling and storage conditions of samples as well as extraction methods need to be optimised. A disadvantage of the current study was the sequencing depth. The shallow depth of sequencing here was not sufficient to investigate *L. pneumophila* strain heterogeneity, as studied on other bacterial pathogens (Bachmann *et al.,* 2015, Pinto *et al.,* 2016). The next chapter will therefore address this challenge by investigation the use of the target capture approach in the investigation of clinical samples from Legionnaires disease clusters using a higher depth of sequencing.

Chapter 6.

Investigating Legionnaires' Disease Outbreaks using Metagenomic Methods

6.1 Introduction

Legionnaires' Disease (LD) as highlighted earlier (Chapter 1) is a severe atypical pneumonia that occurs in susceptible individuals exposed to aerosols from natural or man-made environments containing *Legionella* bacteria. LD cases can occur sporadically as individual cases or may be classified as part of a cluster or outbreak if epidemiological, spatial and temporal criteria are met.

Public Health England (PHE) defines a LD cluster as two or more cases of confirmed LD that appear to be linked by a common work or residential area, including a healthcare or travel-associated setting. Additionally for a case to belong to a cluster, onset of symptoms between cases should fall within a 6-month period. An outbreak is defined as two or more cases of confirmed LD where the onset of symptoms between cases occurs within weeks rather than months. There should also be epidemiological evidence of exposure to a common environmental source of infection. For further details in relation to case, cluster and outbreak definitions please see Chapter 1, Section 1.6.3.

During 2015, a total of 35 outbreaks/clusters involving cases of LD occurred in England and Wales. Of these, 12 outbreaks/clusters were defined as community-associated, 3 were healthcare-associated and 20 were travel-associated (17 for travel abroad and 3 for travel within the UK) (PHE, 2016). This Chapter explores two anonymised outbreaks that occurred in England during 2015. This was carried out by applying metagenomic sequencing methods to available sputum samples and performing a co-analysis of metagenomic data with whole genome sequences from outbreak isolates. The utility and timeliness of metagenomic methods for outbreak investigations has been previously reported for other pathogenic microorganisms (Loman *et al*., 2014, Quick *et al*., 2016, Faria *et al*., 2017, Huang *et al*., 2017, Kafetzopoulou *et al*., 2019). Additionally, pilot data from Chapter 5 demonstrated that high quality *L. pneumophila* draft genomes could be captured directly from clinical samples. Furthermore, the study of LD outbreaks or clusters using metagenomic methods could provide additional information on the landscape of infection in LD cases.

Case Study 1 involved a spatio-temporal outbreak of LD in a low incidence area, comprising two local authorities with a population of 480,000. Two cases of LD per year are expected in these areas combined (Naik *et al*., 2015). The outbreak was identified over a period of 5 months and involved 8 confirmed LD cases. One confirmed case (*L. pneumophila* serogroup 1 ST62) was excluded as symptoms began during a period of travel. A 9th case which was urinary antigen positive for *L. pneumophila* serogroup 1 upon initial NHS hospital laboratory testing but negative by UAT and qPCR at Public Health England (PHE) was also excluded from the cluster however the sample was available for sequencing for the purpose of the present study. Of the 8 cases studied, respiratory specimens were available for 6 patients and clinical/microbiological information for an additional 1 patient. No sputum sample or isolate was available for the 8th patient.

From the information available for cases 1 to 7, a full sequence type had been obtained for three of the cases by routine methods: ST47 (1 patient), ST82 (1 patient) and the novel ST2110 (1 patient). Partial sequence types had been obtained from two cases, one consistent with an ST47 profile and the other consistent with an ST1554 or ST501 profile. Four cases lived within an area of 10 km radius, two of whom reported very limited movement away from their home address whilst three either visited or worked in the stated 10 km radius area. Of note, two cases were especially closely linked through a shared potential source defined as medium/high risk. This site was a residential address (not the home of the cases) that had two pools; a plunge pool and spa pool. Although not used by either case, these were identified as potential exposure sources. A further case lived within 100 metres of this address. Three swabs from the spa pool that had been screened were found to be PCR positive for *L. pneumophila* serogroup 1 but direct SBT was not possible due to mixed *L pneumophila* populations and *Legionella* was not isolated by culture due to presence of significant concentrations of *Pseudomonas spp*. in the swab samples. After investigation of other suspected sites, an environmental source could not be confirmed. The incident management team (IMT) concluded there was no common environmental source for the cases and the cluster cases were classified as sporadic and unrelated (sometimes referred to as a pseudo outbreak).

Case Study 2 involved an outbreak of LD cases over a 1-year period associated with an industrial site. The site had 3 cooling towers providing cooling to machines within the complex. Furthermore, the nature of the industry promoted exposure of workers to aerosols. A total of 7 cases screened had been found to be *L. pneumophila* serogroup 1 positive and an ST37 was confirmed in 3 cases. Respiratory samples from 5 patients and isolates from two patients and two environmental sources within the complex (also confirmed as ST37) were available for this present study. The IMT for the outbreak concluded that the industrial site was the source of infection in all LD cases involved.

The primary aim of the current study therefore was to carry out an exploratory analysis of Case Study 1 and Case Study 2 by combining metagenomic and isolate data to understand the utility of applying metagenomic methods for the investigation of LD outbreaks.

6.2 Aims and Objectives

1. *Legionella* abundance in the bacterial community of sputum samples:
    a. Determine the proportion of the bacterial community of the cluster/outbreak sputum samples represented by the genus *Legionella* through 16S rRNA gene sequencing and analysis.
2. *L. pneumophila* capture and sequencing from sputum DNA:
    a. Perform targeted capture and sequencing from available cluster/outbreak DNA extracts, a dilution series with known copy numbers of *L. pneumophila* to determine analytical cut-offs and a mock community containing two *L. pneumophila* sequence types and five other *Legionella* species.
    b. Examine the sequence coverage across target regions, depth of coverage, the proportion of reads mapping to the intended target and typeability of captured regions by *in silico* traditional SBT, 50 core gene MLST and extended 1,455 core gene MLST.
3. Isolate sequencing:
    a. Carry out sequencing of available clinical and environmental cluster isolates.
4. Phylogenetic analysis:
    a. Carry out phylogenetic analysis based on core genes.
5. Heterozygous SNP analysis:
    a. Investigate captured and isolate data for heterozygous SNPs in core genes as evidence of mixed *L. pneumophila* populations.
    b. Establish cut-offs based on the analysis of single-strain controls.
6. Direct Oxford Nanopore Sequencing:
    a. Determine if direct Oxford Nanopore Sequencing from sputum DNA provides additional information of LD cluster investigations.
7. Genome Assembly and Pangenome Visualisation:
    a. Assemble *Legionella* partial/draft genomes from generated metagenomes.
    b. Perform a co-assembly with Oxford Nanopore generated sequences.
    c. Carry out a pangenome visualisation of captured and isolate assemblies.

6.3 Methods

6.3.1 Ethical Approval

Ethical approval was granted from the Research Ethics Committee (REC), as described in Chapter 2, Section 2.3, for the sequencing and analysis of clinical samples.

6.3.2 Microbiological Methods

All patient urine samples were initially tested by the local NHS hospital trust laboratory using a urinary antigen test (UAT). Urine and sputum samples from UAT positive patients were sent from the local laboratory to the National *Legionella* Reference Laboratory, PHE Respiratory and Vaccine Preventable Bacterial Reference Unit (RVPBRU), PHE Microbiology Reference Services, Colindale. At the RVPBRU, urine samples underwent further analysis using two commercial assays (Bartels EIA and Binax EIA). The urine sample was tested as untreated and boiled. Samples found positive by both assays after boiling were considered positive for *L. pneumophila.*

Respiratory specimens were cultured, using standard methods. Eight culture plates per sample were incubated at 35 – 37 ^{o}C in humidified air. Culture plates were read after 48 hours incubation and every 48 hours thereafter for up to 10 days. Characteristic ground glass colonies were sub-cultured onto BCYE and BCYE from which L-cysteine was omitted (BCYE-cys). Those that grew on BCYE but not on BCYE-cys were presumptively identified as Legionellae and confirmed as *L pneumophila* serogroup 1 by PCR.

DNA was extracted from respiratory samples at PHE using the MagnaPure Compact (Roche) and examined by real time PCR using a triplex assay specific for *L. pneumophila* (targeting the *mip* gene) and *L. pneumophila* serogroup 1 (targeting the *wzm* gene). The genotype of each *L. pneumophila* isolate was determined using the M13 modification of EWGLI standard SBT method where part of seven target genes, comprising *flaA*, *pilE*, *asd*, *mip*, *momp*, *proA* and *neuA*, were amplified by PCR and sequenced (Gaia *et al*., 2005, Ratzow *et al*., 2007, Mentasti *et al*., 2014). All allele designations were confirmed and a sequence type (ST) determined using the sequence quality tool (http://www.hpa-bioinformatics.org.uk/*Legionella*/*Legionella*_sbt/php/sbt_homepage.php). In some cases, as indicated below, the sequence type was determined using a direct nested SBT approach on sputum DNA extracts. Monoclonal Antibody typing was carried out using the Dresden panel (Lück *et al*., 2013) as indicated. All steps above were carried out by RVPBRU as part of routine reference service work. Phenol-Chloroform extraction was carried out by myself (Sharon Carney) at the Genomic Medicine Section, NHLI, Imperial College London on available residual respiratory specimens and isolate material. Sputum specimens were transferred to lysis matrix tubes containing CTAB lysis buffer. Samples were then bead beaten and nucleic acid was extracted by the Phenol-Chloroform method. Nucleic acid was precipitated and purified. Extraction and purification steps are described in detail in Chapter 2, Section 2.5. DNA concentration of all extracts was measured by PicoGreen assay as described in Chapter 2, Section 2.6.1.

Tests and extraction methods carried out for each sample are summarised in **Table 6.1** for Case Study 1 samples and **Table 6.2** for Case Study 2 samples.

Table 6.1. Case Study 1: Sample types available and tests/extraction methods performed.

Patient ID	Sample Type	DNA Extraction Method	qPCR	UAT	Culture	SBT	MAb Typing
Patient1	DNA Extract	MagnaPure Compact	Yes	Yes	Yes	Yes	Yes
	Isolate	Phenol-Chloroform	NA		NA	No	
Patient2	DNA Extract	MagnaPure Compact	Yes	Yes	Yes	Yes	Yes
	Isolate	Phenol-Chloroform	NA		NA	No	
Patient3	Sputum Specimen	MagnaPure Compact	Yes	Yes	Yes	Yes	No
	DNA Extract	Phenol-Chloroform	No		No	No	
Patient4	Sputum Specimen	MagnaPure Compact	Yes	Yes	Yes	No	No
	DNA Extract	Phenol-Chloroform	No		No	No	
Patient5	Sputum Specimen	MagnaPure Compact	Yes	Yes	Yes	Yes	Yes
	DNA Extract	Phenol-Choroform	No		No	No	
Patient6	Sputum Specimen	MagnaPure Compact	Yes	Yes	Yes	Yes	No
	DNA Extract	Phenol-Chloroform	No		No	No	

UAT = urinary antigen testing

SBT = sequence-based typing

MAb = monoclonal antibody

NA = not applicable

Table 6.2. Case Study 2: Sample types available and tests/extraction methods performed.

ID	Sample Type	DNA Extraction Method	qPCR	UAT	Culture	SBT	MAb Typing
Patient1	DNA Extract	MagnaPure Compact	Yes	Yes	Yes	Yes	Unknown
Patient2	DNA Extract	MagnaPure Compact	Yes	Yes	Yes	Yes	Yes
	Isolate	Phenol-Choroform	NA		NA	Yes	
Patient3	DNA Extract	MagnaPure Compact	Yes	Yes	Yes	Yes	Yes
	Isolate	Phenol-Chloroform	NA		NA	Yes	
Patient4	DNA Extract	MagnaPure Compact	Yes	Yes	Yes	Yes	No
	Isolate DNA Extract	MagnaPure Compact	NA		NA	No	
Patient5	DNA Extract	MagnaPure Compact	Yes	Yes	Yes	Yes	No
Environ1	Isolate	Phenol-Chloroform	NA	NA	NA	Yes	Yes
Environ2	Isolate	Phenol-Chloroform	NA	NA	NA	Yes	Yes

UAT = urinary antigen testing

SBT = sequence-based typing

MAb = monoclonal antibody

NA = not applicable

6.3.3 Clinical and Epidemiological Data

Clinical and epidemiological data associated with Case Study 1 and 2 were accessed from PHE. Patient data included age group of individuals, sex, details regarding hospitalisation and clinical comments on referral. Epidemiological data included the number of days between symptom onset and sample collection.

6.3.4 Bacterial Community Profiling by 16S rRNA gene Sequencing and Data Analysis

Bacterial community profiling was carried out on all available sputum DNA extracts (from both extraction methods) by 16S rRNA gene amplification and sequencing as described in Chapter 2, Section 2.15.1. Sample pooling and contamination checks were carried out as described in Chapter 2, Section 2.15.2. Sample purification, DNA quantification and equimolar library pooling were carried out as described in Chapter 2, Section 2.15.3. The integrity of the pooled library was assessed using a Bioanalyzer High Sensitivity DNA chip (Chapter 2, Section 2.8.1) and the pooled library was quantified by qPCR and prepared for sequencing (Chapter 2, Section 2.15.4). Data from the 16S rRNA gene data was cleaned and processed using `QIIME` (Version 1.9.1) (Caporaso *et al.,* 2010). The final data was then imported into `R` (Version 3.4.2) and analysed using `phyloseq` (Version 1.22.3) (McMurdie and Holmes, 2013). The full code for data processing and analysis is available in Appendix Section 9.1.21.

6.3.5 Target Capture for *Legionella pneumophila* and Sequencing

Database preparation and bait design were carried out as described in Chapter 2, Sections 2.11.1 and 2.11.2. Library preparation, hybridisation capture (Chapter 2, Section 2.11.3) were carried out on the sputum DNA extracts from Case Study 1 and 2 and a dilution series and mock community (described below in Section 6.3.6). Post-capture prepared libraries were sequenced on one lane of an Illumina 4500 HiSeq (2 x 150 bps) by the Imperial College London BRC Genomics Facility.

6.3.6 Preparation of Dilution Series and Mock Community for Target Capture

The following genomic material was used in the preparation of the dilution series and mock community for target capture: human genomic DNA, *L. pneumophila* Philadelphia-1, *L. pneumophila* France 5811, *L. longbeachae*, *L. anisa, L. feelei, L. micdadei, L. cherii*, *S. pneumoniae*, *H. influenzae* and *V. dispar*. The strain designations, sources and ethical considerations regarding the genomic material are detailed in Chapter 2, Section 2.4. The dilution series contained human genomic DNA spiked with a defined copy number of *L. pneumophila* Philadelphia-1 as outlined in **Table 6.3** below. The composition of the mock community is detailed in **Table 6.4**.

Table 6.3. Composition of the Dilution Series.

Dilution ID	Composition	*L. pneumophila* Copy Number
D1	Human genomic DNA *L. pneumophila* Philadelphia-1 (ST36)	5 X 10^4
D2		1 X 10^4
D3		5 X 10^3
D4		1 X 10^3

Table 6.4. Composition of the Mock Community.

Mock Community Components	Composition (%)
Human Genomic DNA	80
L. pneumophila Phil-1 (ST36)	2
L. pneumophila OLDA (ST1)	2
L. longbeachae	2
L. anisa	2
L. feelei	2
L. micdadei	2
L. cherii	2
S. pneumoniae	2
H. influenzae	2
V. dispar	2

6.3.7 Oxford Nanopore Library Preparation and Sequencing

Oxford Nanopore Libraries were prepared for direct metagenomic sequencing of 3 sputum DNA extracts from Case Study 1 (Patients 1, 2 and 5) and two sputum DNA extracts from Case Study 2 (Patients 2 and 3). Libraries were prepared following the 1D Low Input Genomic DNA with PCR protocol from ONT with modifications. Each library was prepared individually and sequenced on one flow cell for up to 46 hours. The input quantity of DNA for library preparation for Case Study 1 was 892.4 ng for Patient1, 492.2 ng for Patient2, and 547.4 ng for Patient5 and for Case Study 2 was 324 ng for Patient2 and 194ng for Patient3. For each sample, FFPE DNA repair was carried out as described in Chapter 2, Section 2.14.2. End-repair was carried out as described in Chapter 2, Section 2.14.3. PCR adapter ligation was carried out as described in Chapter 2, Section 2.14.4.

An amplification step for long fragments was carried on 20 ng of each PCR adapter-ligated library in duplicate. Each reaction was setup as follows: 46 µl nuclease-free water, 2 µl PCR primers (SQK-LSK108 kit), 2 µl of adapter ligated template (10 ng/µl), 50 µl LongAmp Taq 2x Master Mix (NEB). The PCR tube was mixed gently and centrifuged briefly. Amplification was carried out using the following cycling conditions: 95 °C for 3 minutes followed by 18 cycles of 95 °C for 15 seconds, 62 °C for 15 seconds, 65 °C for 8 minutes. A final extension was carried out at 65 °C for 8 minutes and held at 10 °C. The amplified reaction was purified as described in Chapter 2, Section 2.14.1. A post-PCR end-repair and sequencing adapter ligation step was carried out as described in Chapter 2, Section 2.14.4 and the final library was kept on ice prior to loading.

Flow cell priming and library loading was carried out as described in Chapter 2, Section 2.14.5. Base calling was carried out using `Guppy` (Version 1.8.5) with a q-score filter setting of > 7 (see Appendix Section 9.1.19 for full code). Human DNA reads were removed using minimap2 (Version 2.14) (Li *et al.,* 2018). (see Appendix Section 9.1.20 for full code)

6.3.8 Data Analysis

6.3.8.1 Cleaning and Quality Control of Target Capture Data

All data cleaning and quality control steps (adapter trimming for removal of the Illumina Universal adapter sequence, quality filtering, PhiX removal and human DNA removal) were carried out as described Chapter 2, Sections 2.12.2 to 2.12.5 inclusive.

6.3.8.2 Sequence Alignment with *Legionella* Reference Genomes

Sequenced reads from the captured data were aligned against a completed reference sequence (*L. pneumophila* Lorraine strain (ST47) for Case Study 1 (https://www.ncbi.nlm.nih.gov/nuccore/NC_018139.1) and *L. pneumophila* ST37 (https://www.ncbi.nlm.nih.gov/nuccore/NZ_LT632616) for Case Study 2. Sequence alignment was carried out using `Bowtie2` (Version 2.3.2) (Langmead *et al.,* 2012) with default sensitivity parameters and the `--no-unal` option to suppress SAM records for reads that failed to align. The SAM file was converted to BAM and sorted and duplicates marked and removed with all steps using `picard` (Version 2.12.1) (Picard Toolkit, 2019). Metrics for mean depth of coverage, percentage of reads mapping to the reference sequence and total percentage of reference genome covered were generated for samples before and after duplicate removal using the `pileup` script from `BBTools` (Version 37.38) (Bushnell, 2014). Appendix Section 9.1.12 details the alignment analysis code.

6.3.8.3 *In silico* Sequence-Based Typing for *L. pneumophila*

In silico sequence-based typing (SBT) analysis based on the traditional ESGLI *L. pneumophila* scheme (Gaia *et al.,* 2005, Ratzow *et al.,* 2007, Mentasti *et al.,* 2014) was carried out using the ESGLI database (http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php) and `SRST2` (Inouye *et al.,* 2014) on the captured data to determine if a partial or full sequence type could be generated. This was also carried out on the Oxford Nanopore data from case study samples, as indicated. The analysis method implemented is detailed in Chapter 2, Section 2.13.2.

6.3.8.4 Identification of 50- and 1,455-Core MLST Genes

An investigation of gene presence/absence of 50 core genes pertinent to the multi-locus sequence-based typing scheme, as defined by David *et al.,* 2016(b), was performed. Captured data was aligned to a gene database of 50 core genes of *L. pneumophila* Philadelphia-1 reference genome (https://www.ncbi.nlm.nih.gov/genome/416?genome_assembly_id=300116). Sample reads were mapped to the gene database using SRST2 (Version 0.2.0) (Inouye *et al.,* 2014). The presence or absence of the 50 genes was reported. Additionally, reads from target capture and isolate data were aligned to the extended MLST scheme gene database of 1,455 core genes (again as defined by David *et al*., 2016[b]) from *L. pneumophila* Philadelphia-1 using the same method. This was also carried out on Oxford Nanopore data from case study samples, as indicated. Appendix Section 9.1.13 details the full code used.

6.3.8.5 Phylogenetic Analysis

SNPs from the 1,455 core MLST genes were extracted from target capture and isolate data using `snippy` (Version 4.3.2) (Seeman, 2014) and default parameters. Core SNP alignments were performed using `snippy-core` (Version 4.3.2) (Seeman, 2014). `RaXML-NG` (Kozlov *et al.,* 2019) was used to perform maximum-likelihood searches on the core SNP alignments using the GTRGAMMA model with ascertainment bias correction and 1,000 bootstrap inferences. For each alignment, the best scoring maximum-likelihood tree with bootstrap support values was written to file (please see Appendix Section 9.1.17 for the full code used). Maximum likelihood trees were visualised using `FigTree` (Version 1.4.4) (Rambaut, 2008) and annotated using Microsoft PowerPoint. Due to the lack of SNP calls in samples with low depth of coverage, low reference base coverage and lack of overlapping captured regions, five phylogenetic trees (detailed in **Table 6.5)** were generated based on SNPs from core 1,455 genes:

Table 6.5. Phylogenetic Trees for Case Study 1: Samples and Core Gene Compositions.

Phylogenetic Tree	Samples	No. Core Genes*
1	Patient1_TC, Patient1_ISO, Patient2_TC, Patient2_ISO, Patient5_TC	122
2	Patient1_ISO, Patient2_ISO, Patient3_TC, Patient5_TC	25 partial genes
3	Patient1_ISO, Patient2_ISO, Patient4_TC, Patient5_TC	135 partial genes
4	Patient1_ISO, Patient2_ISO, Patient5_TC, Patient6_TC	31 partial genes
5	Patient1_ISO, Patient2_ISO, Patient4_TC, Patient5_TC, Patient6_TC	2 partial genes

*- please see Appendix Section 9.8 for core gene lists.

6.3.8.6 Heterozygous SNP Analysis for Mixed Strain Detection

Heterozygous SNPs from the 1,455 core genes of target capture and isolate data were extracted from VCF files containing all SNPs calls using `bcftools` (Version 1.9) (Li *et al.*, 2011). Filters were applied to remove heterozygous SNP calls below a Phred quality score of 100. Heterozygous allele depth calculations were extracted from the filtered VCF file and sorted into major and minor allele calls. A further filter was applied to remove any heterozygous allele calls where the minor allele was supported by less than 5 reads. Proportions of major and minor allele calls were calculated. Code for this analysis is provided in Appendix Section 9.1.18.

6.3.8.7 Genome Assembly

Isolate assembly was carried out using `Unicycler` (Version 0.4.7) (Wick *et al.*, 2017[b]) with default parameters. Metagenome assembly was carried out on target capture reads using `metaSPAdes` (Version 3.10.1) (Nurk *et al.*, 2017) without error correction and integrating assemblies spanning from a k-mer size of 27 to 127 base pairs. A co-assembly was performed with Oxford Nanopore reads using the `–nanopore` parameter where indicated. Please see Appendix Section 9.1.14 for the full code used. Taxonomic classification of metagenomic assemblies was carried out using `Centrifuge` (Version 1.0.3) (Kim *et al.*, 2016) as described in Chapter 2, Section 2.13.1. Assemblies were decontaminated by extracting contigs belonging to the order Legionellalaes only using a custom shell script (please see Appendix Section 9.1.15). All assemblies were investigated using `CheckM` (Version 1.0.8) (Parks *et al.*, 2015) with default parameters for genome completeness and evidence of further contamination.

6.3.8.8 Pangenome Visualisation

Clusters of Orthologous Groups of proteins (COGs) were assigned to genome assemblies for Case Studies 1 and 2. This was carried out using Anv'io (Version 5) (Eren *et al.*, 2015) and the COG database (Tatusov *et al.*, 2000). A pangenome visualisation of assemblies based on gene cluster presence/absence was then generated using Anv'io (Eren *et al.*, 2015).

6.4 Results

6.4.1 Dilutions and Mock Community Results

6.4.1.1 Dilutions and Mock: *L. pneumophila* Target Capture

Before examination of the LD case studies, a number of validation steps were carried out based on mock material. In Chapter 5, target capture data from mock material containing known copy numbers (10^6, 10^5, 10^4, 10^3 copies) of a single *L. pneumophila* strain were analysed to establish a cut-off for reliable allele number determination and depth of coverage for downstream analyses. The cut-off was determined at 10^4 genome copies.

In the current chapter, mock material containing the same *L. pneumophila* strain (Philadelphia-1 [ST36]) was prepared to contain genome copy numbers of 4 x 10^4, 8 x 10^3, 4 x 10^3 and 8 x 10^2 for target capture and sequencing. The aim was to determine if a more specific cut-off could be established between 10^4 and 10^3 genome copies. Reads from dilutions and mock community post-capture libraries were mapped to a completed *L. pneumophila* Philadelphia-1 (ST36) genome (https://www.ncbi.nlm.nih.gov/genome/416?genome_assembly_id=300116). The percentage of reference bases covered, the percentage of on-target reads, the mean depth of coverage before and after read duplicate removal were all investigated. Dilutions contained 41,667 (Dilution1), 8,333 (Dilution2), 4,167 (Dilution3) and 833 (Dilution4) genome copies. For dilution tests, base coverage of the *L. pneumophila* reference genome varied from 100 % for Dilution1 to 80 % for Dilution4. Percentage of reads mapped to the reference genome varied from 80 % for Dilution1 to 4 % for Dilution4. Mean depth of coverage after duplicate removal varied from 269 times for Dilution1 to 7 times for Dilution4. Read duplication levels varied from 64 % for Dilution1 to 86 % for Dilution4. For the mock community, 100 % of the reference genome bases were covered. A total of 85 % of reads mapped to the reference genome. The mean depth of coverage after duplicate removal was 1,447 times and levels of duplication were 23 % (**Table 6.6**).

Table 6.6. *L. pneumophila* Target Capture Statistics for Dilutions and Mock Data.

Genome	Genome Copy Number Input	Reference Genome covered (%)	Reads mapped to Reference (number of pairs) (%)	Mean Depth of Coverage (SD) BEFORE Duplicate Removal	Mean Depth of Coverage (SD) AFTER Duplicate Removal	Read Duplication Levels (%)
Dilution1	41,667	100	9,203,485 (80)	761 (1,648)	269 (486)	64
Dilution2	8,333	99	3,066,576 (21)	254 (545)	55 (103)	77
Dilution3	4,167	99	3,213,480 (24)	266 (626)	43 (101)	83
Dilution4	833	80	642,427 (4)	53 (134)	7 (18)	86
Mock Community	833,333	100	22,891,107 (85)	1,884 (2,408)	1,447 (1,842)	23

SD = Standard Deviation

6.4.1.2 Dilutions and Mock: *In silico* Sequence-Based Typing

An *in silico* sequence-based typing analysis, analogous to MLST, based on the traditional ESGLI 7-loci scheme was carried out on the QC'ed sequencing reads of each sample. The aim was to determine if a *L. pneumophila* sequence type could be obtained from the dilution and mock captured data. **Table 6.7** shows the sequence type results obtained.

For dilution samples Dilution1, 2, and 3, all 7 alleles were captured and sequenced. For Dilution4 three alleles (*pilE, mompS* and *proA*) were determined. For the mock community, all alleles were determined, however, *mompS* and *proA* were determined as allele numbers 82 and 5, respectively. This is most likely due to the mixture of *L. pneumophila* sequence types present in the tested community.

Table 6.7. *In silico* Traditional Sequence-Based Typing for Dilutions and Mock Data.

Genome	*flaA*	*pilE*	*asd*	*mip*	*mompS*	*proA*	*neuA/h*	ST
Dilution1	3	4	1	1	14	9	1	36
Dilution2	3	4	1	1	14	9	1	36
Dilution3	3	4	1	1	14	9	1	36
Dilution4	0	4	0	0	14	9	0	-
Mock Community	3	4	1	1	82	5	1	-

6.4.1.3 Dilutions and Mock: Identification of 50- and 1,455-Core MLST Genes

A gene presence/absence analysis was carried to determine the typeability of genes from the dilutions and mock target capture data based on the 50-gene and the extended 1,455-gene MLST scheme proposed by David *et al.*, 2016(b). Samples were mapped to 50- and 1,455-gene databases based on the Philadelphia-1 reference sequence and the presence of genes with >= 90 % coverage only were reported (**Table 6.8**). From the 50-core gene scheme, all 50 genes were typeable for Dilution1, 2, and 3 and thirteen genes (out of 50) were typeable for Dilution4. In the case of the Mock sample, all 50 genes were typeable. From the 1,455-core gene scheme, all 1,455 genes were typeable for Dilution1 and the Mock Community. For Dilution2, 3 and 4, 1,453, 1,451 and 453 genes were typeable, respectively.

Table 6.8 Typeability of 50- and 1,455-core MLST genes for Dilutions and Mock Data.

Target Capture Genome	50 core genes	1,455 core genes
Dilution1	50	1,455
Dilution2	50	1,453
Dilution3	50	1,451
Dilution4	13	453
Mock Community	50	1,455

6.4.1.4 Dilutions and Mock: Heterozygous SNP Analysis for Mixed Strain Detection

Heterozygous SNPs from 1,455 core genes of single *L. pneumophila* strain and mixed strain controls were analysed as a means of investigating mixed infection patterns. The total number of heterozygous SNPs was determined after quality and read depth filtering. This was then compared against the total number of SNPs (homozygous and heterozygous) and the proportion of heterozygous SNPs was calculated (**Table 6.9**). From single strain controls (Dilution tests), the heterozygous SNPs proportion varied from 1 to 3 % of total SNPs (**Figure 6.1(a)**, **(b)** and **(c)**). In the mixed strain control (Mock Community) the initial mixture of which contains two *L. pneumophila* sequence types, the heterozygous SNPs proportion was 98 % of total SNPs. When plotted, the mixed strain control demonstrated a clustering pattern between major and minor alleles from heterozygous SNP calls (**Figure 6.1(d)**).

Table 6.9 Heterozygous SNPs in Dilution Tests and Mock Community.

ID	No. Het SNPs	No. Total SNPs	Proportion Het SNPs
Dilution1	5	484	1 %
Dilution2	12	401	3 %
Dilution3	5	371	1 %
Dilution4	1	49	2 %
Mock Community	20,280	20,762	98 %

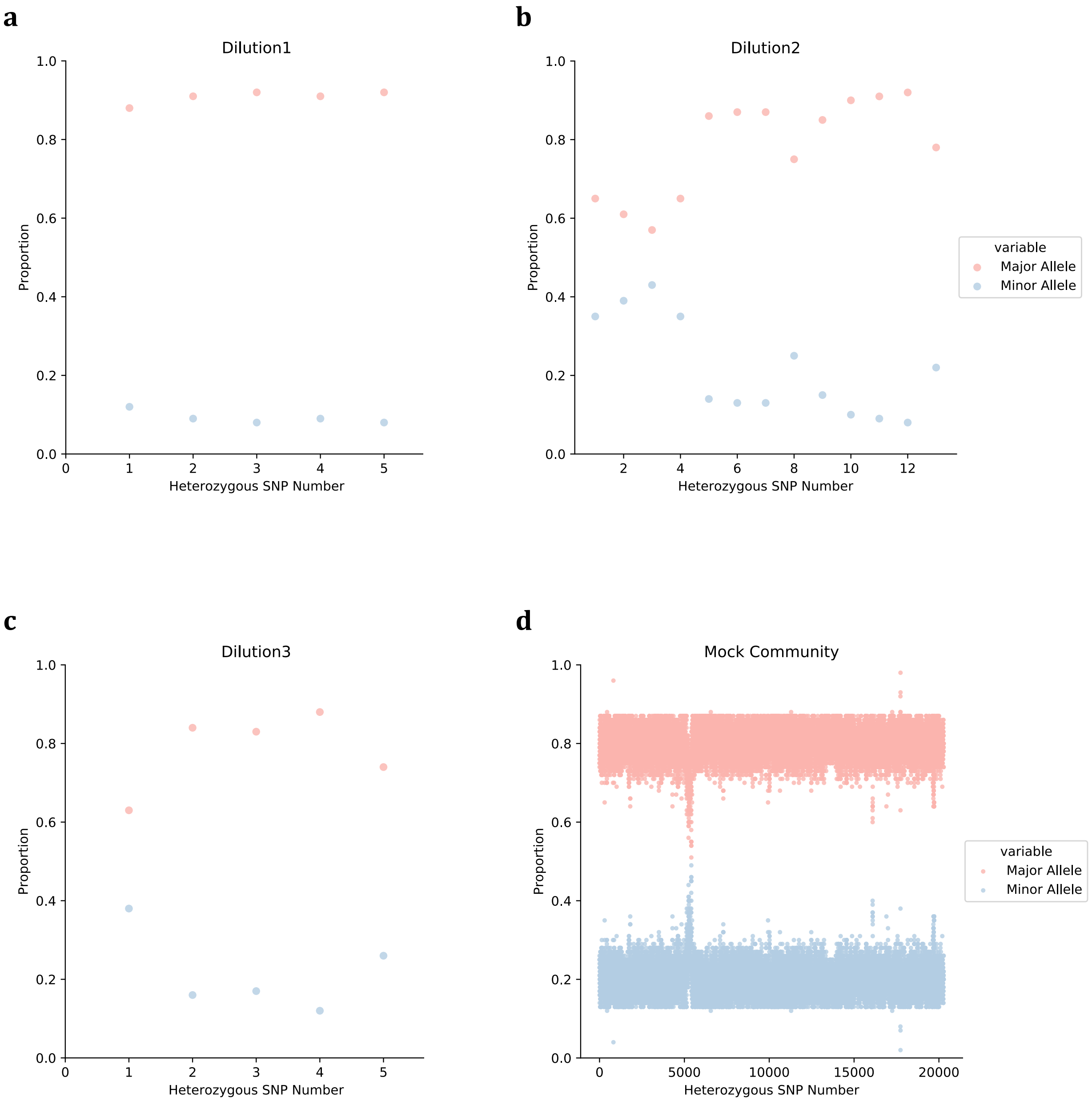


Figure 6.1. Heterozygous SNP proportions in the single strain controls **(a)** Dilution1, **(b)** Dilution2 and **(c)** Dilution3 and **(d)** the Mock Community containing mixed *L. pneumophila* populations. From single strain controls (**a, b** and **c**), heterozygous SNPs proportion varied from 1 to 3 % of total SNPs. In the mixed strain control (**d**) the initial mixture of which contains two *L. pneumophila* sequence types, the heterozygous SNPs proportion was 98 % of total SNPs.

In conclusion, for Dilution1 (4 x 10^4 genome copies), reliable typing data and a high depth of coverage was obtained. Contrastingly, for Dilution4 (8 x 10^2), a partial sequence type was obtained, and genome depth of coverage was low. These results corroborate findings from Chapter 5 (Section 5.5) based on high and low genome copy number input.

Upon analysis of Dilution 2 (8 x 10^3) and 3 (4 x 10^3), all allele numbers were obtained, providing a full 7-loci sequence type in both cases, as well as a high depth of coverage on average. Additionally, all 50 core MLST genes were typeable and from the 1,455 core MLST scheme, 1,455 and 1,453 genes were typeable for Dilution2 and Dilution3, respectively. Data from the dilutions and mock samples demonstrated that the analysis of heterozygous SNPs in 1,455 core genome regions from single *L. pneumophila* strain and mixed strain controls can provide indications of mixed infection patterns

6.4.2 Case Study 1 Results

6.4.2.1 Case Study 1: Epidemiological, Clinical and Microbiological Data

For Case Study 1, respiratory specimens were available for 6 patients and matched isolates available for 2 of the 6. There was no clinical sample or isolate available for a 7th patient, however clinical and microbiological data was available from PHE for this case and is included in this Chapter. From the 7 patients, 1 was female and 6 were male, all aged between 40 and 80 years (mean: 64.8 years). The number of days between reported symptom onset were available for 6 out of 7 patients and sample collection ranged from 4 to 24 days (mean: 10.6 days). Three patients experienced respiratory failure and one experienced multi-organ failure. Seven cases reported shortness of breath, four a cough, four experienced confusion, two cases reported chest pain and two diarrhoea. All cases were alive at the close of the investigation (at least 28 days after their respective onset dates). Half of the cases were current smokers and one an ex-smoker. Co-morbidities were reported for a number of the cases and included hypertension, Chronic Obstructive Pulmonary Disease (COPD), musculoskeletal problems (causing low mobility) and long-term steroid use. Isolates (single) were obtained from samples from 3 patients. Urinary antigen testing of Patient4 was positive as reported by NHS local hospital trust however was negative by UAT and qPCR at the PHE *Legionella* reference lab and was not isolated. **Table 6.10** provides the full information about individual patient samples and diagnostic test results that was available for this current study.

Table 6.10 Epidemiological, Clinical and Microbiology Data for Case Study 1

Patient ID	**Sex**	**Age Group (Years)**	**No. days symptom onset to sample collection**	**Further hospital/clinical details**	**UAT**	**qPCR (CT: *mip*, *wzm*)**	**Culture**	**SBT**	**MAb**
Patient1	F	60 – 70	18	ITU, respiratory failure	Positive	Positive (23.6, 24.9)	Isolated	ST47* (*5,10,22,15,6,2,6*)	Not determined
Patient2	M	40 – 50	24	ITU, type1 respiratory failure	Positive	Positive (23.3, 24.5)	Isolated	ST2110* (*6,10,2,10,13,4,9*)	Knoxville
Patient3	M	50 – 60	5	ITU	Positive	Positive (33, 34)	Not Isolated	Partial ST* (*6,0,0,10,13,14,6*)	Not tested
Patient4	M	70 – 80	No info	Orthopaedic ward patient	Positive Hospital test Not confirmed (PHE)	Negative	Not Isolated	No info	No info
Patient5	M	50 – 60	7	ITU, ventilated, lower lobular pneumonia, type 1 respiratory failure	Positive	Positive (25, 26)	Isolated	ST82 (*5,1,22,10,6,10,6*)	Not determined
Patient6	M	70 – 80	4	ITU, COPD	Positive	Positive (36, 37)	Not Isolated	Not obtained	Not tested
Patient7	M	50 – 60	6	ITU, multi-organ failure	Positive	Positive (34, 33)	Not Isolated	Partial ST* (*5,0,0,0,6,2,6*) (ST47/109 PCR positive)	Not tested

*-obtained by direct nested SBT, UAT = urinary antigen testing, SBT = sequence-based typing, MAb = monoclonal antibody, CT = cycle threshold

6.4.2.2 Case Study 1: *Legionella* Abundance in Bacterial Community

The relative abundance of *Legionella* in the bacterial component of Case Study 1 sputum nucleic acid extracts was examined by sequencing the 16S rRNA gene and comparing the proportion of sequenced *Legionella* operational taxonomic units (OTUs) to the total number of bacterial OTUs in each sample, rarefied to 2,000 OTUs. In addition to the patient samples, DNA from mock communities were sequenced - Mock1 representing the in-house positive sequencing control and Mock2 representing the *Legionella* positive control. The bacterial community for each sequenced sample is visualised in **Figure 6.2**. The relative abundance of *Legionella* in the bacterial communities was: 0.35 % for Patient1, 91.5 % for Patient2, 0.1 % for Patient3a and 0 % for Patient3b (where 'a' signifies extraction by MagnaPure Compact and 'b' signifies extraction by the Phenol-Chloroform method), 22.6 % for Patient5a and 30 % for Patient5b, 0 % for Patient6a and 6b, 0 % for the in-house *Legionella*-negative Mock1 and 16.3 % for the *Legionella*-positive Mock2. Samples for Patient4a and 4b did not contain sufficient reads to include in the analysis and as expected, negative controls for extraction and sequencing did not contain sufficient reads to include in the visualisation.

Communities for Patient3 and Patient6 were dominated by *S. pneumoniae* and the community for Patient1 was dominated by *H. influenzae*.

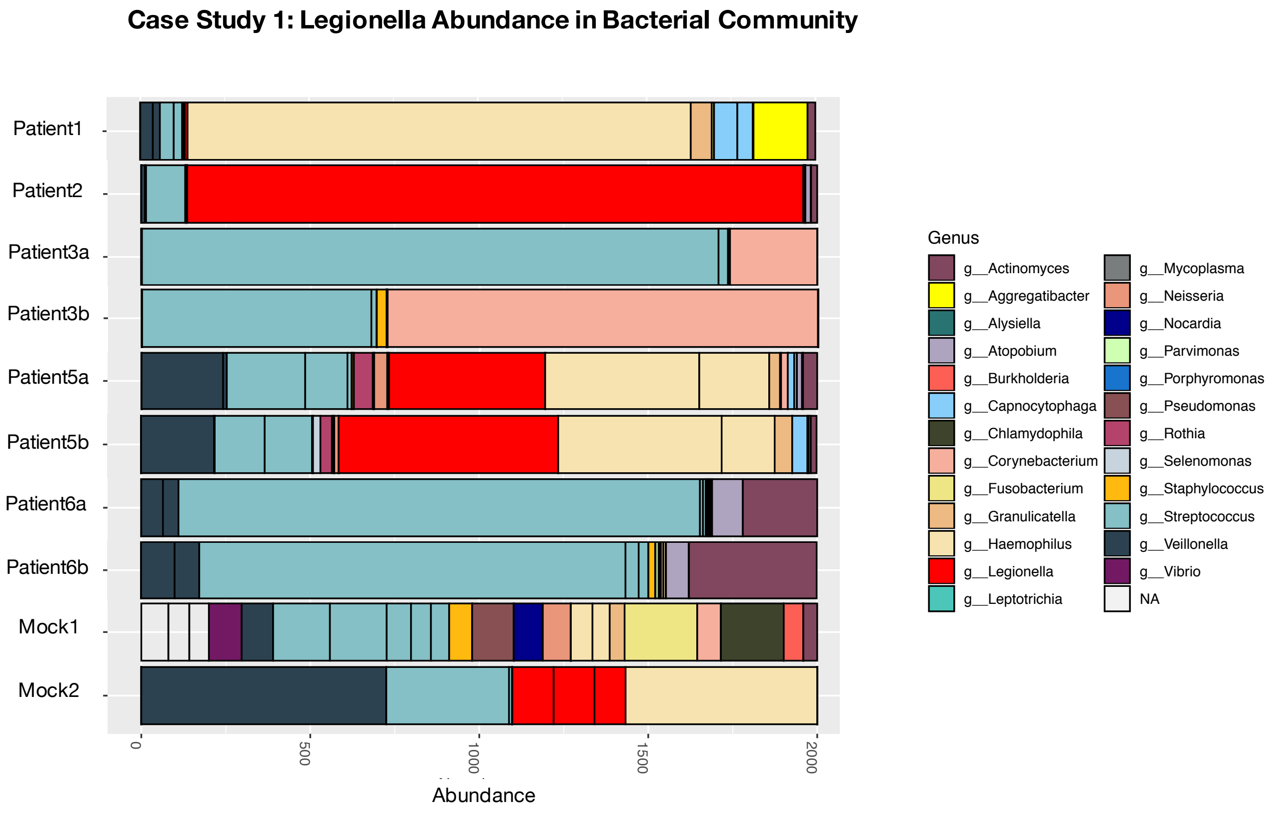


Figure 6.2 Bacterial communities demonstrating *Legionella* genus OTU abundance for Case Study 1 samples.

6.4.2.3 Case Study 1: *L. pneumophila* Target Capture

Reads from Case Study 1 post-capture libraries were mapped to a completed *L. pneumophila* Lorraine (ST47) genome (https://www.ncbi.nlm.nih.gov/nuccore/NC_018139.1). The percentage of reference bases covered, the percentage of on-target reads, the mean depth of coverage before and after read duplicate removal were all investigated (**Table 6.11**). Base coverage of the *L. pneumophila* reference genome varied from 2 % for Patient4a ('a' signifies the MagnaPure Compact extracted sample) to 96 % for Patient5b ('b' signifies the Phenol-Chloroform extracted sample). The proportion of reads mapping to the *Legionella* reference genome varied from 0.0004 % for Patient4a to 48 % for Patient5b. Mean depth of coverage varied from 0.02 times for Patient4a to 31 times for Patient5b post-duplicate removal. Read duplication levels varied from 9 % to 88 %.

Table 6.11 *L. pneumophila* Target Capture Statistics for Case Study 1.

Genome	Genome Copy Number Input	Reference Genome covered (%)	Reads mapped to Reference (number of pairs) (%)	Mean Depth of Coverage (SD) BEFORE Duplicate Removal	Mean Depth of Coverage (SD) AFTER Duplicate Removal	Read Duplication Levels (%)
Patient1	116,667	70	1,721,390 (17)	136 (1,853)	23 (561)	82
Patient2	83,333	50	288,111 (3)	22 (136)	4 (28)	82
Patient3a	167	3	74,693 (0.5)	4 (194)	0.7 (27)	84
Patient3b	Est. 167	7	50,277 (0.4)	3 (101)	0.5 (16)	83
Patient4a	ND	2	347 (0.0004)	0.02 (0.27)	0.02 (0.2)	9
Patient4b	Est. ND	20	69,045 (0.7)	5 (16)	0.9 (3)	83
Patient5a	41,667	44	73,810 (0.8)	5 (10)	3 (6)	44
Patient5b	Est. 41,667	96	3,316,515 (48)	244 (242)	31 (45)	86
Patient6a	250	4	126,548 (1)	8 (346)	0.8 (32)	88
Patient6b	Est. 250	27	179,742 (2)	14 (78)	2 (13)	86

6.4.2.4 Case Study 1: *In silico* Sequence-Based Typing

An *in silico* sequence-based typing analysis, analogous to MLST and based on the traditional ESGLI 7-loci scheme, was carried out on the QC'ed sequencing reads of each sample. Reads from patient samples sequenced twice (two different extraction methods) were merged at this point. **Table 6.12** shows the sequence type results from the target capture data for Case Study 1. In the case of Patient3 and Patient4, no allele number was determined. A partial sequence type was determined for Patient1 (alleles *0, 0, 0, 15, 6, 0, 6*) and Patient2 (alleles *6, 0, 2, 0, 13, 0, 0*) and one allele (*proA*) was determined for Patient6 (alleles *0, 0, 0, 0, 0, 2, 0*). A full sequence type was determined for Patient5 matching ST82 as previously determined by PHE. Oxford Nanopore reads sequenced directly from Patient1, 2 and 5 samples did not yield a sequence type.

Table 6.12 *In silico* Sequence-Based Typing Result of Case Study 1 Target Capture Data.

Genome	*flaA*	*pilE*	*asd*	*mip*	*mompS*	*proA*	*neuA/h*	ST
Patient1	0	0	0	15	6	0	6	-
Patient2	6	0	2	0	13	0	0	-
Patient3	0	0	0	0	0	0	0	-
Patient4	0	0	0	0	0	0	0	-
Patient5	5	1	22	10	6	10	6	82
Patient6	0	0	0	0	0	2	0	-

6.4.2.5 Case Study 1: Identification of 50- and 1,455-Core MLST Genes

A gene presence/absence analysis was carried to determine the typeability of genes from the target capture data based on the 50-gene and the extended 1,455-gene MLST scheme. Samples were mapped to 50- and 1,455-gene databases based on the Philadelphia-1 reference sequence and the presence of genes with >= 90 % coverage only were reported (**Table 6.13**). From the 50-core gene scheme, 10 genes were sequenced from Patient1 capture data, 3 from Patient2, 48 from Patient5 and 1 from Patient6. No gene was sequenced from the 50-gene scheme for Patient3 or Patient4 capture data. From the 1,455 core gene scheme, 243 genes were sequenced from Patient1 capture data, 107 from Patient2, 2 from

Patient4, 1,356 from Patient5 and 13 from Patient6. No gene was sequenced from the 1,455-gene scheme for Patient3 capture data. Oxford Nanopore reads sequenced directly from Patient1, 2 and 5 samples did not yield full gene sequence information from either 50 or 1,455 core genes.

Table 6.13 Core 50- and 1,455-gene MLST of Case Study 1 Target Capture Data.

Genome	50 core genes	1,455 core genes
Patient1	10	243
Patient2	3	107
Patient3	0	0
Patient4	0	2
Patient5	48	1,356
Patient6	1	13

Genes marked as present if > 90 % bases covered

6.4.2.6 Case Study 1: Phylogenetic Analysis

Due to lack of overlapping genomic regions captured and sequenced from some samples, five maximum-likelihood phylogenetic trees were constructed. This was carried out to examine the clustering of the target capture and isolate data together and to investigate the phylogenetic localisation of target capture data particularly in cases where *L. pneumophila* was not isolated or a sequence type was not obtained.

Tree 1 (**Figure 6.3**) was generated based on an alignment of overlapping regions from 1455 core genes of Patient1 and Patient2 target capture and isolate data, Patient5 target capture data and a reference (*L. pneumophila* Philadelphia-1 [ST36]). In addition to regions of similarity, the alignment contained 280 SNPs from 123 of the core genes (See Appendix Section 9.8 for the list of 123 core genes containing SNPs). The samples formed two clades: one with Patient2 (ST2110) and the other with Patient1 (ST47) and Patient5 (ST82). The clustering of ST47 with ST82 is anticipated as both sequence types share 4 alleles based on 7-loci SBT.

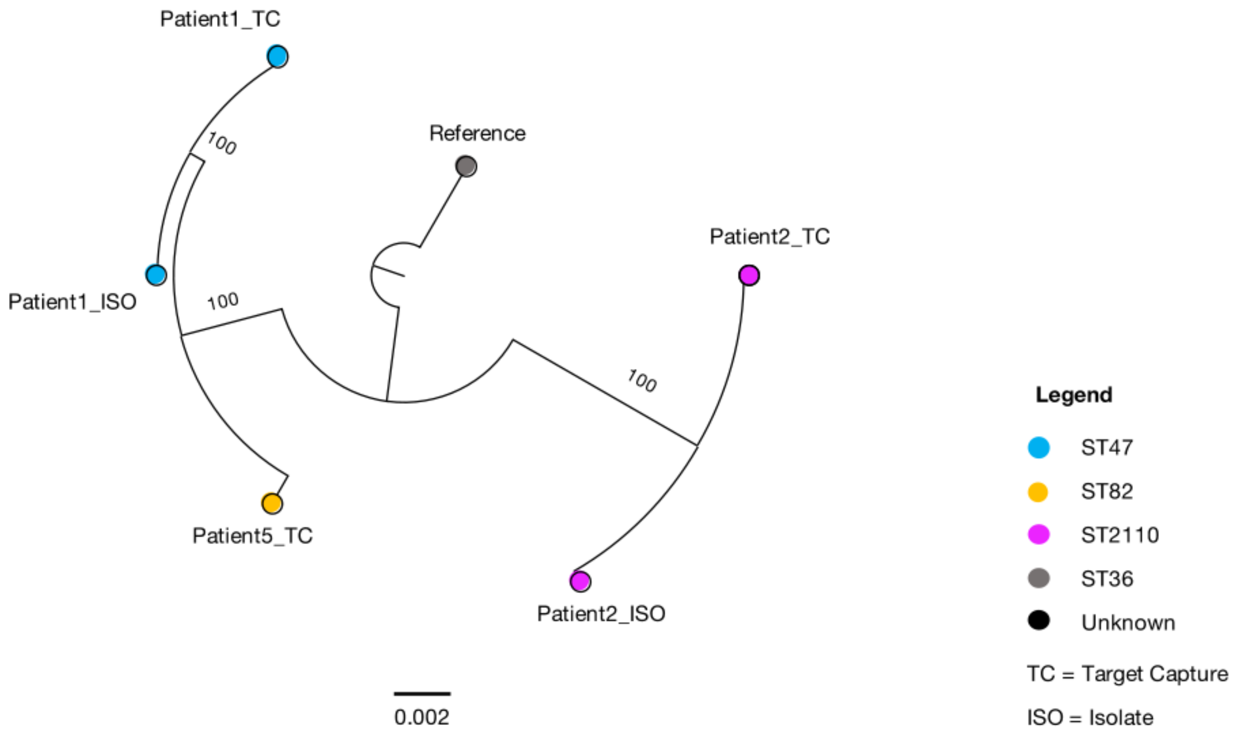


Figure 6.3. Maximum-likelihood Phylogenetic Tree 1. The tree was computed based on an alignment of overlapping regions from 1455 core genes of Patient1 and Patient2 target capture and isolate data, Patient5 target capture data and a reference (*L. pneumophila* Philadelphia-1 [ST36]). In addition to regions of similarity, the alignment contained 280 SNPs from 123 of the core genes. The internal nodes are labelled with the bootstrap values.

Tree 2 (**Figure 6.4**) was constructed based on an alignment of overlapping regions from 1455 core genes from Patient1, 2, 3 and 5 data. In addition to regions of similarity, the alignment contained 76 SNPs from 25 of the core genes (See Appendix Section 9.8 for the list of 25 genes core genes containing SNPs). Target capture data for Patient3 did not cluster with known STs involved in Case Study 1. The same pattern was observed in Tree 3 (which contained 407 SNPs from 135 of the core genes) for Patient4 target capture data (**Figure 6.5**) (See Appendix Section 9.8 for the list of 135 core genes) and Tree 4 (which contained 91 SNPs from 31 of the core genes) for Patient6 target capture data (**Figure 6.6**) which did not cluster with data from Patient1, 2 or 5 (See Appendix Section 9.8 for the list of 31 core genes).

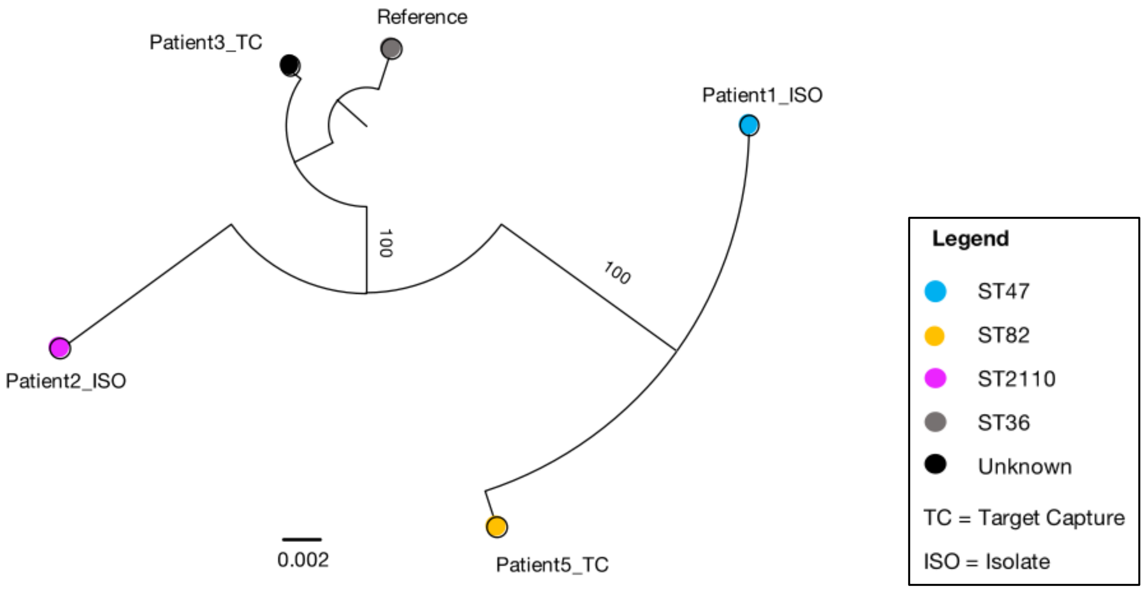


Figure 6.4. Maximum-likelihood Phylogenetic Tree 2. The tree is based on an alignment of overlapping genomic regions from 1455 core genes of Patient1 and Patient2 isolate data, Patient3 and Patient5 target capture data and a reference (*L. pneumophila* Philadelphia-1 [ST36]). In addition to regions of similarity, the alignment contained 76 SNPs from 25 of the core genes. The internal nodes are labelled with the bootstrap values.

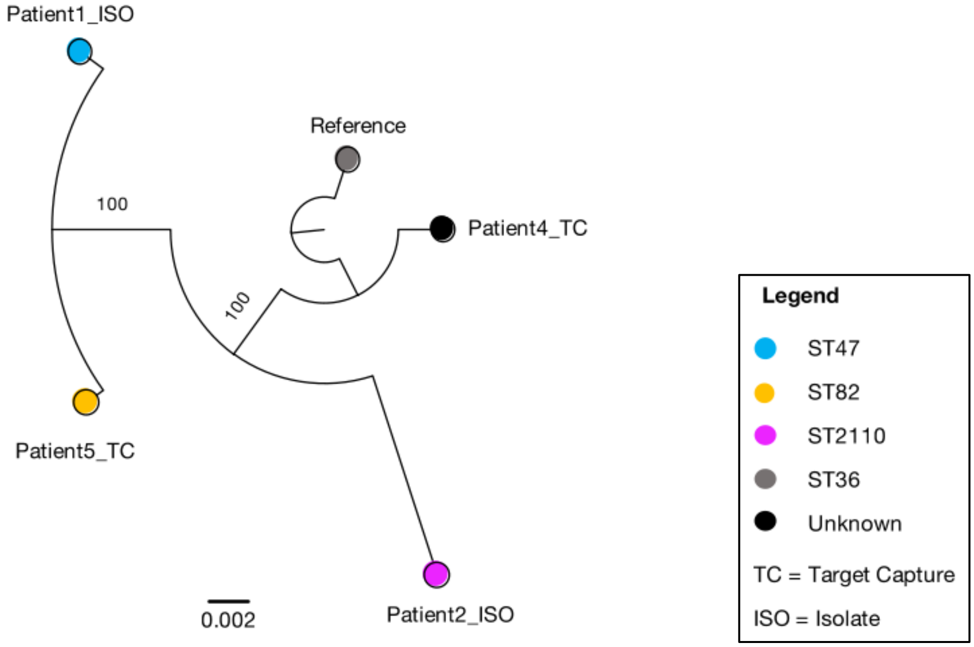


Figure 6.5. Maximum-likelihood Phylogenetic Tree 3. The tree is based on an alignment of overlapping genomic regions from 1455 core genes of Patient1 and Patient2 isolate data, Patient4 and Patient5 target capture data and a reference (*L. pneumophila* Philadelphia-1 [ST36]). In addition to regions of similarity, the alignment contined 407 SNPs from 135 core genes. The internal nodes are labelled with the bootstrap values.

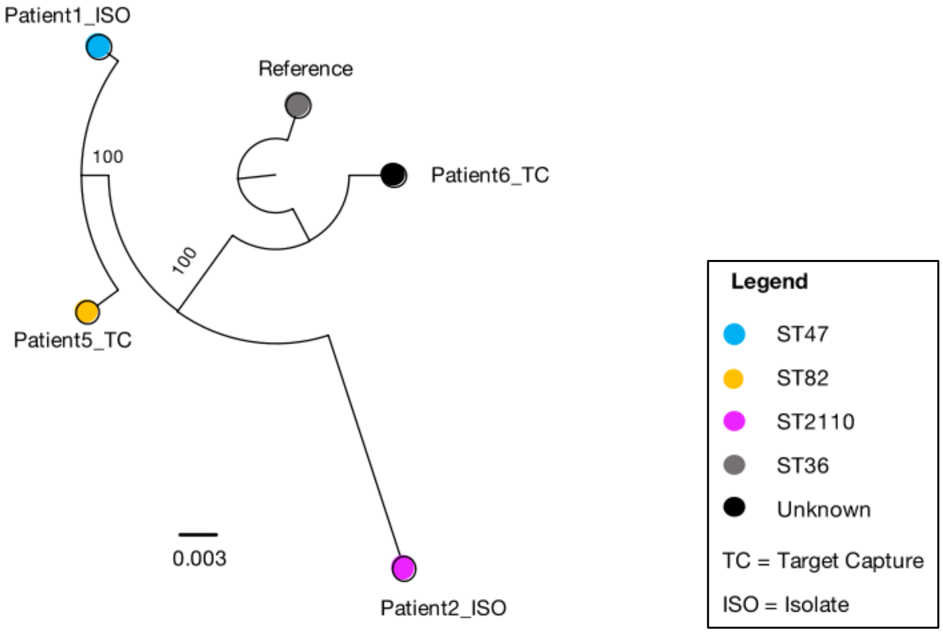


Figure 6.6. Maximum-likelihood Phylogenetic Tree 4. The tree is based on an alignment of overlapping genomic regions from 1455 core genes of Patient1 and Patient2 isolate data, Patient5 and Patient6 target capture data and a reference (*L. pneumophila* Philadelphia-1 [ST36]). In addition to regions of similarity, the alignment contined 91 SNPs from 31 of the core genes. The internal nodes are labelled with the bootstrap values.

A final tree (Tree 5, **Figure 6.7**) based on two partial (70 %) core genes was constructed for five out of six patient samples: Patient1, 2, 4, 5 and 6. (See Appendix Section 9.8) Despite the low resolution of the tree, data for Patient1 and 5 formed the same clustering pattern as in the previous 4 trees. Interestingly, Patient4 and Patient6 clustered together however Patient4 demonstrated a diverging pattern.

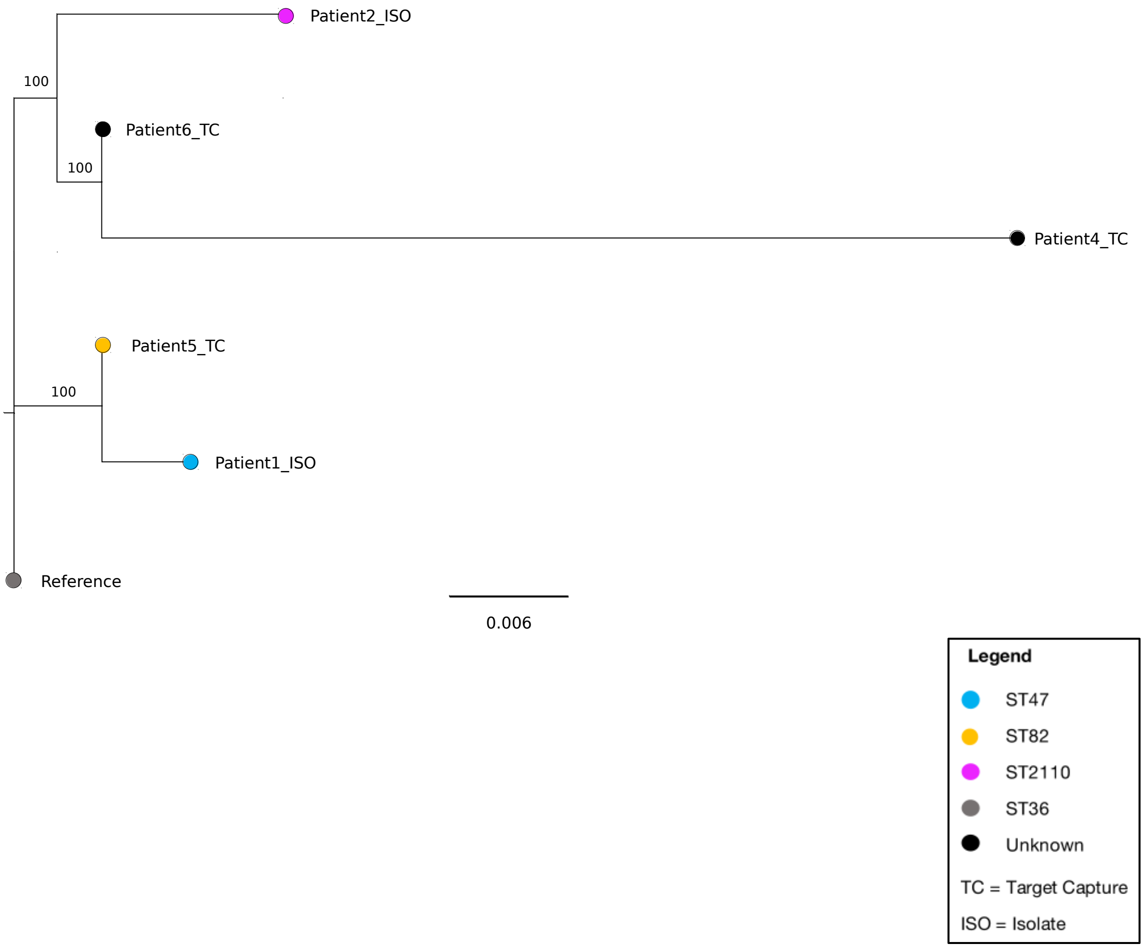


Figure 6.7. Maximum-likelihood Phylogenetic Tree 5. The tree is based on SNPs from 2 partial core genes of Patient1 and Patient2 isolate data, Patient4, Patient5 and Patient6 target capture data and a reference (*L. pneumophila* Philadelphia-1 [ST36]). The internal nodes are labelled with the bootstrap values.

6.4.2.7 Case Study 1: Heterozygous SNP Analysis for Mixed Strain Detection

Heterozygous SNP analysis, as carried out in on mock samples in Section 6.4.1.4., was carried out on target capture and isolate data from Case Study 1 samples. The total number of heterozygous SNPs was determined after quality and read depth filtering. This was then compared against the total number of SNPs (homozygous and heterozygous) and the proportion of heterozygous SNPs was calculated (**Table 6.14**). Heterozygous SNP proportions representing less than 10 % of total SNPs were not further analysed for mixed strains due to increased uncertainty. From Case Study 1 samples, no heterozygous SNPs were present in target capture data from Patient1, 2, 3, 6 or isolate data from Patient2. SNP data from the Patient1 isolate had 0.04 % heterozygous SNPs. Plotting the heterozygous SNP output for the Patient1 isolate (**Figure 6.8(a)**) demonstrated a similar profile to that of single strain controls in Section 6.4.1.4. Target capture data from Patient4 had 4 % heterozygous SNP calls however it was not analysed further due to the sparsity of the data (3 heterozygous SNPs only).

Target capture data from Patient5 provided an interesting pattern. From total SNPs, 13 % were heterozygous. Furthermore, when the data was plotted (**Figure 6.8(b)**), major and minor allele clustering was observed indicating that the sample may indeed contain mixed *L. pneumophila* populations.

Table 6.14 Proportions of Heterozygous SNPs in 1,455 core genes of Case Study 1 samples.

Genome	No. Het SNPs	No. Total SNPs	Proportion Het SNPs
Patient1_TC	0	1,183	0 %
Patient1_ISO	8	19,728	0.04 %
Patient2_TC	0	1,130	0 %
Patient2_ISO	0	22,320	0 %
Patient3_TC	0	9	0 %
Patient4_TC	3	70	4%
Patient5_TC	1,915	15,013	13 %
Patient6_TC	0	26	0 %

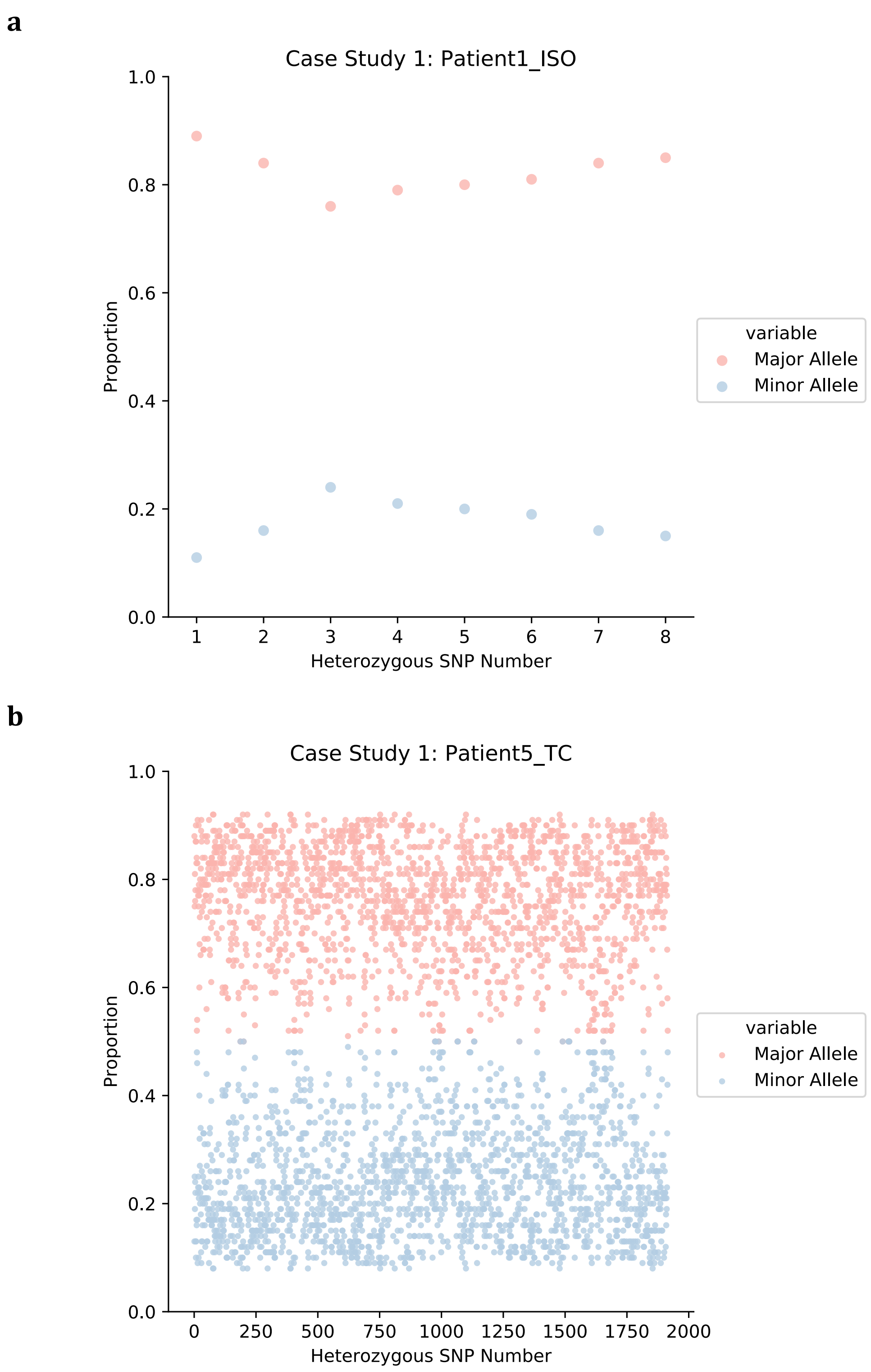


Figure 6.8. Heterozygous SNP proportions (Case Study 1) in **(a)** Patient1 isolate indicating a single strain and **(b)** Patient5 target capture data indicating a mixed *L. pneumophila* population.

6.4.2.8 Case Study 1: Genome Assembly and Pangenome Visualisation

Genomes from the target capture and isolate data were assembled and target capture assemblies were decontaminated by removing contigs belonging to taxonomic orders other than the order Legionellales. A quality analysis was carried out to determine genome length, maximum contig length, N50 (the minimum contig length needed to cover 50 % of the genome), GC content, genome completeness, residual contamination and the number of predicted genes. Genome draft quality (high, medium or low) was assigned based on criteria defined by Bowers *et al.*, 2017. For target capture data, genome length varied from 3,701,560 for Patient5 to 27,534 bases for Patient3. Maximum contig length varied from 16,216 for Patient5 to 732 for Patient3. Estimations of genome completeness based on the presence of single copy genes varied from 98 % for Patient5 to 1.7 % for Patient4. Estimations of contamination levels varied from 0 % for Patient3 and 4 to 13 % for Patient5. Upon closer inspection, it was observed that the majority of contamination for Patient5 was representative of strain heterogeneity (**Table 6.15**). Assembly of isolate data for Patient1 and Patient2 produced two high quality draft assemblies (**Table 6.16**).

As means of viewing genome assemblies within the context of a LD cluster, a pangenome visualisation was created for Case Study 1 target capture and isolate data (**Figure 6.9**). The visualisation was based on the distribution of gene clusters across the genomes. Gene clusters were identified among the genomes and each cluster was annotated with a COG function which was displayed, if known (in green) or unknown (in grey). Additionally, a functional homogeneity plot was incorporated into the visualisation as a means of displaying the level of conservation of aligned amino acid residues across genes. A gene cluster may contain amino acid sequences from different genomes that are almost identical, which would indicate functional homogeneity. Contrastingly, the gene cluster may contain highly divergent amino acid sequences from different genomes which may indicate functional heterogeneity or possibly immune selection.

A total of 4,087 gene clusters were assigned based on 19,102 genes from Case Study 1 genomes. A total of 1330 (26 %) gene clusters had >= 99 % functional homogeneity, 1,590 (38 %) had >= 75 % functional homogeneity and 2,791 (68 %) had >= 50 % functional homogeneity.

Table 6.15 Genome Assembly Quality Statistics: Target Capture Data for Case Study 1.

Genome	Genome Length (Bases)	Max Contig Length (Bases)	N50	GC Content	Completeness	Contamination	No. Predicted Genes	Draft Quality	Coding Density	No. contigs
Patient1*	1,392,561	2,656	417	0.39	43.3 %	3 %	3,720	Low	0.88	3,343
Patient2*	906,117	2,450	366	0.39	25.7 %	1.6 %	2,563	Low	0.84	2,442
Patient3	27,534	732	332	0.41	4.2 %	0 %	82	-	0.88	81
Patient4	49,864	750	335	0.41	1.7 %	0 %	153	-	0.91	144
Patient5*	3,701,560	16,216	2,179	0.39	98 %	13 %	5,394	Medium	0.88	3,087
Patient6	278,399	1,639	351	0.41	7 %	3.5 %	834		0.91	764

*- genome co-assembled with *Legionella* reads from Oxford Nanopore sequencing

Table 6.16 Genome Assembly Quality Statistics: Isolate Data for Case Study 1.

Genome	Genome Length (Bases)	Max Contig Length (Bases)	N50	GC Content	Completeness	Contamination	No. Predicted genes	Draft Quality	Coding Density	No. contigs
Patient1	3,572,548	239,729	81,017	0.38	100	0	3,198	High	0.887	75
Patient2	3,581,525	581,213	241,284	0.38	100	0.585	3,232	High	0.883	35

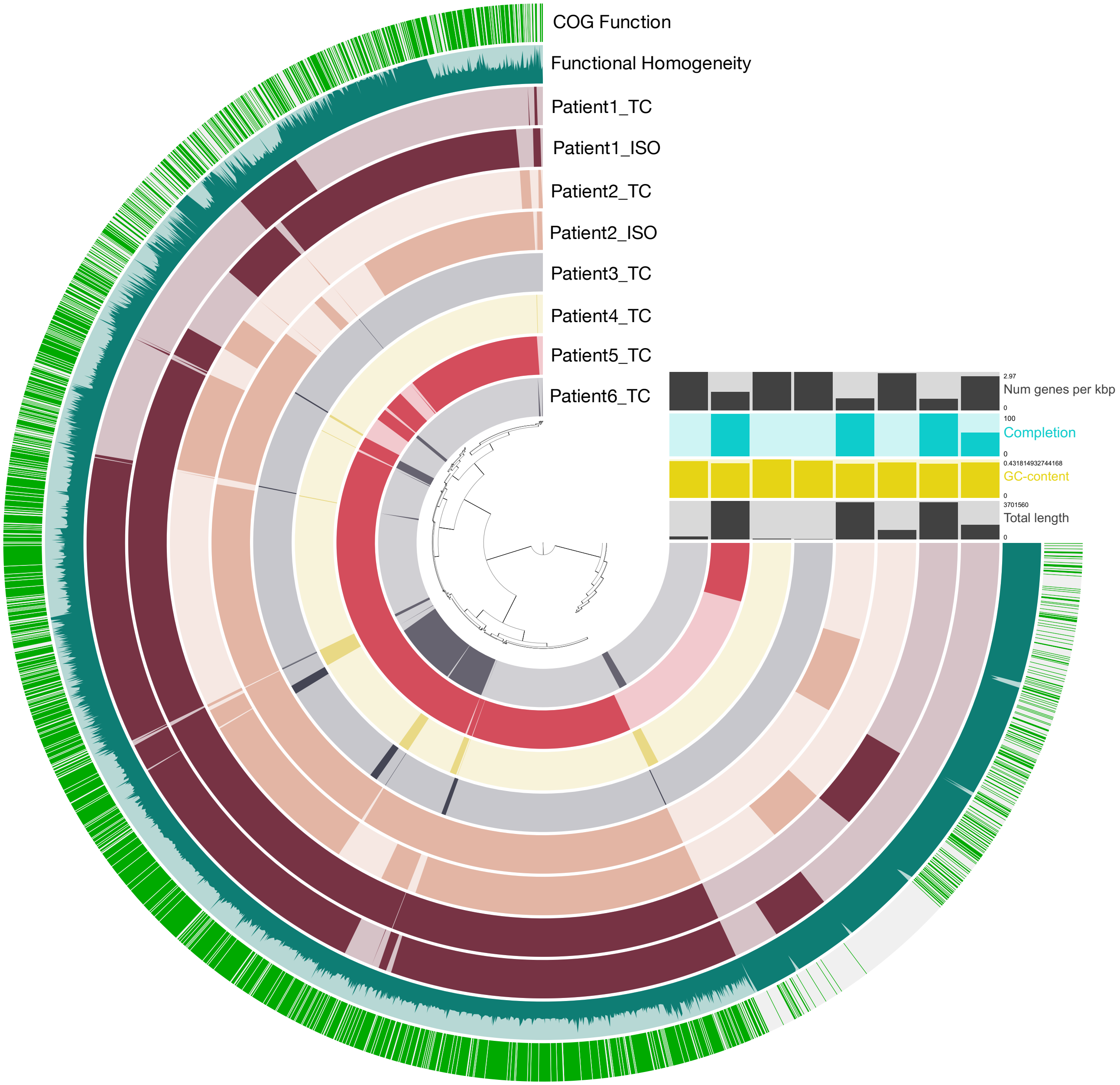


Figure 6.9. Pangenome visualisation of genomes assemblies from Case Study 1. The visualisation displays the gene clusters present in the target capture and isolate genomes, COG function assignment (green for known and grey for unknown), functional homogeneity (homogeneity represented by spikes in green) and layers containing information on number of gene per kbp, GC content, levels of genome completion and total length. TC = target capture and ISO = isolate. Dark shading indicates gene clusters that were sequenced. Light shading indicates gene clusters that are absent from the assembly.

6.4.3 Case Study 2 Results

6.4.3.1 Case Study 2: Epidemiological, Clinical and Microbiological Data

For this study, respiratory specimens were available for 5 patients with isolates available for 2 of the 5. Additionally, environmental isolates were available from two different sources. Of the 5 patients, 2 were female and 3 were male, all aged between 40 and 60 years (mean: 52.4 years). The number of days between reported symptom onset and sample collection ranged from 5 to 9 days (mean: 7.4 days). One patient experienced multi-organ failure. No deaths were reported. An isolate had been obtained for three of the patients. **Table 6.17** provides full information about individual patient and environmental samples and diagnostic test results.

Table 6.17. Epidemiological, Clinical and Microbiological Data for Case Study 2.

Patient/ Environmental ID	Sex	Age Group (Years)	No. days symptom onset to sample collection	Further details	UAT	qPCR (CT: *mip*, *wzm*)	Culture	SBT	MAb Type
Patient1	M	50 - 60	5	ITU, Multi-organ failure	Positive	Positive (31,31)	Isolated	ST37	No info
Patient2	F	40 - 50	8	ITU	Positive	Positive (23, 24)	Isolated	ST37	Philadelphia
Patient3	M	50 - 60	7	ITU	Positive	Positive (27, 28)	Isolated	ST37	Philadelphia
Patient4	F	50 - 60	8	-	Positive	Positive (33, 32)	Not Isolated	ST37*	Not tested
Patient5	M	50 - 60	9	ITU	Positive	Positive (32, 32)	Not Isolated	Partial ST* (3,0,0,1,14,9,11)	No info
Environmental1	NA	NA	NA	Cooling tower infection system?	NA	No info	Isolated	ST37	Philadelphia
Environmental2	NA	NA	NA	Cooling Tower	NA	No info	Isolated	ST37	Philadelphia

UAT = urinary antigen testing

SBT = sequence-based typing

MAb = monoclonal antibody

NA = not applicable

* - indicates direct nested SBT

6.4.3.2 Case Study 2: *Legionella* Abundance in Bacterial Communities

The relative abundance of *Legionella* in the bacterial component of Case Study 2 sputum nucleic acid extracts was examined by sequencing the 16S rRNA gene and comparing the proportion of sequenced *Legionella* OTUs to the total number of bacterial OTUs in each sample, rarefied to 2,000 OTUs. Also included were the mock samples as described in Section 6.4.2.2.

The bacterial community for each sequenced sample is visualised and shown in **Figure 6.10**. The relative abundance of *Legionella* in the bacterial communities was as follows: 99.3 % for Patient2, 0.2 % for Patient3, 0.05 % for Patient4 and 0 % for Patient5. The sample for Patient1 did not contain sufficient reads to allow its inclusion in the analysis.

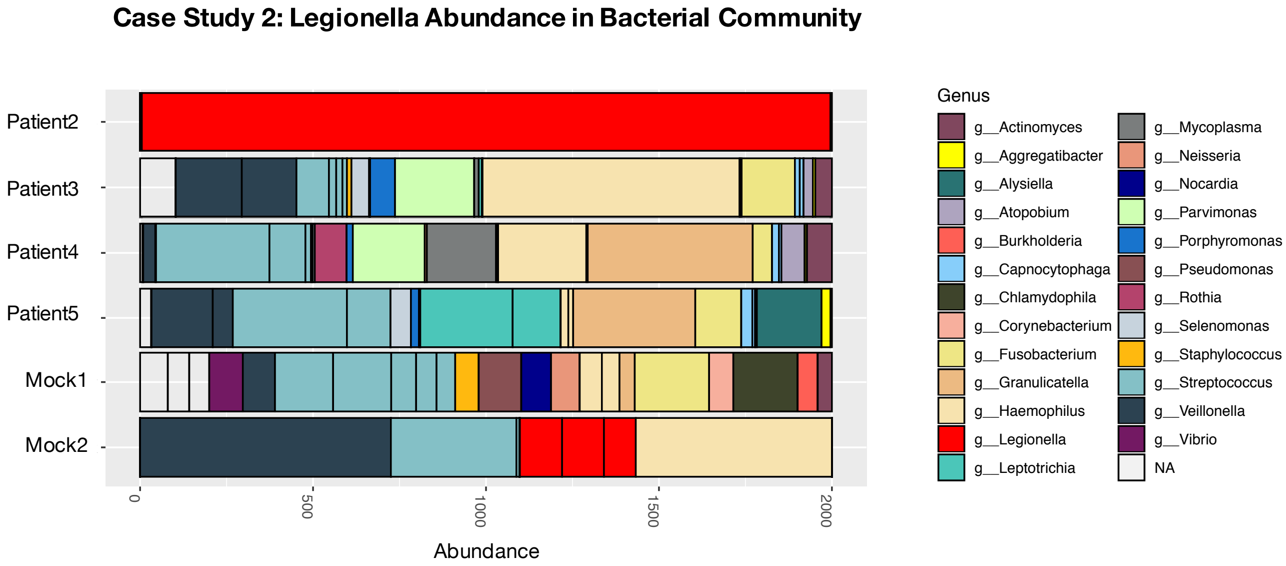


Figure 6.10. *Legionella* relative abundance in the bacterial communities of sputum extracts from Case Study 2.

6.4.3.3 Case Study 2: *L. pneumophila* Target Capture

Reads from Case Study 2 post-capture libraries were mapped to a *L. pneumophila* completed ST37 reference genome (https://www.ncbi.nlm.nih.gov/nuccore/NZ_LT632616). The percentage of reference bases covered, the percentage of on-target reads, the mean depth of coverage before and after read duplicate removal were all investigated (**Table 6.18**). The percentage of the *Legionella* reference genome covered varied from 2 % for Patient1 to 90 % for Patient2. The proportion of reads mapping to the *Legionella* reference genome varied from 0.006 % for Patient1 to 26 % for Patient2. Mean depth of coverage varied from 0.05 times for Patient1 to 56 times for Patient5 after duplicate removal. Read duplication levels varied from 28 % for Patient1 to 85 % for Patient5.

Table 6.18 *L. pneumophila* Target Capture Statistics for Case Study 2.

Genome	Genome Copy Number Input	Reference Genome covered (%)	Reads mapped to Reference (%)	Mean Depth of Coverage (SD) BEFORE Duplicate Removal	Mean Depth of Coverage (SD) AFTER Duplicate Removal	Read Duplication Levels (%)
Patient1	2,778	2	890 (0.006)	0.07 (0.9)	0.05 (0.5)	28
Patient2	37,383	90	2,193,382 (26)	173 (2133)	56 (927)	67
Patient3	16,667	16	83,700 (1)	6 (87)	1 (11)	82
Patient4	1,250	4	41,790 (0.3)	3 (126)	0.7 (23)	78
Patient5	833	4	257,390 (1.2)	20 (818)	3 (127)	85

6.4.3.4 Case Study 2: *In silico* Sequence-Based Typing

An *in silico* sequence-based typing analysis was carried out on the QC'ed sequencing reads of each sample. **Table 6.19** details the sequence type results from the target capture data for Case Study 2. In the case of target capture data from Patient2, six alleles were obtained (*3,4,1,1,14,9,0*). The *neuA* allele number was not determined however the profile is consistent with ST37, as confirmed from the matched isolate. No alleles numbers were determined for

Patient1, 3, 4 or 5. Oxford Nanopore reads sequenced directly from Patient2 and 3 samples did not yield a sequence type.

Table 6.19 *In silico* Sequence-Based Typing for Case Study 2 Target Capture Data.

Genome	*flaA*	*pilE*	*asd*	*mip*	*mompS*	*proA*	*neuA/h*	ST
Patient1	0	0	0	0	0	0	0	-
Patient2	3	4	1	1	14	9	-	-
Patient3	0	0	0	0	0	0	0	-
Patient4	0	0	0	0	0	0	0	-
Patient5	0	0	0	0	0	0	0	-

6.4.3.5 Case Study 2: Identification of 50- and 1,455-Core MLST Genes

A gene presence/absence analysis was carried to determine the typeability of genes from the target capture data based on the 50-gene and the extended 1,455-gene MLST scheme proposed by David *et al.*, 2016(b). From the 50-core gene scheme, 40 genes were sequenced from Patient2 target capture data and 1,171 genes from the 1,455 core gene scheme. No full length core genes were sequenced from the 50-gene or 1,455 gene schemes for Patient1, 3, 4 or 5 data (**Table 6.20**). Oxford Nanopore reads sequenced directly from Patient2 and 3 samples did not yield full genes from the 50- or 1,455 core gene sets.

Table 6.20 Core 50- and 1,455-gene MLST of Case Study 2 Target Capture Data.

Genome	50 core genes	1,455 core genes
Patient1	0	0
Patient2	40	1,171
Patient3	0	0
Patient4	0	0
Patient5	0	0

6.4.3.6 Case Study 2: Phylogenetic Analysis

Due to lack of overlapping data captured and sequenced from Patient samples 1, 3, 4 and 5, a phylogenetic analysis could not be performed on core gene data from these samples. A maximum-likelihood phylogenetic tree based on alignment of overlapping regions from 1455 core genes was therefore constructed containing target capture data from Patient2 only and isolate data from Patient2 and 3 and environmental isolates 1 and 2 (**Figure 6.11**). In addition to regions of similarity, the alignment contained 1009 SNPs from 87 of the core genes (See Appendix Section 9.8 for list of 87 core genes). There were no SNP differences between the aligned core gene regions of isolates (both patient and environmental) from Case Study 2. Patient2 target capture data however differed from Case Study 2 isolate data by 9 SNPs.

For context, a second phylogenetic tree was constructed (**Figure 6.12**) containing the Case Study 2 genomes and including complete and draft ST37 genome assemblies deposited in the NCBI RefSeq server on the 5th of October 2018 (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/).

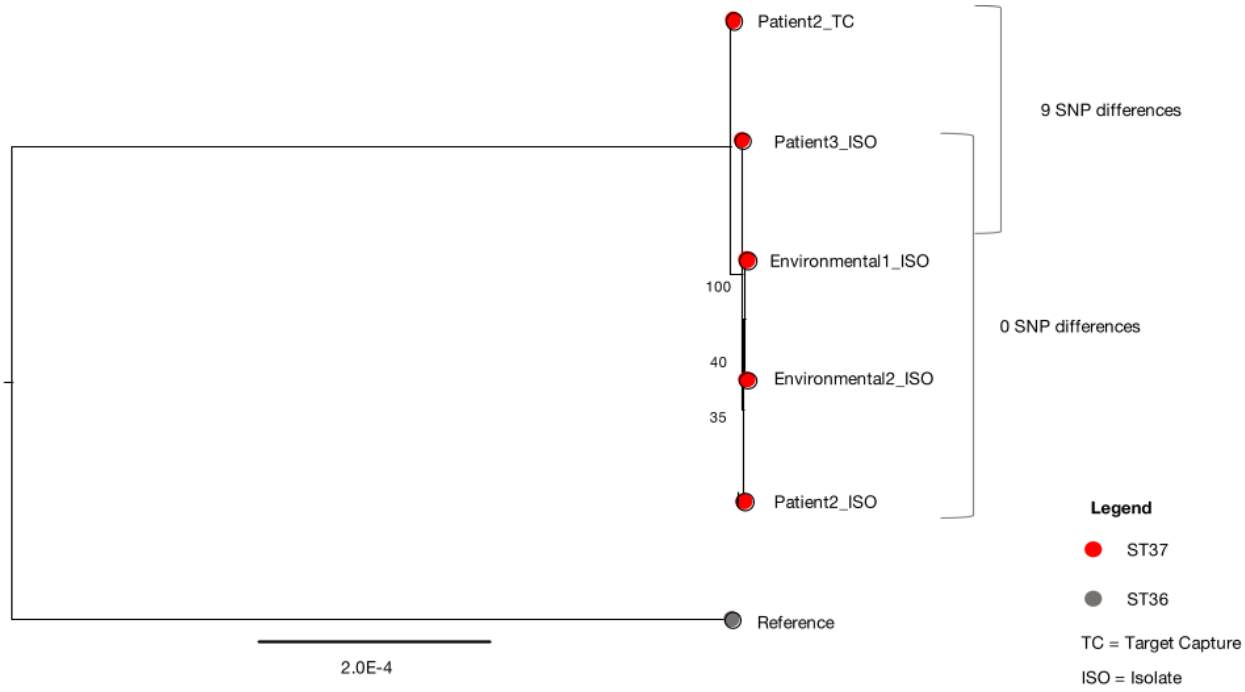


Figure 6.11 Maximum-likelihood Phylogenetic Tree for Case Study 2. Tree is based on an alignment of overlapping genomic regions from 1455 core genes of Patient2 target capture data, isolate data from Patient2 and 3 and isolate data from two environmental sources. A reference (*L. pneumophila* Philadelphia-1 [ST36]) was also included. In addition to regions of similarity, the alignment contained 1009 SNPs from 87 of the core genes. The internal nodes are labelled with the bootstrap values.

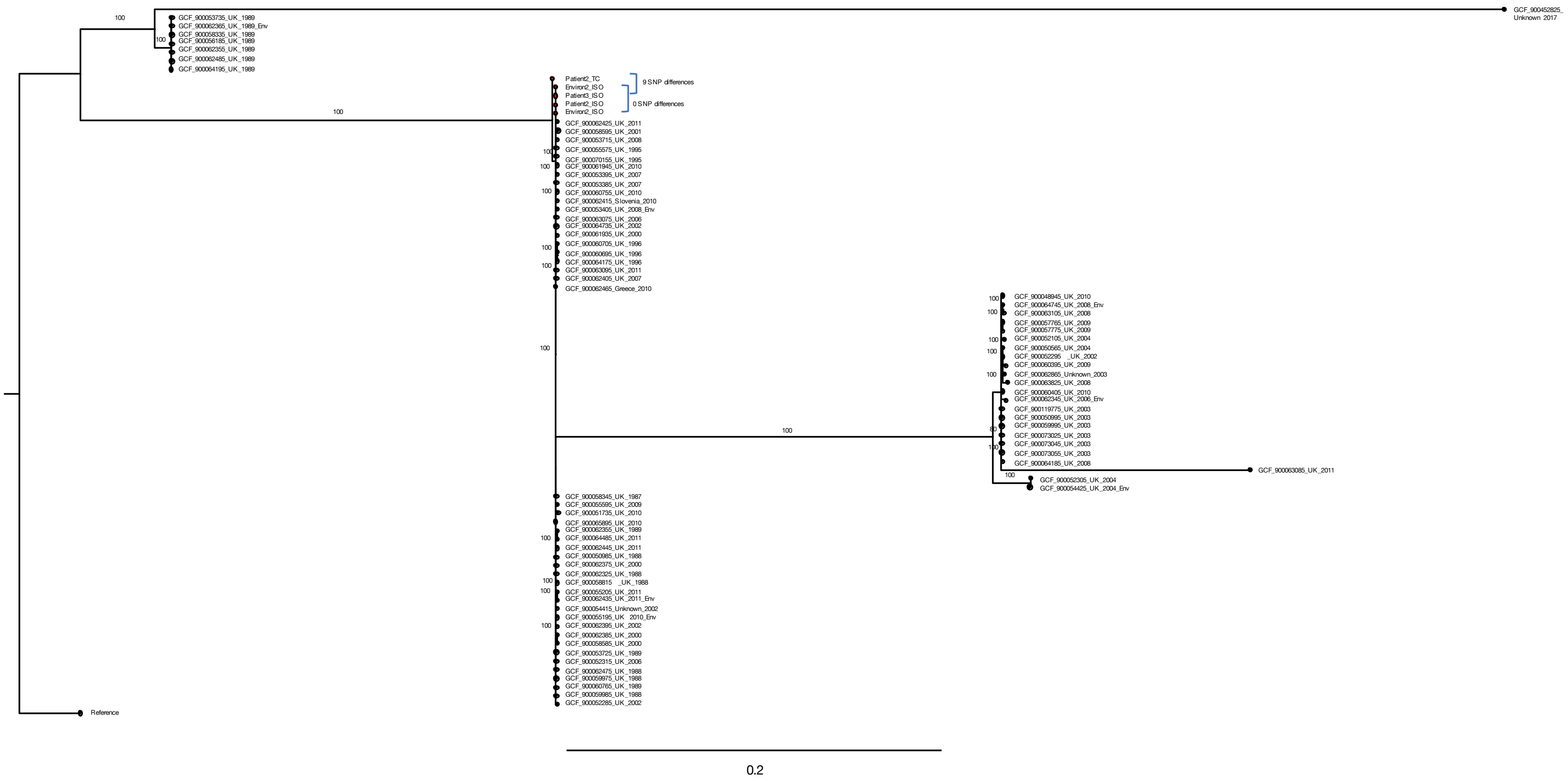


Figure 6.12 Maximum-likelihood Phylogenetic Tree for Case Study 2 and ST37 RefSeq Genomes. Tree is based on an alignment of overlapping genomic regions from 1455 core genes of Patient2 target capture data, isolate data from Patient2 and 3 and isolate data from two environmental sources. A reference (*L. pneumophila* Philadelphia-1 [ST36]) was also included. Also included are ST37 genomes from RefSeq, for context. The internal nodes are labelled with the bootstrap values. The larger image of Case Study 2 sample phylogeny is visualised in **Figure 6.11**.

6.4.3.7 Case Study 2: Heterozygous SNP Analysis for Mixed Strain Detection

Heterozygous SNP analysis was carried out on target capture and isolate data from Case Study 2 samples. From Case Study 2 samples, no heterozygous SNPs were present in target capture data from Patient1, 3, 4 and 5 or isolate data. Target capture SNP data from the Patient2 had 3 heterozygous SNPs (representing 0.4 % of the total SNP profile) therefore it was determined that Patient2 target capture sample did not contain a mixed profile (**Table 6.21**).

Table 6.21 Proportions of Heterozygous SNPs in 1455 core genes from Case Study 2 samples.

ID	No. Het SNPs	No. Total SNPs	Proportion Het SNPs
Patient1_TC	0	0	0 %
Patient2_TC	3	789	0.4 %
Patient2_ISO	0	1,631	0 %
Patient3_TC	0	6	0 %
Patient3_ISO	0	1,626	0 %
Patient4_TC	0	1	0 %
Patient5_TC	0	0	0 %
Environ1_ISO	0	1,630	0 %
Environ2_ISO	0	1,633	0 %

6.4.3.8 Case Study 2: Genome Assembly and Pangenome Visualisation

Genomes from the target capture and isolate data were assembled and target capture assemblies were decontaminated by removing contigs belonging to taxonomic orders other than the order Legionellales. For target capture data, genome length varied from 3,055,819 bases for Patient2 to 3,282 bases for Patient1. The maximum contig length varied from 15,311 for Patient2 to 637 for Patient1. Estimated levels of completeness based on the presence of single copy genes was 95.5 % for Patient2 and 0 % for Patient1, 3, 4 and 5 (**Table 6.22**). High quality drafts were assembled from isolate data for Patient2 and 3 and Environmental isolates 1 and 2 (**Table 6.23**).

As means of viewing genome assemblies within the context of an LD outbreak, a pangenome visualisation was created for Case Study 2 target capture and isolate data (**Figure 6.13**). The visualisation was based on the distribution of gene clusters across the genomes as detailed in Section 6.4.2.8. A total of 3,137 gene clusters were assigned based on 17,024 genes from Case Study 2 genomes. A total of 1,346 (42 %) of the gene clusters had >= 99 % functional homogeneity, 2,085 (66 %) had >= 75 % functional homogeneity and 3,035 (97 %) had >= 50 % functional homogeneity.

Table 6.22 Genome Assembly Quality Statistics: Target Capture Data for Case Study 2.

Genome	Genome Length (Bases)	Max Contig Length (Bases)	N50	GC Content	Completeness	Contamination	No. Predicted genes	Draft Quality	Coding Density	No. contigs
Patient1	3,282	644	483	48	0	0	7	-	0.9	8
Patient2*	3,055,819	15,311	1,573	39	95.5	5	4,407	High	0.898	2,759
Patient3*	58,289	679	305	40	0	0	184	-	0.83	182
Patient4	7,298	668	314	0.44	0	0	23	-	0.9	22
Patient5	3,722	637	456	0.5	0	0	9	-	0.93	9

* genome co-assembled with *Legionella* reads from Oxford Nanopore sequencing

Table 6.23 Genome Quality Statistics: Isolate Data for Case Study 2.

Genome	Genome Length (Bases)	Max Contig Length (Bases)	N50	GC Content	Completeness	Contamination	No. Predicted genes	Draft Quality	Coding Density	No. contigs
Patient2	3,440,549	648,422	336,253	38	100	0	3,077	High	0.888	32
Patient3	3,427,511	300,132	134,342	38	100	0	3,068	High	0.889	54
Environ1	3,441,134	648,389	253,366	38	100	0	3,078	High	0.888	37
Environ2	3,443,015	648,379	300,140	38	100	0	3,082	High	0.888	34

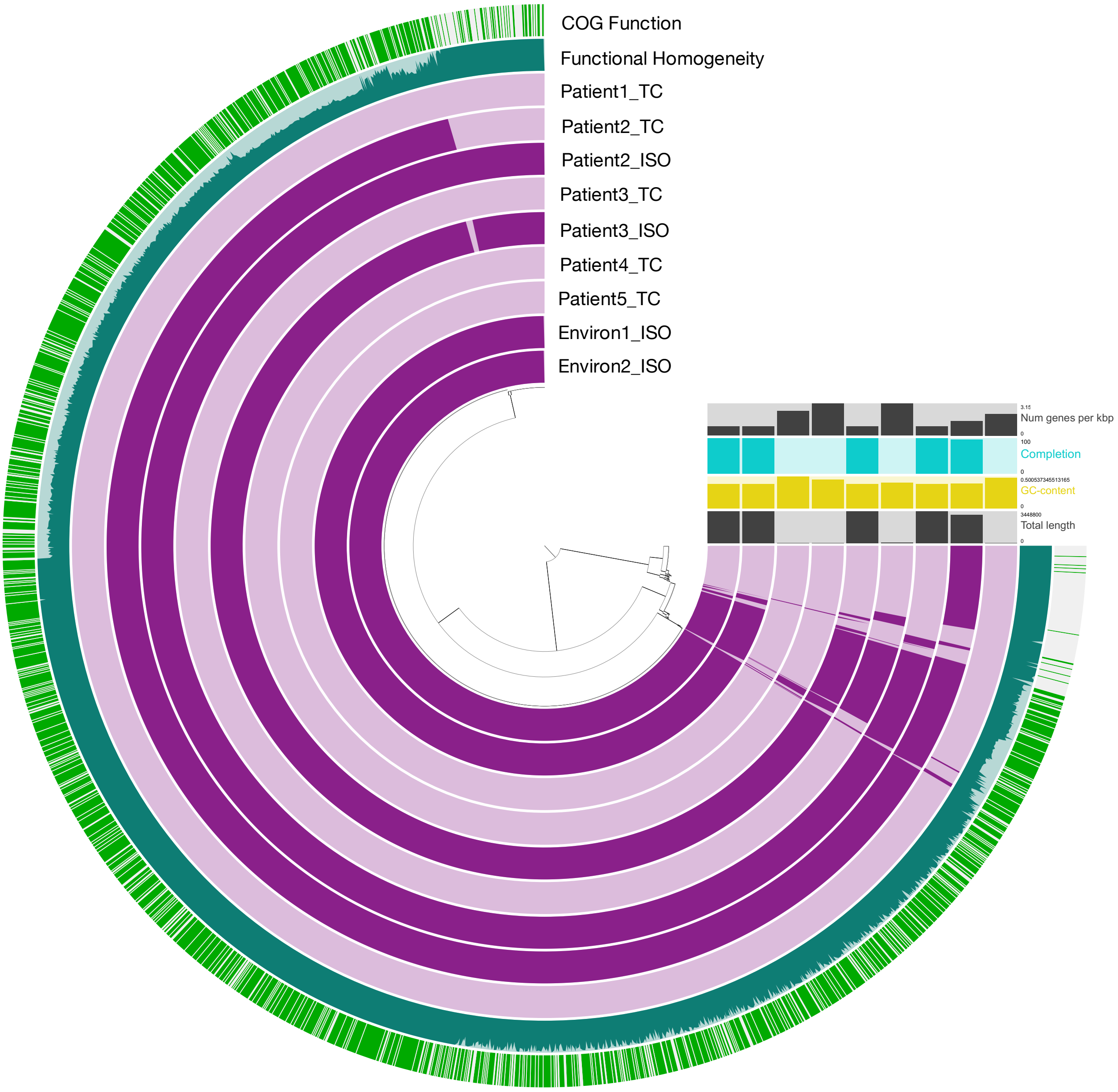


Figure 6.13. Pangenome visualisation of genomes assemblies from Case Study 2. The visualisation displays gene clusters present in the target capture and isolate genomes, COG function assignment (green for known and grey for unknown), functional homogeneity (homogeneity represented by spikes in green) and layers containing information on number of gene per kbp, GC content, levels of genome completion and total length. TC = target capture and ISO = isolate, COG = clusters of orthologous groups of proteins. Dark shading indicates gene clusters that were sequenced. Light shading indicates gene clusters that are absent from the assembly.

6.5 Discussion

Pilot data from Chapter 5 indicated that high quality, typeable genomes could be captured and sequenced from sputum DNA extracts. The aim of the current chapter was therefore to apply the target capture method to patient samples from two Legionnaires' Disease Outbreaks (Case Study 1 and Case Study 2) to examine the utility of the approach within a public health scenario requiring prompt investigation. A secondary aim was to determine if the targeted metagenomic method could provide additional information on the landscape of infection in the LD cases.

In the current chapter, mock material containing the *L. pneumophila* Philadelphia-1 strain was prepared to contain genome copy numbers of 4 x 10^4, 8 x 10^3, 4 x 10^3 and 8 x 10^2 for target capture and sequencing. The aim was to determine if a specific cut-off could be established between 10^4 and 10^3 genome copies. Reliable SBT data was obtained down to 4 x 10^3 genomes copies as well as a high depth of coverage on average. Additionally, 1,453 core genes were typeable at this level. This information is useful as it provides a more specific genome copy input cut-off (when compared to data generated in Chapter 5) with the aim of achieving high quality sequence data in potential future work.

An investigation of heterozygous SNPs from mock material (dilutions and mixed *Legionella* population mock community) was carried out as a control measure for further investigations of strain heterogeneity in case study samples This method was described by Sobkowiak *et al.,* for the investigation of mixed *M. tuberculosis* strains. *M. tuberculosis*, however, is highly clonal and is known to undergo very little recombination. Since the *L. pneumophila* genome is known to undergo frequent recombination events, heterozygous SNP analysis here was restricted to the 1,455 core MLST genes. Mixed strains sequenced in the mock community could be distinguished from single strain samples based on the frequency of heterozygous to homogenous SNPs. Furthermore, the mixed strain plot demonstrated a clustering profile based on minor and major allele frequencies. For the purpose of this study, the proportion of heterozygous SNPs in single strain controls provided a proportional cut-off for analysis of real samples from the LD case studies.

Case Study 1 involved a cluster of confirmed LD cases and one suspected LD case from England. This outbreak involved a confirmed ST47, a novel ST2110 and ST82. Whilst LD patients from Case Study 1 lived or worked in close proximity, some with daily cross-over, 2 were largely immobile. During the initial investigation in 2015, one common source of infection for all individuals could not be established. In the current study, target capture was carried out on five *L. pneumophila* positive patient cases and one *L. pneumophila* negative patient case (by UAT and qPCR at PHE). In addition, isolates from two patient samples were sequenced (no target capture).

Firstly, *Legionella* abundance in the bacterial community of the sputum samples was investigated by 16S rRNA gene sequencing and this revealed that *L. pneumophila* was present in extremely low abundance or was undetectable in 3 patient samples. Two patient samples had a high proportion of *Legionella* OTUs and one sample dropped out due to insufficient total number of OTUs. As previously discussed in Chapter 3 (Section 3.5), in a study by Mizrahi *et al*., 2017 on the legionellosis microbiome, *Legionella* was never the dominant genus in the bacterial communities of sputa from LD patients. This, in addition to the abundance of human DNA present in sputa, can hinder the capture of *Legionella* genomic regions and is a major caveat for metagenomic investigations of LD cases.

Analysis of target capture results demonstrated that good reference genome base coverage and depth of coverage was achieved for the genome from Patient5. Poor quality of capture was observed for samples from Patient3, 4 and 6, however the input genome copy number was low for these samples. Surprisingly, capture of genomes from Patient1 and Patient2 (extracted by MagnaPure Compact) was less efficient than expected as these contained a high genome copy number input. When examining duplicate extracts for Patient3, 4, 5 and 6, it can be seen that capture was more efficient for samples extracted by the Phenol-Chloroform method ('b') than MagnaPure Compact ('a'). It is uncertain why this is the case. It may be due to the long-term storage of the MagnaPure Compact extracted DNA, however, further studies would be needed to confirm this. Even though DNA quantity was normalised for targeted capture, qPCR was not carried out on the newly extracted samples (by Phenol-Chloroform) therefore it may be that this extract contained more genome copies. Of interest from this analysis was that *Legionella* genomic regions were captured from Patient4, a suspected LD case which was previously confirmed negative and excluded from the cluster investigation. Despite low depth of coverage, approximately 20 % of reference genome bases were covered.

After analysis of target capture data, duplicate extractions from the same patients were merged and *in silico* SBT and core genome MLST (50 and 1,455 genes) analyses were carried out. From 7-loci SBT analysis, while a partial sequence type was obtained from

Patient1 and Patient2 and a full sequence type from Patient5, no allele numbers were determined for Patients 3 or 4. However, a *proA*-2 was obtained from Patient6, adding new SBT information to the cluster profile as no alleles were determined during previous laboratory analysis. Through application of the 50- and 1,455-core gene MLST schemes, typeable genes were obtained for nearly all patient samples apart from Patient3. Ideally all 50 genes should be sequenced with 100 % coverage. However, analysis of partial type data may be useful particularly for data generated by metagenomic sequencing which is often fragmented.

Next phylogenetic analysis was carried out on SNPs from partial genes. Isolate and target capture genomes from Patient1 and Patient2 formed two clades with the clustering of the Patient5 target capture genomes with the Patient1 clade as expected. Samples from Patient3, 4 and 6 did not group with Patient1 or 2 samples. A low-resolution tree grouped Patient4 and Patient6 samples together however their relationship could not be further confirmed. A caveat here is the low genome copy number available in these samples for target capture which was not sufficient to carry out a high-resolution phylogenetic analysis. Heterozygous SNP analysis revealed an interesting result for Patient5. There were good indications of mixed *L. pneumophila* populations in this sample with heterozygous SNPs representing a 13 % proportion of total SNPs. To further corroborate this, the genome was assembled, and quality analysis indicated a degree of contamination (13 %) that was predominantly attributable to strain heterogeneity (> 90 % amino acid identity). This is an interesting result as Patient5 was linked to visiting a residential site where a spa pool was located. The spa pool was found to contain mixed *L. pneumophila* populations and due to this sequence types could not be confirmed by SBT.

Based on the ability to capture and sequence mixed *L. pneumophila* populations from mock samples and clinical samples as reported in this present study as well as pilot data on capture from environmental specimens in Chapter 5 (Section 5.4.4), this could represent an excellent application of metagenomic sequencing to an environmental sample for LD cluster source investigation.

A medium quality draft genome was assembled from Patient5 and low-quality draft genomes were assembled from Patients 1 and 2. Analysis and visualisation of the Case Study 1 pangenome showed the captured gene clusters demonstrated a good degree of functional heterogeneity. While an attempt was made to extract the gene clusters sequenced for all samples for alignment and phylogenetic analysis, this was not possible due to partial gene calls in some of the clusters.

Case Study 2 involved an outbreak of ST37 LD cases from an industrial site in England. A number of locations on site were attributed as environmental sources of infection. *Legionella* abundance was examined in the bacterial communities' sputum nucleic acid extracts from 5 patients by 16S rRNA sequencing. *Legionella* abundance was high in one sample (Patient2) and extremely low in 3 samples. One sample (Patient1) did not contain sufficient number of total OTUs and was therefore excluded from further analysis. Target capture was successful for the high abundance *Legionella* sample from Patient2. Minor capture only was achieved for the samples from Patients 1, 3, 4 and 5 and this was also reflected in the genome assembly results. Phylogenetic analysis was carried out for Patient2 target capture data with the available isolate data on SNP sites from 1,171 core genes revealing no SNPs between environmental and clinical isolates and 9 SNP differences between target capture data for Patient2 and the isolates (including a matched isolate for Patient2). No heterozygous sites were observed in isolate genomes and the low proportion (3 heterozygous sites representing a 0.4 % proportion of total SNPs) in Patient2 target capture data did not warrant further investigation based on single strain cut-offs.

From phylogenetic and pangenome analysis, the available Case Study 2 genomes appeared to be genetically and functionally homogenous. Unfortunately, the low copy number in samples other than those of Patient2 did not allow for adequate capture and sequencing of genomic regions for analysis here.

For both case studies, Oxford Nanopore Technologies (ONT) sequencing was investigated directly on sputum samples (without human DNA depletion). A number of long reads were sequenced and incorporated into assemblies; however, the long reads alone were not sufficient to allow allele number determination. Recent efforts using ONT sequencing directly on sputum DNA extracts with human DNA depletion has demonstrated effectiveness in determining the aetiology of respiratory infection, when compared to isolate results (Charalampous *et al*., 2019). In the case of LD cases and low abundance of *L. pneumophila* sputum samples, this may be a challenge but is currently being investigated by other researchers (personal communication – Dr. Victoria Chalker).

The results from this Chapter demonstrate both the utility of metagenomic sequencing for LD cluster investigations and the caveats encountered. Despite the challenges, the use of metagenomic data in addition to isolate data may provide an additional layer of evidence during LD cluster and outbreaks, enhancing the resolution of an urgent public health investigation.

Chapter 7.

Discussion

7.1 General Discussion

Legionella species, which cause Legionnaires' Disease (LD), are both difficult-to-grow and slow-growing bacteria which may delay efforts in applying typing methods, particularly core genome typing to a whole genome sequence from isolate material in the case of endemic clones. For residents of England and Wales, cases of LD must be notified to Public Health England (PHE). PHE coordinate the national surveillance scheme for LD with the primary objectives of detecting clusters and outbreaks of *Legionella* in England and Wales, identifying the source of infection so that control measures are implemented and collaborating with the European surveillance network in the detection, control and prevention of clusters and outbreaks of LD in European countries by reporting travel-associated cases.

The aim of this thesis was to develop metagenomic methods for the sequencing of *Legionella* from clinical and environmental specimens that would provide a more timely approach for the detection and identification of *Legionella*, reduce diagnostic selection bias and provide insights into potential mixtures of *L. pneumophila* subtypes, sequence types and *Legionella* species in samples which might aid investigations.

This was addressed in four parts by:

1. Validating a metagenomic pipeline for the sequencing and analysis of *Legionella* from mock samples and examining the proportion of contaminating human DNA in *Legionella pneumophila* (*L. pneumophila*) positive and negative sputum samples (Chapter 3).
2. Exploring the development of hybridisation- and PCR-based methods for the depletion of human DNA from clinical samples through targeting the abundance of repetitive regions in the human genome (Chapter 4).
3. Investigating the use of a targeted capture approach (Agilent SureSelect™) for the enrichment of *L. pneumophila* from clinical and environmental samples using biotinylated RNA baits (Chapter 5).
4. Exploring a Legionnaires' Disease cluster and outbreak in England using the Agilent SureSelect™ targeted capture approach and direct Oxford Nanopore sequencing (Chapter 6).

Technical validation of diagnostic tests is an evidence-based fitness-for-purpose analysis of the performance of a test in the laboratory. Validation should be performed before the routine introduction of either non-standard methods or methods developed in-house or used beyond their intended purpose as well as validated methods subsequently modified (PHE, 2017[b]).

In **Chapter 3**, technical validation steps were carried out using mock communities containing *Legionella* at varying proportions. The results demonstrated that > 97 % of *L. pneumophila* genome base coverage could be achieved when *L. pneumophila* was present at a proportion of 10 % in a sample (with an average depth of 7 times) based on 24 times multiplexing on a MiSeq. Furthermore, all 7-loci relevant to the *L. pneumophila* traditional sequence-based typing (SBT) approach could be determined at that proportion. From a rapid surveillance point of view, this may be relevant information if *Legionella* qPCR and total DNA quantification results are available for the sample of interest as it may be possible to then sequence *Legionella* directly from the sample. A caveat here however is that analysis was based on human DNA spiked with *L. pneumophila* DNA only. Defining cut-offs based on real specimens spiked with *Legionella* DNA would be advantageous over this approach. Furthermore, based on direct sequencing of sputum samples in **Chapter 3** as well as previous studies (Doughty *et al.,* 2014, Pendleton *et al*., 2017), the burden of host DNA overwhelms the microbial community in sputum specimens. A further caveat is that in sputum samples analysed in **Chapter 3**, *L. pneumophila* was never the dominant microorganism.

The Clinical Microbiology Laboratory (be it at PHE or hospital based) requires cost-effective, validated and actionable tools. The use of culture, antimicrobial resistance tests, qPCR, epidemiological gene typing (or whole genome typing from isolate data), multi-target panels and serology are important to satisfy these functions. Respiratory infections are the fourth largest cause of mortality worldwide and account for 60 % of all antibiotics prescribed in general practice in the United Kingdom (NICE, 2008). It is likely that the rapid and unbiased detection of respiratory pathogens in one process would positively impact public health microbiology. While clinical metagenomic sequencing for pathogen detection is still in its infancy, a number of case studies have generated insight into its applicability in the detection of respiratory pathogens (Yan *et al*., 2016, Leo *et al*., 2017, Charalampous *et al*., 2019). But respiratory pathogen detection by metagenomics is confounded by the low microbial biomass nature of specimens and therefore is an implicit risk for ambiguous results due to sample contamination. The introduction of contaminants can occur at any stage from sample collection to library preparation and sequencing (Salter *et al.*, 2014). A number of groups have validated metagenomic sequencing pipelines in licensed microbiology laboratories (Schlaberg *et al*., 2017[b], Hong *et al.,* 2018, Blauwkamp *et al.,* 2019, Miller *et al*., 2019). Schlaberg *et al.,* 2017(a) provided example workflows for validating the detection of pathogens by metagenomic sequencing in clinical microbiology laboratories including defining sensitivity and specificity for pathogen detection, reproducibility, specimen stability, bioinformatic validation factors and defining quality cut-offs. Additionally, Miller *et al.*, 2019 developed and validated a clinical metagenomic sequencing pipeline for the diagnosis of pathogens causing encephalitis and meningitis from cerebrospinal fluid with the purpose of implementing it in a licensed microbiology laboratory. To date, however, there is no gold standard approach for the validation of metagenomic pipelines. In the future it is likely that study groups will define appropriate cut-offs and standards for diagnostic metagenomic sequencing approaches specific to individual or panels of pathogenic microorganisms as well as standards for the interpretation of the clinical significance of the reported findings. In the case of *Legionella*, this may be more challenging due to the low abundance of the bacteria within the microbial community particularly as current diagnostic metagenomic approaches for respiratory pathogen detection focus on the identification of the dominant microorganism(s) (Charalampous *et al*., 2019). Consequently, a number of approaches may need to be implemented for direct *Legionella* sequencing and characterisation.

One approach to address this challenge relies on the use of methods for the depletion of human DNA from clinical samples. Human DNA sequences often comprise the majority of reads when a lower respiratory sample is sequenced. This may be due to neutrophils present in the airway (Brinkmann *et al.,* 2004), release of contents from the human cells during cell death (Wartha *et al.,* 2007) or the size of the human genome which is approximately one thousand times larger than the average bacterial genome. This can rapidly create a disproportionate noise to signal ratio.

The development of a number of methods for human DNA depletion was explored in **Chapter 4**. Methods investigated were hybridisation-based (using biotinylated Cot-1 DNA, *Alu* DNA and *Alu* RNA) and PCR-based (*Alu* primers with biotinylated dTTP). Despite the number of strategies investigated, there was insufficient or no depletion of human DNA from mock and real samples. A number of solutions for human depletion have been addressed by others both in the form of commercial kits and other non-commercialised methods developed in-house. One commercial solution offered is the NEBNext Microbiome DNA Enrichment Kit (New England Biolabs) which targets and removes human DNA by immunomagnetic separation based on the methylation differences between prokaryotic and eukaryotic DNA (Feehery *et al*., 2013). Variable results have however been reported for this method (Thoendel *et al*., 2016, Marotz *et al*., 2018) and the kit is likely not to be cost-effective for routine application in a clinical microbiology laboratory. Other methods are based on the immunoprecipitation of DNA with inactive methyl-specific restriction endonucleases (Barnes *et al*., 2014, Liu *et al*., 2016). Due to cost-effectiveness and speed, the most studied methods are based on the lysis of human cells and the subsequent elimination of the released human genomic DNA. Two commercial kits (MolYsis Basic5 kit [Molzyme] and QIAamp DNA Microbiome Kit [Qiagen]) are based on differential human cell lysis followed by DNase treatment. Published in-house lysis methods are based on osmotic lysis followed by propium monoazide treatment (Marotz *et al*., 2018), saponin-based lysis followed by DNase treatment (Hasan *et al*., 2016), saponin-based lysis followed by HL-SAN treatment (Charalampous *et al.,* 2019) and hypotonic lysis and endonuclease digestion with benzonase2 (Nelson *et al*., 2019). These methods have demonstrated very good efficiency of removal of DNA, significantly improving the depth and overall proportion of sequenced microbial reads. Future studies however will be needed to further evaluate the effects lysis methods have on the microbial metagenome and if the methods introduce compositional biases particularly in low microbial biomass samples such as sputum. Furthermore, as discussed in **Chapter 1 Section 1.10.3**, human cell lysis can lead to the elimination of bacteria without cell walls or removal of cell-free nucleic acid from bacteria that have lysed during antibiotic treatment.

In the context of *Legionella*-specific detection by metagenomic sequencing, a human cell lysis method may be appropriate however this will require validation studies using mock communities and real samples. These evaluations were beyond the scope of this thesis but are being addressed by other investigators (personal communication Dr. Victoria Chalker).

Due to the insufficient depletion of human DNA by the approaches investigated in **Chapter 4**, a pilot study was performed in **Chapter 5** to evaluate a targeted capture approach for the enrichment of *L. pneumophila* genomes directly from clinical and environmental specimens. To my knowledge, results from **Chapter 5** are the first to demonstrate that draft *L. pneumophila* genomes can be captured and sequenced from patients with LD and from environmental source samples without prior culture. The data generated also demonstrated that *Legionella* diversity (partial single copy genes from other *Legionella* species) within environmental sources could be captured. Importantly, the work has additionally demonstrated the first successful application of *in silico* 7-loci SBT (Gaia *et al*., 2005, Ratzow *et al*., 2007, Mentasti *et al*., 2014) and 50 core gene MLST (David *et al*., 2016[b]) to *Legionella* data generated by a metagenomic method. Ultimately, **Chapter 5** demonstrated the proof of concept of targeted metagenomic sequencing of *L. pneumophila* directly from multiple patients and environmental sources as well as the ability to capture a variety of sequence types.

These results may be informative for *Legionella* surveillance laboratories as this approach is rapid and typing results are obtained in one process. There are however two main challenges to successful *Legionella* bait capture. The first and foremost challenge is the quality of the DNA extract. Ideally DNA should be extracted from as much of the sample as possible, using a non-kit-based approach to avoid the introduction of reagent contaminants and DNA should be eluted in a concentrated manner. This is important as the minimum requirement to sequence a good quality draft genome was identified as > 4,000 *L. pneumophila* genome copies. In future studies, an assessment of archived qPCR results from previous years for *L. pneumophila* positive specimens may give an indication of the percentage of specimens from which one would expect to obtain a good quality draft genome using the targeted metagenomics sequencing approach. A cost analysis could then be carried out to determine the feasibility of this approach compared to culture and whole genome sequencing or direct nested SBT.

The second challenge is the recombinogenic nature of the *Legionella* genome which may impede the cross-reactivity of the RNA capture baits. The disparity in genomic regions captured and sequenced across the specimen panels made it challenging to identify variances in hybridisation efficiency between the different *L. pneumophila* sequence types. This aspect will require evaluation in future work. A solution that could address this is that a panel of core genes (e.g. 1,455 core genes [David *et al*., 2016(b)]) could be targeted and sequenced, rather than a whole genome. This would, for example, allow sufficient discrimination between clusters of endemic clones without the requirement of culture or a whole genome sequence. Furthermore, this approach would be more cost-effective and it is likely that sequence depth would be higher for the targeted genes thereby allowing greater confidence in the assignment of allele numbers.

The application of the targeted capture approach was extended to the investigation of a cluster and outbreak of Legionnaires Disease (LD) in **Chapter 6**. To my knowledge, this is the first investigation of LD outbreaks using metagenomic methods.

The Case Study 1 LD outbreak from England in 2015 involved a ST47, ST82 and the novel ST2110. A sequence type was not identified from the three other available patient samples. Through the application of *L. pneumophila* target capture, it was found that one patient sample, confirmed negative by routine methods, was positive using target capture - although a sequence type could not be confirmed. Additionally, a *proA*-2 allele was retrieved from another of the patient samples. No previous allele data was obtained from this sample using traditional approaches. Finally, good evidence of a mixed *L. pneumophila* infection profile was observed in the captured data from the ST82 patient sample. This result was particularly interesting as the patient had visited a residential property with a spa pool confirmed to contain mixtures of *L. pneumophila*. Unfortunately, investigation of patient samples from Case Study 2 did not yield additional information due to the limited quantity of *L. pneumophila* genome copies in the extracts. Since samples were extracted and stored in 2015 and stored sputum samples were available for re-extraction only for 4 cases of Case Study 1, this represents a caveat to the current investigation.

A real-time investigation of a LD cluster or outbreak using *L. pneumophila* target capture would be an optimal approach for assessing the capabilities of the process in this context as fresh samples would be available for immediate DNA extraction, *L. pneumophila* capture, sequencing and analysis. There are however some ethical and legal implications to performing this optimal type of study. As discussed previously, sputum typically has a large amount of human DNA relative to microbial DNA. A significant proportion of sequences generated by metagenomics are off-target and include human DNA reads (data shown in **Chapter 5** and data from **Chapter 6** [not shown]). Human DNA sequences represent patient-identifiable data therefore ethical approval must be sought before sequencing to guarantee patient privacy. This would delay the investigation of a cluster or outbreak considerably.

While this requirement holds true for the metagenomic investigation of patient samples for any public health purpose, a further hurdle is encountered during a Legionnaires' Disease investigation. Legal action may be taken by individuals who contracted LD against businesses, travel-facilities, etc, if an uncontrolled source is identified. A legal hearing may require microbiological evidence obtained from validated diagnostic tests. If incidental evidence is generated by metagenomic sequencing this too will need to be put forward. Due to the current difficulty of interpretation of metagenomic data as well as sparsity of data in some cases, such data may be of low validity. As a consequence of this, samples from more recent LD clusters and outbreaks could not be included in the present study.

A number of recent studies have highlighted the utility of metagenomic sequencing during outbreak investigations (Loman *et al*., 2014, Quick *et al*., 2016, Faria *et al*., 2017, Huang *et al*., 2017 Kafetzopoulou *et al*., 2019). In the future, with greater clarity regarding legalities and patient privacy, it is expected that the same can be applied to LD clusters and outbreaks leading to the rapid recovery of genomic regions and the timely resolution of cases and environmental sources.

7.2 Future Directions

The future of metagenomics and case, cluster and outbreak investigations lie within the ability to rapidly detect a causative pathogen. Increased availability of sequencing technology in clinical laboratories and the increased expertise within the microbiological community on the application of bioinformatics for analysis are key to future adoption. The speed and portability of Oxford Nanopore Technology (ONT) sequencing has already impacted public health microbiology investigations and is continuing to generate more reliable, high quality data with continual improvements in sequencing accuracy. Recently, ONT released protocols for the application of the Agilent SureSelect™ RNA capture approach to their sequencing pipeline. Based on the proof-of-concept of the approach described in this thesis, it could be a very interesting avenue of investigation for *Legionella* cases, further improving time-to-results.

In the case of an agnostic, untargeted approach to *Legionella* sequencing, future development or further validation of host depletion methods will be required. Future work will surely address validation studies for pathogen presence or absence and developing standards for the clinical interpretability of the data.

Currently, metagenomic sequencing datasets are being generated at a rapid rate and require massive amounts of space for data processing and analysis. Bioinformatic infrastructure has lagged behind this. In the future, the standardisation or development of bioinformatic tools will be required along with defining metrics for critical assessment and validation of the tools. Further to this, effort and funding will be required to maintain bioinformatic tools and infrastructure. Additionally, genome reference databases are incomplete or contain sequences that are mis-identified or of poor quality. The curation of high-quality databases for metagenomic data analysis will be paramount, particularly for clinical microbiology laboratories.

To conclude:

This thesis has demonstrated the applicability as well as the challenges of metagenomic sequencing for *Legionella* detection. As it stands, the targeted metagenomic sequencing from LD cases and environmental source samples can provide typeable, draft genome information when extracts contain an adequate quantity of *Legionella* genome copies. Furthermore, the diversity of *Legionella* within the samples can be detected and genomic information from non-culturable or diagnostically ambiguous samples can be obtained.

8. Bibliography

Abbott JC, 2017. BugBuilder – An Automated Microbial Genome Assembly and Analysis Pipeline. bioRxiv 148783; doi: https://doi.org/10.1101/148783.

Abdel-Nour M, Duncan C, Low DE, Guyard C. Biofilms: the stronghold of *Legionella pneumophila*. *Int J Mol Sci*. 2013;14(11):21660–21675.

Abu Kwaik Y, Gao LY, Stone BJ, Venkataraman C, Harb OS. Invasion of protozoa by *Legionella pneumophila* and its role in bacterial ecology and pathogenesis. *Appl Environ Microbiol.* 1998;64(9):3127-3133.

Adeleke AA, Fields BS, Benson RF, Daneshvar MI, Pruckler JM, Ratcliff RM, Harrison TG, Weyant RS, Birtles RJ, Raoult D, Halablab MA. *Legionella drozanskii* sp. nov., *Legionella rowbothamii* sp. nov. and *Legionella fallonii* sp. nov.: three unusual new *Legionella* species. *Int J Syst Evol Microbiol.* 2001;51:1151-1160.

Ahn T-H, Chai J, Pan C. Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics*. 2015;31(2):170–177.

Albanese D, Donati C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nat Commun.* 2017;8(1):1–13.

Allen KW, Prempeh H, Osman MS. *Legionella pneumonia* from a novel industrial aerosol. *Commun Dis Public Heal.* 1999;2:294-296.

Ampel NM, Ruben FL, Norden CW. Cutaneous abscess caused by *Legionella micdadei* in an immunosuppressed patient. *Ann Intern Med.* 1985;102:630–632.

Andersen BB, Sogaard I. Legionnaires' disease and brain abscess. *Neurology.* 1987;37:333–334.

Arnow PM, Chou T, Weil D, Shapiro EN, Kretzschmar C. Nosocomial Legionnaires' disease caused by aerosolized tap water from respiratory devices. *J Infect Dis.* 1982;146:460 – 467.

Arnow PM, Boyko EJ, Friedman EL. Perirectal abscess caused by *Legionella pneumophila* and mixed anaerobic bacteria. *Ann Intern Med.* 1983;98:184–185.

Bachmann N, Sullivan M, Jelocnik M, Myers GS, Timms P, Polkinghorne A. Culture-independent genome sequencing of clinical samples reveals an unexpected heterogeneity of infections by *Chlamydia pecorum*. *J Clin Microbiol.* 2015; 53(5):1573-1581.

Bajrai LH, Azhar EI, Yasir M, Jardot P, Barrassi L, Raoult D, La Scola B, Pagnier I. 2016. *Legionella saoudiensis* sp. nov., isolated from a sewage water sample. *Int J Syst Evol. Microbiol.* 2016;66:4367-4371.

Bangsborg JM, Uldum S, Jensen JS, Bruun BG. Nosocomial legionellosis in three heart-lung transplant patients: Case reports and environmental observations. *Eur J Clin Microbiol Infect Dis.* 1995;14(2):99-104.

Barnes HE, Liu G, Weston CQ, King P, Pham LK, Waltz S, Helzer KT, Day L, Sphar D, Yamamoto RT, Forsyth RA. Selective microbial genomic DNA isolation using restriction endonucleases. *PLoS One.* 2014;9:e109061.

Bashir M, Ahmed M, Weinmaier T, Ciobanu D, Ivanova N, Pieber TR, Vaishampayan PA. Functional Metagenomics of Spacecraft Assembly Cleanrooms: Presence of Virulence Factors Associated with Human Pathogens. *Front Microbiol*. 2016;7:1321.

Batzer MA & Deininger PL. *Alu* repeats and human genomic diversity. *Nature Rev. Genet.* 2002;3:370–379.

Benson RF, Thacker WL, Waters RP, Quinlivan PA, Mayberry WR, Brenner DJ, Wilkinson HW. *Legionella quinlivanii* sp. nov. isolated from water. *Curr Microbiol.* 1989;18:195-197.

Benson RF, Thacker WL, Fang FC, Kanter B, Mayberry WR, Brenner DJ. *Legionella sainthelensi* serogroup 2 isolated from patients with pneumonia. *Res Microbiol.* 1990;141(4):453-463.

Benson RF, Thacker WL, Lanser JA, Sangster N, Mayberry WR, Brenner DJ. *Legionella adelaidensis*, a new species isolated from cooling tower water. *J Clin Microbiol.* 1991;29:1004-1006.

Benson RF, Thacker WL, Daneshvar MI, Brenner DJ. *Legionella waltersii* sp. nov. and an unnamed *Legionella genomospecies* isolated from water in Australia. *Int J Syst Bacteriol.* 1996;46:631-634.

Bercovier H, Steigerwalt AG, Derhi-Cochin M, Moss CW, Wilkinson HW, Benson RF, Brenner DJ. Isolation of legionellae from oxidation ponds and fishponds in Israel and description of *Legionella israelensis* sp. nov. *Int J Syst Bacteriol.* 1986;36: 368-371.

Berger KH, Isberg RR. Two distinct defects in intracellular growth complemented by a single genetic locus in *Legionella pneumophila*. *Mol Microbiol.* 1993;7:7–19.

Bibb WF, Sorg RJ, Thomason BM, Hicklin MD, Steigerwalt AG, Brenner DJ, Wulf MR. Recognition of a second serogroup of *Legionella longbeachae*. *J Clin Microbiol.* 1981;14(6):674-677.

Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, Hogle SL, Coe A, Bergauer K, Bouman HA, Browning TJ, De Corte D, Hassler C, Hulston D, Jacquot JE, Maas EW, Reinthaler T, Sintes E, Yokokawa T, Chisholm SW. Marine microbial metagenomes sampled across space and time. *Sci Data*. 2018;4;5:180176.

Blauwkamp TA, Thair S, Rosen MJ, Blair L, Lindner MS, Vilfan ID, Kawli T, Christians FC, Venkatasubrahmanyam S, Wall GD, Cheung A, Rogers ZN, Meshulam-Simon G, Huijse L, Balakrishnan S, Quinn JV, Hollemon D, Hong DK, Vaughn ML, Kertesz M, Bercovici S, Wilber JC, Yang S. Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nat Microbiol.* 2019;4(4):663-674.

Boamah DK, Zhou G, Ensminger AQ, O'Connor TJ. From many hosts, one accidental pathogen: The diverse protozoan hosts of *Legionella*. *Front Cell Infect Microbiol.* 2017;7:477.

Borchardt J, Helbig JH, Lück PC. 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. *Eur J Clin Microbiol Infect Dis.* 2008;27(1):29-36.

Bornstein N, Marmet D, Surgot M, Norwicki M, Meugnier H, Fleurette J, Ageron E, Grimont F, Grimont PAD, Thacker WL, Benson RF, Brenner DJ. *Legionella gratiana* sp. nov. isolated from French spa water. *Res Microbiol.* 1989;140:541-552.

Bornstein N, Mercatello A, Marmet D, Surgot M, Deveaux Y, Fleurette J. Pleural infection caused by *Legionella anisa*. *J Clin Microbiol.* 1989;27(9):2100-2101.

Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, Ngan CY, Cheng JF, Tringe SG, Woyke T. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics.* 2015;16(1):1–12.

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C; Genome Standards Consortium, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol.* 2017; 8;35(8):725-731.

Bozue JA, Johnson W. Interaction of *Legionella pneumophila* with *Acanthamoeba castellanii*: Uptake by coiling phagocytosis and inhibition of phagosome-lysosome fusion. *Infect Immun.* 1996;64:668–673.

Brabender W, Hinthorn DR, Asher M, Lindsey NJ, Liu C. *Legionella pneumophila* wound infection. *JAMA*. 1983;250:3091–3092.

Breiman RF, Cozen W, Fields BS, Mastro TD, Carr SJ, Spika JS, Mascola L. Role of air sampling in investigation of an outbreak of legionnaires' disease associated with exposure to aerosols from an evaporative condenser. *J Infect Dis.* 1990;161(6):1257-1261.

Brenner DJ, Steigerwalt AG, McDade JE. Classification of the Legionnaires' disease bacterium: *Legionella pneumophila*, genus novum, species nova, of the family Legionellaceae, familia nova. *Ann Intern Med.* 1979;90:656-658.

Brenner DJ, Steigerwalt AG, Gorman GW, Weaver RE, Feeley JC, Cordes LG, Wilkinson HW, Patton C, Thomason BM, Lewallen Sasseville KR. *Legionella bozemanii* sp. nov. and *Legionella dumoffii* sp. nov.: classification of two additional species of *Legionella* associated with human pneumonia. *Curr Microbiol.* 1980;4:111-116.

Brenner DJ, Steigerwalt AG, Gorman GW, Wilkinson HW, Bibb WF, Hackel M, Tyndall RL, Campbell J, Feeley JC, Thacker WL, Skaliy P, Martin WT, Brake BJ, Fields BS, McEachern HV, Corcoran LK. Ten new species of *Legionella*. *Int J Syst Bacteriol.* 1985;35:50-59.

Brenner DJ, Steigerwalt AG, Epple P, Bibb WF, McKinney RM, Starnes RW, Coleville JM, Selander RK, Edelstein PH, Moss CW. *Legionella pneumophila* serogroup Lansing 3 isolated from a patient with fatal pneumonia, and descriptions of *L. pneumophila* subsp. *pneumophila* subsp. nov., *L. pneumophila* subsp. *fraseri* subsp. nov., and *L. pneumophila* subsp. *pascullei* subsp. nov. *J Clin Microbiol.* 1988;26:1695-1703.

Brinkmann V, Reichard U, Goosmann C, Fauler B, Uhlemann Y, Weiss DS Weinrauch Y, Zychlinsky A. Neutrophil Extracellular Traps Kill Bacteria. *Science.* 2004;303(5663):1532-1535.

Britten RJ & Kohne DE. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science.* 1968;161:529–540.

Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZ, Depledge DP, Nikolayevskyy V, Broda A, Stone MJ, Christiansen MT, Williams R, McAndrew MB, Tutill H, Brown J, Melzer M, Rosmarin C, McHugh TD, Shorten RJ, Drobniewski F, Speight G, Breuer J. Rapid Whole-Genome Sequencing of *Mycobacterium tuberculosis* Isolates Directly from Clinical Samples. *J Clin Microbiol*. 2015; 53(7):2230-2237.

Buchbinder S, Leitritz L, Trebesius K, Banas B, Heesemann. J. Mixed lung infection by *Legionella pneumophila* and *Legionella gormanii* detected by fluorescent *in situ* hybridization. *Infection.* 2004;32:242–245.

Bushnell B, 2014. https://sourceforge.net/projects/bbmap.

Buultjens AH, Chua KYL, Baines SL, Kwong J, Gao W, Cutcher Z, Adcock S, Ballard S, Schultz MB, Tomita T, Subasinghe N, Carter GP, Pidot SJ, Franklin L, Seemann T, Gonçalves Da Silva A, Howden BP, Stinear TP. A Supervised Statistical Learning Approach for Accurate *Legionella pneumophila* Source Attribution during Outbreaks. *Appl Environ Microbiol.* 2017;83(21):e01482-17.

Burstein D, Amaro F, Zusman T, Lifshitz Z, Cohen O, Gilbert JA, Pupko T, Shuman HA, Segal G. Genomic analysis of 38 *Legionella* species identifies large and diverse effector repertoires. *Nat Genet.* 2016;48:167–175.

Campbell J, Bibb WF, Lambert MA, Eng S, Steigerwalt AG, Allard J, Moss CW, Brenner DJ. *Legionella sainthelensi*: a new species of *Legionella* isolated from water near Mt. St. Helens. *Appl Environ Microbiol.* 1984;47:369-373.

Campocasso A, Boughalmi M, Fournous G, Raoult B, La Scola B. *Legionella tunisiensis* sp. nov. and *Legionella massiliensis* sp. nov., isolated from environmental water samples. *Int. J Syst Evol Microbiol.* 2012;62:3003-3006.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7(5):335-6.

Casati S, Gioria-Martinoni A, Gaia V. Commercial potting soil as an alternative infection source of *Legionella pneumophila* and other *Legionella* species in Switzerland. *Clin Microbio Infect.* 2009;15(6):571-575.

Castellani Pastoris M, Lo Monaco R, Goldoni P, Mentore B, Balestra G, Ciceroni L, Visca P. Legionnaires' disease on a cruise ship linked to the water supply system: clinical and public health implications. *Clin Infect Dis.* 1999;28(1):33-38.

Cazalet C, Rusniok C, Brüggemann H, Zidane N, Magnier A, Ma L, Tichit M, Jarraud S, Bouchier C, Vandenesch F, Kunst F, Etienne J, Glaser P, Buchrieser C. Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. *Nat Genet*. 2004;36(11):1165-73.

CDC (Centre for Disease Control), 2018. Legionnaires' Disease Surveillance Summary Report, United States - 2014 and 2015. CDC; 2018.

Chan JZ, Sergeant MJ, Lee OY, Minnikin DE, Besra GS, Pap I, Spigelman M, Donoghue HD, Pallen MJ. Metagenomic analysis of tuberculosis in a mummy. *N Engl J Med.* 2013;369:289–290.

Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, Rae D, Grundy S, Turner DJ, Wain J, Leggett RM, Livermore DM, O'Grady J. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat Biotechnol.* 2019;37(7):783-792.

Charles M, Johnson E, Macyk-Davey A, Henry M, Nilsson JE, Miedzinski L, Zahariadis G. *Legionella micdadei* brain abscess. *J Clin Microbiol.* 2013;51(2):701-704.

Chee CE, Baddour LM. *Legionella maceachernii* soft tissue infection. *Am J Med Sci.* 2007;334:410–413.

Chen DJ, Procop GW, Vogel S, Yen-Lieberman B, Richter SS. Utility of PCR, Culture, and Antigen Detection Methods for Diagnosis of Legionellosis. *J Clin Microbiol.* 2015;53(11):3474-3477.

Chereshsky AY, Bettelheim KA. Infections due to *Legionella sainthelensi* in New Zealand. *N Z Med J.* 1986;99(801):335.

Cherry WB, Gorman GW, Orrison LH, Moss CW, Steigerwalt AG, Wilkinson HW, Johnson SE, McKinney RM, Brenner DJ. *Legionella jordanis*: a new species of *Legionella* isolated from water and sewage. *J Clin Microbiol.* 1982;15:290-297.

Chien M, Morozova I, Shi S, Sheng H, Chen J, Gomez SM, Asamani G, Hill K, Nuara J, Feder M, Rineer J, Greenberg JJ, Steshenko V, Park SH, Zhao B, Teplitskaya E, Edwards JR, Pampou S, Georghiou A, Chou IC, Iannuccilli W, Ulz ME, Kim DH, Geringer-Sameth A, Goldsberry C, Morozov P, Fischer SG, Segal G, Qu X, Rzhetsky A, Zhang P, Cayanis E, De Jong PJ, Ju J, Kalachikov S, Shuman HA, Russo JJ. The genomic sequence of the accidental pathogen *Legionella pneumophila*. *Science*. 2004;305(5692):1966-8.

Christiansen MT, Brown AC, Kundu S, Tutill HJ, Williams R, Brown JR. Whole-genome enrichment and sequencing of *Chlamydia trachomatis* directly from clinical samples. *BMC Infect Dis.* 2014;14:591.

Cianciotto NP, Fields BS. *Legionella pneumophila mip* gene potentiates intracellular infection of protozoa and human macrophages. *Proc Natl Acad Sci*. USA 1992;89:5188–5191.

Clark SA, Doyle R, Lucidarme J, Borrow R, Breuer, J. Targeted DNA enrichment and whole genome sequencing of *Neisseria meningitidis* directly from clinical specimens. *Int. J. Med. Microbiol.* 2018; 308:256–262.

Collins SL, Afshar B, Walker JT, Aird H, Naik F, Parry-Ford F, Phin N, Harrison TG, Chalker VJ, Sorrell S, Cresswell T. Heated birthing pools as a source of Legionnaires' disease. *Epidemiol Infect*. 2016;144(4):796-802.

Compain F, Bruneval P, Jarraud S, Perrot S, Aubert S, Napoly V, Ramahefasolo A, Mainardi J-L, Podglajen I.Chronic endocarditis due to *Legionella anisa*: the first case difficult to diagnose. *New Microbes New Infect.* 2015;8:113-115.

Cooke HJ & Hindley J. Cloning of human satellite IIIDNA: different components are on different chromosomes. *Nucleic Acids Res*. 1979; 10: 3177–3197.
Correia AM, Ferreira JS, Borges V, Nunes A, Gomes B, Capucho R, Gonçalves J, Antunes DM, Almeida S, Mendes A, Guerreiro M, Sampaio DA, Vieira L, Machado J, Simões MJ, Gonçalves P, Gomes JP. Probable Person-to-Person Transmission of Legionnaires' Disease*. N Engl J Med.* 2016;374(5):497-498.

Coscollá M & González-Candelas F. Direct sequencing of *Legionella pneumophila* from respiratory samples for sequence-based typing analysis. *J Clin Microbiol.* 2009; 47:2901–2905.

Coscollá M, Fernández C, Colomina J, Sánchez-Busó L, González-Candelas F. Mixed infection by *Legionella pneumophila* in outbreak patients. *Int J Med Microbiol*. 2014;304(3-4):307-313.

Cramp GJ, Harte D, Douglas NM, Graham F, Schousboe M, Sykes K. An outbreak of Pontiac fever due to *Legionella* longbeachae serogroup 2 found in potting mix in a horticultural nursery in New Zealand. *Epidemiol Infect.* 2010;138(1):15-20.

Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res*. 2015;43(3):e15.

Cunha BA, Burillo A, Bouza E. Legionnaires' disease. *Lancet*. 2016;23;387(10016):376-385.

Currie SL, Beattie TK, Knapp CW, Lindsay DSJ. *Legionella spp*. in UK composts – a potential public health issue? *Clin Microbiol Infect.* 2014;20(4):O224-229.

David *et al.,* 2016(a): David S, Rusniok C, Mentasti M, Gomez-Valero L, Harris SR, Lechat P, Lees J, Ginevra C, Glaser P, Ma L, Bouchier C, Underwood A, Jarraud S, Harrison TG, Parkhill J, Buchrieser C. Multiple major disease-associated clones of *Legionella pneumophila* have emerged recently and independently. *Genome Res*. 2016;26(11):1555-1564.

David *et al.,* 2016(b): David S, Mentasti M, Tewolde R, Aslett M, Harris SR, Afshar B, Underwood A, Fry NK, Parkhill J, Harrison TG. Evaluation of an optimal epidemiologic typing scheme for *Legionella pneumophila* with whole genome sequence data using validation guidelines. *J Clin Microbiol.* 2016;54(8):2135-2148.

David *et al.,* 2017(a): David S, Afshar B, Mentasti M, Ginevra C, Podglajen I, Harris SR, Chalker VJ, Jarraud S, Harrison TG, Parkhill J. Seeding and Establishment of *Legionella pneumophila* in Hospitals: Implications for Genomic Investigations of Nosocomial Legionnaires' Disease. *Clin Infect Dis.* 2017;64(9):1251-1259.

David *et al.,* 2017(b): David S, Sánchez-Busó L, Harris SR, Marttinen P, Rusniok C, Buchrieser C, Harrison TG, Parkhill J. Dynamics and impact of homologous recombination on the evolution of *Legionella pneumophila*. *PLoS Genet.* 2017;13(6):e1006855.

David S, Mentasti M, Parkhill J, Chalker VJ. Low genomic diversity of *Legionella pneumophila* within clinical specimens. *Clin Microbiol Infect*. 2018;24(9):1020.e1-1020.e4.

DeAngelis KM, Brodie, EL, DeSantis TZ, Andersen, GL, Lindow SE, Firestone MK. Selective progressive response of soil microbial community to wild oat roots. *The ISME Journal*. 2009;3(2):168-178.

Declerck P, Behets J, Delaedt Y, Margineanu A, Lammertyn E, Ollevier F. Impact of non-*Legionella* bacteria on the uptake and intracellular replication of *Legionella* *pneumophila* in *Acanthamoeba castellanii* and *Naegleria lovaniensis*. *Microb. Ecol.* 2005;50:536–549.

Declerck P. Behets J, De Keersmaecker B, Ollevier F. Receptor-mediated uptake of *Legionella pneumophila* by *Acanthamoeba castellanii* and *Naegleria lovaniensis*. *J. Appl. Microbiol.* 2007; 103:2697–2703.

de Felipe KS, Pampou S, Jovanovic OS, Pericone CD, Ye SF, Kalachikov S, Shuman HA. Evidence for acquisition of *Legionella* type IV secretion substrates via interdomain horizontal gene transfer. *J Bacteriol.* 2005;187(22):7716-7726.

De Filippis F, Parente E, Ercolini D. Metagenomics insights into food fermentations. *Microb Biotechnol*. 2016;10(1):91–102.

Deininger P. *Alu* elements: know the SINEs. *Genome Biol*. 2011;12(12):236.

de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet*. 2011;7(12):e1002384.

Dennis PJ, Brenner DJ, Thacker WL, Wait R, Vesey G, Steigerwalt AG, Benson RF. Five new *Legionella* species isolated from water. *Int J Syst Bacteriol.* 1993;43:329-337.

DH, 2010: Department of Health: Health Protection Legislation (England) Guidance 2010.

Diederen BM, Peeters MF. Evaluation of the SAS *Legionella* Test, a new immunochromatographic assay for the detection of *Legionella pneumophila* serogroup 1 antigen in urine. *Clin Microbiol Infect.* 2007;13:86–88.

Dominguez J, Gali N , Matas L , Pedroso P , Hernandez A , Padilla E , Ausina V . Evaluation of a rapid immunochromatographic assay for the detection of *Legionella* antigen in urine samples. *Eur J Clin Microbiol Infect Dis.* 1999;18:896–898.

Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. Culture-independent detection and characterisation of Mycobacterium tuberculosis and *M. africanum* in sputum samples using shotgun metagenomics on a benchtop sequencer. *PeerJ*. 2014;2:e585.

Dournon E. Isolation of legionellae from clinical specimens. In: Harrison TG, Taylor AG, eds. A laboratory manual for *Legionella*. London, United Kingdom, John Wiley & Sons Ltd, 1988;13–30.

Dowling JN, Kroboth FJ, Karpf M, Yee RB, Pasculle AW. Pneumonia and multiple lung abscesses caused by dual infection with *Legionella micdadei* and *Legionella pneumophila*. *Am Rev Respir Dis.* 1983;127:121-125.

Doyle RM, Burgess C, Williams R, Gorton R, Booth H, Brown J, Bryant JM, Chan J, Creer D, Holdstock J, Kunst H, Lozewicz S, Platt G, Romero EY, Speight G, Tiberi S, Abubakar I, Lipman M, McHugh TD, Breuer J. Direct whole-genome sequencing of sputum accurately identifies drug-resistant *Mycobacterium tuberculosis* faster than MGIT culture sequencing. *J Clin Microbiol.* 2018; 56:e00666-18.

Dreyfus LA, Iglewski BH. Conjugation-mediated genetic exchange in *Legionella pneumophila*. Journal of Bacteriology. 1985;161(1):80–4.

ECDC, 2017: European Centre for Disease Prevention and Control. Legionnaires' disease in Europe, 2015. Stockholm: ECDC; 2017.

ECDC, 2019: European Centre for Disease Prevention and Control. Legionnaires' disease. In: ECDC. Annual epidemiological report for 2017. Stockholm: ECDC; 2019.

Edelstein PH. Improved semiselective medium for isolation of *Legionella pneumophila* from contaminated clinical and environmental specimens. *J Clin Microbiol.* 1981;14:298–303.

Edelstein PH, Brenner DJ, Moss CW, Steigerwalt AG, Francis EM, George WL. *Legionella wadsworthii* species nova: a cause of human pneumonia. *Ann Intern Med.* 1982;97:809-813.

Edelstein PH. The laboratory diagnosis of Legionnaires' disease. *Semin Respir Infect.* 1987;2:235–241.

Edelstein PH, Edelstein MA, Shephard LJ, Ward KW, Ratcliff RM. *Legionella steelei* sp. nov., isolated from human respiratory specimens in California, USA, and South Australia. *Int J Syst Evol Microbiol.* 2012;62:1766-1771.

Edelstein PH. *Legionella jamestowniensis* fatal pneumonia in an immunosuppressed man. *J Infect Chemother.* 2017;23(1):59-61.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. The Pfam Protein Families Database in 2019. *Nucleic Acids Res*. 2019;47:D427-D432.

Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ.* 2015;3:e1319

ESR, 2017: The Institute of Environmental Science and Research Ltd. Notifiable Diseases in New Zealand: Annual Report 2017 Porirua, New Zealand ISSN: 1179-3058.

European Technical Guidelines, 2017: minimising the risk from *Legionella* infections in building water systems.

Evans GE, Murdoch DR, Anderson TP, Potter HC, George PM, Chambers ST. Contamination of Qiagen DNA extraction kits with *Legionella* DNA. *J Clin Microbiol.* 2003;41(7):3452–3453.

Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, Kraemer MUG, Hill SC, Black A, da Costa AC, Franco LC, Silva SP, Wu CH, Raghwani J, Cauchemez S, du Plessis L, Verotti MP, de Oliveira WK, Carmo EH, Coelho GE, Santelli ACFS, Vinhal LC, Henriques CM, Simpson JT, Loose M, Andersen KG, Grubaugh ND, Somasekar S, Chiu CY, Muñoz-Medina JE, Gonzalez-Bonilla CR, Arias CF, Lewis-Ximenez LL, Baylis SA, Chieppe AO, Aguiar SF, Fernandes CA, Lemos PS, Nascimento BLS, Monteiro HAO, Siqueira IC, de Queiroz MG, de Souza TR, Bezerra JF, Lemos MR, Pereira GF, Loudal D, Moura LC, Dhalia R, França RF, Magalhães T, Marques ET Jr, Jaenisch T, Wallau GL, de Lima MC, Nascimento V, de Cerqueira EM, de Lima MM, Mascarenhas DL, Neto JPM, Levin AS, Tozetto-Mendoza TR, Fonseca SN, Mendes-Correa MC, Milagres FP, Segurado A, Holmes EC, Rambaut A, Bedford T, Nunes MRT, Sabino EC, Alcantara LCJ, Loman NJ, Pybus OG. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature*. 2017;546(7658):406-410.

Federhen S. Type material in the NCBI taxonomy database. *Nucleic Acids Res.* 2015;43: D1086–1098.

Feehery GR, Yigit E, Oyola SO, Langhorst BW, Schmidt VT, Stewart FJ, Dimalanta ET, Amaral-Zettler LA, Davis T, Quail MA, Pradhan S. A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PLoS One*. 2013;8(10):e76096.

Feeley JC, Gibson RJ, Gorman GW, Langford NC, Rasheed JK, Mackel DC, Baine WB. Charcoal-yeast extract agar: primary isolation medium for *Legionella pneumophila*. *J Clin Microbiol.* 1979;10:437–441.

Fields BS. 2005. Procedures for the recovery of *Legionella* from the environment. Respiratory Disease Laboratory Section, US Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, Atlanta, GA.

Finn RD, Clements J, Eddy SR. HMMER Web Server: Interactive Sequence Similarity Searching. *Nucleic Acids Res*. 2011;39:W29-37.

Fisman DN, Lim S, Wellenius GA, Johnson C, Britz P, Gaskins M, Maher J, Mittleman MA, Spain CV, Haas CN, Newbern C. It’s not the heat, it’s the humidity: wet weather increases legionellosis risk in the greater Philadelphia metropolitan area. *J Infect Dis.* 2005;192:2066–2073.

Flendrie M, Jeurissen M, Franssen M, Kwa D, Klaassen C, Vos F. Septic arthritis caused by *Legionella dumoffii* in a patient with systemic lupus erythematosus-like disease. *J Clin Microbiol.* 2011;49(2):746-749.

Fliermans CB, Cherry WB, Orrison LH, Thacker L. Isolation of *Legionella pneumophila* from non-epidemic-related aquatic habitats. *Appl Environ Microbiol.* 1979;37:1239-1242.

Fliermans CB, Cherry WB, Orrison LH, Smith SJ, Tison DL, Pope DH. Ecological distribution of *Legionella pneumophila*. *Appl. Environ. Microbiol.* 1981;41:9-16.

Fliersmans CB, Cherry WB, Orrison LH, Smith SJ, Tison DL, Pope DH. Ecological Distribution of *Legionella pneumophila. Appl. Environ. Microbiol.* 2015;41:9-16.

Fraser DW, Tsai TR, Orenstein W, Parkin WE, Beecham HJ, Sharrar RG, Harris J, Mallison GF, Martin SM, McDade JE, Shepard CC, Brachman PS. Legionnaires' disease: description of an epidemic of pneumonia. *N Engl J Med.* 1977;1189–1197.

Fry NK, Rowbotham TJ, Saunders NA, Embley TM. Direct amplification and sequencing of the 16S ribosomal DNA of an intracellular *Legionella* species recovered by amoebal enrichment from the sputum of a patient with pneumonia. *FEMS Microbiol Lett.* 1991:1;67(2):165-168.

Fumarola D, Miragliotta C, Logroscino G, Castellani Pastoris M. Simultaneous infection with *Legionella pneumophila* and *Legionella micdadei* in an immunologically intact host. A case report. *Boll. Ist. Sieroter. Milan.* 1984;63:165-166.

Gaia V, Fry NK, Afshar B, Lück PC, Meugnier H, Etienne J, Peduzzi R, Harrison TG. Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila*. *J Clin Microbiol.* 2005;43(5):2047-2052.

García-Fulgueiras A, Navarro C, Fenoll D, García J, González-Diego P, Jiménez-Buñuales T, Rodriguez M, Lopez R, Pacheco F, Ruiz J, Segovia M, Balandrón B, Pelaz C. Legionnaires' disease outbreak in Murcia, Spain. *Emerg Infect Dis*. 2003;9(8):915-921.

Ginevra C, Lopez M, Forey F, Reyrolle M, Meugnier H, Vandenesch F, Etienne J, Jarraud S, Molmeret M. Evaluation of a nested-PCR-derived sequence-based typing method applied directly to respiratory samples from patients with Legionnaires' disease. *J Clin Microbiol.* 2009;47(4):981-987.

Glick TH, Gregg MB, Berman B, Mallison G, Rhodes WW Jr, Kassanoff I. Pontiac fever. An epidemic of unknown etiology in a health department. Clinical and epidemiologic aspects. *Am. J. Epidemiol.* 1978;107:149-160.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* 2009; 27:182–9.

Gomez-Valero L, Rusniok C, Jarraud S, Vacherie B, Rouy Z, Barbe V, Medigue C, Etienne J, Buchrieser C. Extensive recombination events and horizontal gene transfer shaped the *Legionella pneumophila* genomes. *BMC Genomics*. 2011;12:536.

Gomez-Valero L, Rusniok C, Cazalet C, Buchrieser C. Comparative and functional genomics of *legionella* identified eukaryotic like proteins as key players in host-pathogen interactions. *Front Microbiol.* 2011;28;2:208. (a)

Gomez-Valero L, Rusniok C, Carson D, Mondino S, Pérez-Cobas AE, Rolando M, Pasricha S, Reuter S, Demirtas J, Crumbach J, Descorps-Declere S, Hartland EL, Jarraud S, Dougan G, Schroeder GN, Frankel G, Buchrieser C. More than 18,000 effectors in the *Legionella* genus genome provide multiple, independent combinations for replication in human cells. *Proc Natl Acad Sci U S A.* 2019;116(6):2265-2273.

Gordon A, 2009. FASTX-Toolkit. Github https://github.com/agordon/fastx_toolkit.

Gordon M, Yakunin E, Valinsky L, Chalifa-Caspi V, Moran-Gilad J. A bioinformatics tool for ensuring the backwards compatibility of *Legionella pneumophila* typing in the genomic era. *Clin Microbiol Infect.* 2017;23(5):306–10.

Gorman GW, Feeley JC, Steigerwalt A, Edelstein PH, Moss CW, Brenner DJ. *Legionella anisa*: a new species of *Legionella* isolated from potable waters and a cooling tower. *Appl Environ Microbiol.* 1985;49:305-309.

Greay TL, Gofton AW, Paparini A, Ryan UM, Oskam CL, Irwin PJ. Recent insights into the tick microbiome gained through next-generation sequencing. *Parasit Vectors*. 2018;4;11(1):12.

Griffith ME, Lindquist DS, Benson RF, Thacker WL, Brenner DJ, Wilkinson HW. First isolation of *Legionella gormanii* from human disease. *J Clin Microbiol.* 1988;26(2):380-381.

Gubler JGH, Schorr M, Gaia V, Zbinden R, Altwegg M. Recurrent soft tissue abscesses caused by *Legionella cincinnatiensis. J Clin Microbiol.* 2001;39:4568–4570.

Gupta SK, Shin H, Han D, Hur HG, Unno T. Metagenomic analysis reveals the prevalence and persistence of antibiotic- and heavy metal-resistance genes in wastewater treatment plant. *J Microbiol.* 2018;56(6):408-415.

Guyot S, Goy JJ, Gersbach P, Jaton K, Blanc DS, Zanetti G. *Legionella* pneumonia aortitis in a heart transplant recipient. *Transpl Infect Dis*. 2007;9:58-59.

Gyi JI, Lane AN, Conn GL, Brown T. The orientation and dynamics of the C2'-OH and hydration of RNA and DNA.RNA hybrids. *Nucleic Acids Res.* 1998;26(31):3104-3110.

Hampton-Marcell JT, Lopez JV, Gilbert JA. The human microbiome: an emerging tool in forensics. *Microb. Biotechnol.* 2017;10:228–230.

Han XY, Ihegword A, Evans SE, Zhang J, Li L, Cao H, Tarrand JJ, El-Kweifi O. Microbiological and Clinical Studies of Legionellosis in 33 Patients with Cancer. *J Clin Microbiol.* 2015;53(7):2180-2187.

Harb OS, Venkataraman C, Haack BJ, Gao LY, Kwaik YA. Heterogeneity in the attachment and uptake mechanisms of the Legionnaires' disease bacterium, *Legionella pneumophila*, by protozoan hosts. *Appl Environ Microbiol.* 1998;64:126–132.

Harrison T , Uldum S , Alexiou-Daniel S , Bangsborg J , Bernander S , Draŝar V , Etienne J , Helbig J , Lindsay D , Lochman I , Marques T , de Ory F , Tartakovskii I , Wewalka G , Fehrenbach F . A multicenter evaluation of the Biotest *Legionella* urinary antigen EIA. *Clin Microbiol Infect.* 1998;4:359–365.

Harrison TG, Afshar B, Doshi N, Fry NK, Lee JV. Distribution of *Legionella pneumophila* serogroups, monoclonal antibody subgroups and DNA sequence types in recent clinical and environmental isolates from England and Wales (2000-2008). *Eur J Clin Microbiol Infect Dis.* 2009;28(7):781-791.

Hasan MR, Rawat A, Tang P, Jithesh PV, Thomas E, Tan R, Tilley P. Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing. *J Clin Microbiol*. 2016;54(4):919-927.

Hayes-Phillips D, Bentham R, Ross K, Whiley H. Factors Influencing *Legionella* Contamination of Domestic Household Showers. *Pathogens.* 2019;26;8(1):pii:E27.

Hebert GA, Steigerwalt AG, Brenner DJ. *Legionella micdadei* species nova: classification of a third species of *Legionella* associated with human pneumonia. *Curr Microbiol.* 1980;3:255-257.

Hebert GA, Moss CW, McDougal LK, Bozeman FM, McKinney RM, Brenner DJ. The rickettsia-like organisms TATLOCK (1943) and HEBA (1959): bacteria phenotypically similar to but genetically distinct from *Legionella pneumophila* and the WIGA bacterium. *Ann. Intern. Med.* 1980;92:45-52.

Helbig JH, Uldum SA , Luck PC , Harrison TG. Detection of *Legionella pneumophila* antigen in urine samples by the BinaxNOW immunochromatographic assay and comparison with both Binax *Legionella* Urinary Enzyme Immunoassay (EIA) and Biotest *Legionella* Urine Antigen EIA. *J Med Microbiol.* 2001;50:509–516.

Helbig JH, Bernander S, Castellani Pastoris M, Etienne J, Gaia V, Lauwers S, Lindsay D, Lück PC, Marques T, Mentula S, Peeters MF, Pelaz C, Struelens M, Uldum SA, Wewalka G, Harrison TG. Pan-European study on culture-proven Legionnaires' disease: distribution of *Legionella pneumophila* serogroups and monoclonal subgroups. *Eur J Clin Microbiol Infect Dis.* 2002;21(10):710-716.

Heriot WJ, Mack HG, Stawell R. Ocular involvement in a patient with *Legionella longbeachae* 1 infection. *Clin Exp Ophthalmol.* 2014;42(5):497-499.

Herwaldt LA, Gorman GW, McGrath T, Toma S, Brake B, Hightower AW, Jones J, Reingold AL, Boxer PA, Tang PW *et al.,* A new *Legionella* species, *Legionella feeleii* species nova, causes Pontiac fever in an automobile plant. *Ann*. *Int. Med*, 1984;100:333-338.

Heuner K, Albert-Weissenberger C. The flagellar regulon of *Legionella pneumophila* and the expression of virulence traits in Legionella: Molecular Microbiology, eds Heuner K, Swanson MS (Norfolk, UK: Horizon Scientific Press) 2008;249.

Hicks LA, Rose CE, Jr, Fields BS, Drees ML, Engel JP, Jenkins PR, Rouse BS, Blythe D, Khalifah AP, Feikin DR, Whitney CG. Increased rainfall is associated with increased risk for legionellosis. *Epidemiol Infect*. 2007;135:811–817.

Hoffmann C, Harrison CF, Hilbi H. The natural alternative: protozoa as cellular models for Legionella infection. *Cell Microbiol.* 2014;16(1):15-26.

Hong C, Manimaran S, Shen Y, Perez-Rogers JF, Byrd AL, Castro-Nallar E, Crandall KA, Evan Johnson W. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome.* 2014;2:33.

Hong DK, Blauwkamp TA, Kertesz M, Bercovici S, Truong C, Banaei N. Liquid biopsy for infectious diseases: sequencing of cell-free plasma to detect pathogen DNA in patients with invasive fungal disease. *Diagn Microbiol Infect Dis.* 2018;92(3):210-213.

Hookey JV, Saunders NA, Fry NK, Birtles RJ, Harrison TG. Phylogeny of *Legionellaceae* based on small-subunit ribosomal DNA sequences and proposal of *Legionella lytica* comb. nov. for *Legionella*-like amoebal pathogens. *Int J Syst Bacteriol.* 1996;46:526-531.

Horbach I, Naumann D, Fehrenbach FJ. Simultaneous infections with different serogroups of *Legionella pneumophila* investigated by routine methods and Fourier transform infrared spectroscopy. *J Clin Microbiol.* 1988;26(6):1106-1110.

Horwitz MA. Formation of a novel phagosome by the Legionnaires' disease bacterium (*Legionella pneumophila*) in human monocytes. *J Exp Med.* 1983;**158:** 1319–1331.

Horwitz MA. Phagocytosis of the legionnaires' disease bacterium (*Legionella pneumophila*) occurs by a novel mechanism: Engulfment within a Pseudopod coil. *Cell.* 1984;36:27–33.

HPS (Health Protection Scotland) Weekly Report. Surveillance Report: Legionellosis in Scotland: 2015 – 2016. Volume 51. No. 2017/34.

Huang AD, Luo C, Pena-Gonzalez A, Weigand MR, Tarr CL, Konstantinidis KT. Metagenomics of Two Severe Foodborne Outbreaks Provides Diagnostic Signatures and Signs of Coinfection Not Attainable by Traditional Methods. *Appl Environ Microbiol.* 2017;83(3):e02577-16.

Hughes MS, Steele TW. Occurrence and distribution of *Legionella* species in composted plant materials. *Appl Environ Microbiol.* 1994;60(6):2003-2005.

Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* 2014;6(11):90.

Isberg RR, O'Connor TJ, Heidtman M. The *Legionella pneumophila* replication vacuole: making a cosy niche inside host cells. *Nat Rev Microbiol.* 2009;7:13–24.

Ishimuru N, Suzuki H, Tokuda Y, Takanu T. Severe *Legionella* disease with pneumonia and biopsy confirmed myocarditis most likely caused by *Legionella pneumophilia* serogroup 6. *Intern Med.* 2012;51:3207-3212.

Ishizaki N, Sogawa K, Inoue H, Agata K, Edagawa A, Miyamoto H, Fukuyama M, Furuhata K. *Legionella thermalis* sp. nov., isolated from hot spring water in Tokyo, Japan. *Microbiol Immunol*. 2016; 60:203-208.

Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. Improved data analysis for the MinION nanopore sequencer. *Nat Methods.* 2015;12(4):351-356.

Jefferys AJ, Wilson V & Thein SL. Hypervariable 'mini- satellite' regions in human DNA. *Nature.* 1985; 314: 67–73.

Joly JR, Déry P, Gauvreau L, Coté L, Trépanier C. Legionnaires' disease caused by *Legionella dumoffii* in distilled water. *CMAJ.* 1986;135(11):1274-7.

Jones MB, Highlander SK, Anderson EL, Li W, Dayrit M, Klitgord N, Fabani MM, Seguritan V, Green J, Pride DT, Yooseph S, Biggs W, Nelson KE, Venter JC. Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc Natl Acad Sci U S A.* 2015;112(45):14024-14029.

Joseph SJ, Cox D, Wolff B, Morrison SS, Kozak-Muiznieks NA, Frace M, Didelot X, Castillo-Ramirez S, Winchell J, Read TD, Dean D. Dynamics of genome change among *Legionella* species. *Sci Rep.* 2016;6:33442.

Kafetzopoulou LE, Pullan ST, Lemey P, Suchard MA, Ehichioya DU, Pahlmann M, Thielebein A, Hinzmann J, Oestereich L, Wozniak DM, Efthymiadis K, Schachten D, Koenig F, Matjeschk J, Lorenzen S, Lumley S, Ighodalo Y, Adomeh DI, Olokor T, Omomoh E, Omiunu R, Agbukor J, Ebo B, Aiyepada J, Ebhodaghe P, Osiemi B, Ehikhametalor S, Akhilomen P, Airende M, Esumeh R, Muoebonam E, Giwa R, Ekanem A, Igenegbale G, Odigie G, Okonofua G, Enigbe R, Oyakhilome J, Yerumoh EO, Odia I, Aire C, Okonofua M, Atafo R, Tobin E, Asogun D, Akpede N, Okokhere PO, Rafiu MO, Iraoyah KO, Iruolagbe CO, Akhideno P, Erameh C, Akpede G, Isibor E, Naidoo D, Hewson R, Hiscox JA, Vipond R, Carroll MW, Ihekweazu C, Formenty P, Okogbenin S, Ogbaini-Emovon E, Günther S, Duraffour S. Metagenomic sequencing at the epicentre of the Nigeria 2018 Lassa fever outbreak. *Science*. 2019;363(6422):74-77.

Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*. 2015;27;3:e1165.

Kawashima EH, Farinelli L, Mayer P. (2005-05-12). ”Patent: Method of nucleic acid amplification”.

Kayani MUR, Doyle SM, Sangwan N, Wang G, Gilbert JA, Christner BC, Zhu TF. Metagenomic analysis of basal ice from an Alaskan glacier. *Microbiome*. 2018;5;6(1):123.

Khelef N, Shuman HA, Maxfield FR. Phagocytosis of wild-type *Legionella pneumophila* occurs through a wortmannin-insensitive pathway. *Infect Immun.* 2001;69**:** 5157–5161.

Kilborn JA, Manz LA, O'Brien M, Douglass MC, Horst HM, Kupin W, Fisher EJ. Necrotizing cellulitis caused by *Legionella micdadei. Am J Med.* 1992;92:104–106.

Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Res.* 2016;26(12):1721–1729.

Koide M, Saito A, Okazaki M, Umeda B, Benson RF. Isolation of *Legionella longbeachae* serogroup 1 from potting soils in Japan. *Clin Infect Dis.* 1999;29(4):943-944.

König C, Hebestreit H, Valenza G, Abele-Horn M, Speer CP. *Legionella waltersii*-a novel cause of pneumonia? *Acta Paediatr.* 2005;94(10):1505-1507.

Kozak-Muiznieks NA, Morrison SS, Mercante JW, Ishaq MK, Johnson T, Caravas J, Lucas CE, Brown E, Raphael BH, Winchell JM. Comparative genome analysis reveals a complex population structure of *Legionella pneumophila* subspecies. *Infect Genet Evol.* 2018;59:172-185.

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol*. 2013 Sep;79(17):5112-20.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*. 2019;pii: btz305.

Kroeger ME, Delmont TO, Eren AM, Meyer KM, Guo J, Khan K, Rodrigues JLM, Bohannan BJM, Tringe SG, Borges CD, Tiedje JM, Tsai SM, Nüsslein K. New Biological Insights Into How Deforestation in Amazonia Affects Soil Microbial Communities Using Metagenomics and Metagenome-Assembled Genomes. *Front Microbiol.* 2018;23;9:1635.

Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol.* 2016; 33(7):1870-1874.

Kuroki H, Miyamoto H, Fukuda K, Iihara H, Kawamura Y, Ogawa M, Wang Y, Ezaki T, Taniguchi H. *Legionella impletisoli* sp. nov. and *Legionella yabuuchiae* sp. nov., isolated from soils contaminated with industrial wastes in Japan. *Syst Appl Microbiol.* 2007;30: 273-279.

Kusnetsov J, Neuvonen LK, Korpio T, Uldum SA, Mentula S, Putus T, Tran Minh NN, Martimo KP. Two Legionnaires' disease cases associated with industrial waste water treatment plants: a case report. *BMC Infect Dis.* 2010;2;10:343.

Kwaik YA, Gao L-Y, Stone BJ, Venkataraman C, Harb O.S. Invasion of protozoa by *Legionella pneumophila* and its role in bacterial ecology and pathogenesis. *Appl Environ Microbiol*. 1998;64:3127–3133.

Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie2. *Nat Methods.* 2012;9(4):357-359.

La Scola B, Michel G, Raoult D Isolation of *Legionella pneumophila* by centrifugation of shell vial cell cultures from multiple liver and lung abscesses. *J Clin Microbiol.* 1999;37:785–787.

La Scola B, Birtles RJ, Greub G, Harrison TJ, Ratcliff RM, Raoult D. *Legionella drancourtii* sp. nov., a strictly intracellular amoebal pathogen. *Int J Syst Evol Microbiol.* 2004;54:699-703.

Last AR, Pickering H, Roberts C, Coll F, Phelan J, Burr SE, Cassama E, Nabicassa M, Thomson NR, Holland MJ. Population-based analysis of ocular Chlamydia trachomatis in trachoma- endemic West African communities identifies genomic markers of disease severity. *Genome Medicine* 2018; 10:15.

Lavania M, Singh I, Turankar RP, Ahuja M, Pathak V, Sengupta U, Das L, Kumar A, Darlong J, Nathan R, Maseey A. Molecular detection of multidrug-resistant *Mycobacterium leprae* from Indian leprosy patients*. J Glob Antimicrob Resist.* 2018; 12:214-219.

Leo S., Gaïa N., Ruppé E., Emonet S., Girard M., Lazarevic V. Detection of bacterial pathogens from broncho-alveolar lavage by next-generation sequencing. *Int J Mol Sci.* 2017;18:2011.

Leonard SR, Mammel MK, Lacher DW, Elkins CA. Application of metagenomic sequencing to food safety: detection of Shiga Toxin-producing *Escherichia coli* on fresh bagged spinach. *Appl Environ Microbiol.* 2015;81(23):8183-8191.

Lesnik EA & Freier SM. Relative Thermodynamic Stability of DNA, RNA and DNA:RNA Hybrid Duplexes: Relationship with Base Composition and Structure. *Biochemistry*. 1995;34(34):10807-10815.

Lettinga KD, Verbon A, Nieuwkerk PT, Jonkers RE, Gersons BP, Prins JM, Speelman P. Health-related quality of life and posttraumatic stress disorder among survivors of an outbreak of Legionnaires disease. *Clin Infect Dis.* 2002;1;35(1):11-17.

Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*. 2003;31;299(5607):682-686.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics.* 2009;25(16):2078-9.

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, *Bioinformatics*. 2011;27(21):2987-2993.

Li L, Faucher SP. The membrane protein LasM promotes the culturability of *Legionella pneumophila* in Water. *Front Cell Infect Microbiol.* 2016;6:113.

Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094-3100.

Littrup P, Madsen JK, Lind K. Aortic valve endocarditis associated with *Legionella* infection after *Mycoplasma pneumonia*. *Br Heart J.* 1987;58(3):293-5.

Liu R, Yu Z, Guo H, Liu M, Zhang H, Yang M. Pyrosequencing analysis of eukaryotic and bacterial communities in faucet biofilms. *Sci Total Environ.* 2012;1;435-436:124-31.

Liu G, Weston CQ, Pham LK, Waltz S, Barnes H, King P, Sphar D, Yamamoto RT, Forsyth RA. Epigenetic segregation of microbial genomes from complex samples using restriction endonucleases HpaII and McrB. *PLoS One*. 2016;11:e0146064

Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. *Genome Med.* 2016;8:51.

Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ-M, Quick J, Weir JC, Quince C, Smith GP, Betley, JR, Aepfelbacher M, Pallen MJ. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA*. 2013;309(14):1502–10.

Lo Presti F, Riffard S, Vandenesch F, Reyrolle M, Ronco E, Ichai P, Etienne J. The first clinical isolate of *Legionella parisiensis*, from a liver transplant patient with pneumonia. *J Clin Microbiol.* 1997;35(7):1706-1709.

Lo Presti F, Riffard S, Meugnier H, Reyrolle M, Lasne Y, Grimont PAD, Grimont F, Vandenesch F, Etienne J, Fleurette J, Freney J. *Legionella taurinensis* sp. nov., a new species antigenically similar to *Legionella spiritensis*. *Int J Syst Bacteriol.* 1999;49:397-403.

Lo Presti F, Riffard S, Meugnier H, Reyrolle M, Lasne Y, Grimont PAD, Grimont F, Benson RF, Brenner DJ, Steigerwalt AG, Etienne J, Freney J. *Legionella gresilensis* sp. nov. and *Legionella beliardensis* sp. nov., isolated from water in France. *Int J Syst Evol Microbiol.* 2001;51:1949-1957.

Loridant S, Lagier JC, La Scola B. Identification of *Legionella feeleii* cellulitis. *Emerg Infect Dis*. 2011;17(1):145–146.

Lou X, Hou Y, Liang D, et al. A novel *Alu*-based real-time PCR method for the quantitative detection of plasma circulating cell-free DNA: sensitivity and specificity for the diagnosis of myocardial infarction. *Int J Mol Med*. 2014;35(1):72-80.

Lowry PW, Blankenship RJ, Gridley W, Troup NJ, Tompkins LS. A cluster of sternal wound infections due to postoperative topical exposure to contaminated tap water. *N Engl J Med.* 1991;324:109–113.

Lowry PW, Tompkins LS. Nosocomial legionellosis: a review of pulmonary and extrapulmonary syndromes. *Am J Infect Control.* 1993;21(1):21-27.

Lück PC, Bender L, Ott M, Helbig JH, Hacker J. Analysis of *Legionella pneumophila* serogroup 6 strains isolated from a hospital warm water supply over a three-year period by using genomic long-range mapping techniques and monoclonal antibodies. *Appl Environ Microbiol.* 1991;57(11):3226-3231.

Luck PC, Jacobs E, Roske I, Schroter-Bobsin U, Dumke R, Grownow S. *Legionella dresdenensis* sp. nov., isolated from river water. *Int J Syst Evol Microbiol.* 2010;60:2557-2562.

Lück C, Fry NK, Jürgen HH, Jarraud S, Harrison TG, Buchrieser C, Hilbi H (eds.), Legionella: Methods and Protocols, Methods in Molecular Biology, vol. 954, 001 10.1007/978-1-62703-161-5_6, © Springer Science+Business Media New York 2013.

Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. ConStrains identifies microbial strains in metagenomic datasets. *Nat Biotechnol.* 2015;33:1045.

Mahoney FJ, Hoge CW, Farley TA, Barbaree JM, Breiman RF, Benson RF, McFarland LM. Communitywide outbreak of Legionnaires' disease associated with a grocery store mist machine. *J Infect Dis.* 1992;165**:**736-739.

Marrie TJ, Raoult D, La Scola B, Birtles RJ, de Carolis E; Canadian Community-Acquired Pneumonia Study Group. *Legionella*-like and other amoebal pathogens as agents of community-acquired pneumonia. *Emerg Infect Dis.* 2001;7(6):1026-1029.

Marston BJ, Lipman HB, Breiman RF. Surveillance for Legionnaires' disease. Risk factors for morbidity and mortality. *Arch Intern Med*. 1994;154:2417–2422.

Martinelli F, Carasi S, Scarcella C, Speziani F. Detection of *Legionella pneumophila* at thermal spas. *New Microbiol.* 2001;24(3):259-264.

Matsui M, Fujii S, Shiroiwa R, Amemura-Maekawa J, Chang B, Kura F, Yamauchi K. Isolation of *Legionella rubrilucens* from a pneumonia patient co-infected with *Legionella pneumophila*. *J Med Microbiol.* 2010;59:1242-1246.

McAdam PR, Vander Broek CW, Lindsay DS, Ward MJ, Hanson MF, Gillies M, Watson M, Stevens JM, Edwards GF, Fitzgerald JF. Gene flow in environmental *Legionella pneumophila* leads to genetic and pathogenic heterogeneity within a Legionnaires' disease outbreak. *Genome Biol*. 2014;15(11):504.

McClelland MR, Vaszar LT, Kagawa FT. Pneumonia and osteomyelitis due to *Legionella longbeachae* in a woman with systemic lupus erythematosus. *Clin Infect Dis.* 2004;38:e102–106.

McDade JE, Brenner DJ, Bozeman FM. Legionnaires' Disease Bacterium Isolated in 1947. *Ann Intern Med.* 1979;90**:** 659–661.

McKinney RM, Porschen RK, Edelstein PH, Bissett ML. Harris PP, Bondell SP, Steigerwalt AG, Weaver RE, Ein ME, Lindquist DS, Kops RS, Brenner DJ. *Legionella longbeachae* species nova, another etiologic agent of human pneumonia. *Ann Intern Med.* 1981;94:739-743.

McMurdie and Holmes (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE. 8(4):e61217

McNally C, Hackman B, Fields BS, Plouffe JF. Potential importance of Legionella species as etiologies in community acquired pneumonia (CAP). *Diagn Microbiol Infect Dis.* 2000;38(2):79-82.

Marks M, Fookes M, Wagner J, Ghinai R, Sokana O, Sarkodie YA, Solomon AW, Mabey DCW, Thomson NR.Direct Whole-Genome Sequencing of Cutaneous Strains of Haemophilus ducreyi. *Emerg Infect Dis*. 2018;24(4):786-9.

Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome*. 2018;6(1):42.

Matsui, M. *et al.,* Isolation of *Legionella rubrilucens* from a pneumonia patient co-infected with *Legionella pneumophila*. *J Med Microbiol.* 2010;59**:**1242–1246.

Mentasti M, Fry NK, Afshar B, Palepou-Foxley C, Naik FC, Harrison TG. 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. *Eur J Clin Microbiol Infect Dis.* 2012;31(8):2017-2028.

Mentasti M, Underwood A, Lück C, Kozak-Muiznieks NA, Harrison TG, Fry NK. Extension of the *Legionella pneumophila* sequence-based typing scheme to include strains carrying a variant of the N-acylneuraminate cytidylyltransferase gene. *Clin Microbiol Infect.* 2014;20(7):O435-441.

Mentasti M, Afshar B, Collins S, Walker J, Harrison TG, Chalker V. Rapid investigation of cases and clusters of Legionnaires' disease in England and Wales using direct molecular typing. *J Med Microbiol.* 2016;65(6):484-493.

Mentula S, Pentikäinen J, Perola O, Ruotsalainen E. *Legionella longbeachae* infection in a persistent hand-wound after a gardening accident. *JMM Case Rep*. 2014;1(4):e004374.

Mercante JW, Winchell JM. Current and emerged *Legionella* diagnostics for laboratory and outbreak investigations. *Clin Microbiol Rev.* 2015;28:95-133.

Meyer RD, Edelstein PH, Kirby BD, Louie MH, Mullingan ME, Morgenstein AA, Finegold SM. Legionnaires’ disease: unusual clinical and laboratory features. Ann Intern Med. 1980;93:240-243.

Miller S, Naccache SN, Samayoa E, Messacar K, Arevalo S, Federman S, Stryke D, Pham E, Fung B, Bolosky WJ, Ingebrigtsen D, Lorizio W, Paff SM, Leake JA, Pesano R, DeBiasi R, Dominguez S, Chiu CY. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res.* 2019;29(5):831-842.

Mintz CS, Shuman HA. Transposition of bacteriophage mu in the Legionnaires' disease bacterium. *PNAS.* 1987;84(13):4645–4649.

Mintz CS, Schultz DR, Arnold PI, Johnson W. *Legionella pneumophila* lipopolysaccharide activates the classical complement pathway. *Infect Immun.* 1992;60:2769–2776.

Mintz CS, Arnold PI, Johnson W, Schultz DR. Antibody-independent binding of complement component C1q by *Legionella pneumophila*. *Infect Immun.* 1995;63:4939–4943.

Mitchell RG, Pasvol G, Newnham RS. Pneumonia due to *Legionella bozemanii*: first report of a case in *Europe. J Infect.* 1984;8(3):251-255.

Mizrahi H, Peretz A, Lesnik R, Aizenberg-Gershtein Y, Rodríguez-Martínez S, Sharaby Y, et al. Comparison of sputum microbiome of legionellosis-associated patients and other pneumonia patients: indications for polybacterial infections. *Sci Rep*. 2017;7:40114.

Molmeret M, Horn M, Wagner M, Santic M, Abu Kwaik Y. Amoebae as training grounds for intracellular bacterial pathogens. *Appl Environ Microbiol.* 2005;71(1):20-8.

Moran-Gilad J, Lazarovitch T, Mentasti M, Harrison T, Weinberger M, Mordish Y, Mor Z, Stocki T, Anis E, Sadik C, Amitai Z, Grotto I. Humidifier-associated paediatric Legionnaires' disease, Israel, February 2012. *Euro Surveill.* 2012;11;17(41):20293.

Moran-Gilad J, Prior K, Yakunin E, Harrison TG, Underwood A, Lazarovitch T, Valinsky L, Lueck C, Krux F, Agmon V, Grotto I, Harmsen D. Design and application of a core genome multilocus sequence typing scheme for investigation of Legionnaires' disease incidents. *Euro Surveill.* 2015;20(28):pii:21186.

Morris GK, Steigerwalt A, Feeley JC, Wong ES, Martin WT, Patton CM, Brenner DJ. *Legionella gormanii* sp. nov. *J Clin Microbiol.* 1980;12:718-721.

Muder RR, Yu VL, Woo AH. Mode of transmission of *Legionella pneumophila*. A critical review. *Arch Intern Med.* 1986;146(8):1607-1612.

Munoz M, Martinez Toldos MC, Yague G. Evaluation of three immunochomatographic assays for detection of *Legionella pneumophila* serogroup 1 antigen in urine sample. *de la Sociedad Espanola de Quimioterapia.* 2009;22(4):207–209.

Myerowitz RL, Pasculle AW, Dowling JN, Pazin GJ Sr, Puerzer M, Yee RB, Rinaldo CR Jr, Hakala TR. Opportunistic lung infection due to "Pittsburgh Pneumonia Agent". *N Engl J Med.* 1979;301(18):953-958.

Naik FC, Dabrera G. Legionnaires' Disease in England and Wales 2014. England PH, ed. 2015.

Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* 2016;26:1612-1625.

Nelson DL, Ledbetter SA, Corbo L, Victoria MF, Ramirez-Solis R, Webster TD, Ledbetter DH, Caskey CT. *Alu* polymerase chain reaction: a method for rapid isolation of human-specific sequences from complex DNA sources. *Proc Natl Acad Sci U S A.* 1989;86(17):6686-6690.

Nelson MT, Pope CE, Marsh RL, Wolter DJ, Weiss EJ, Hager KR, Vo AT, Brittnacher MJ, Radey MC, Hayden HS, Eng A, Miller SI, Borenstein E, Hoffman LR. Human and Extracellular DNA Depletion for Metagenomic Analysis of Complex Clinical Infection Samples Yields Optimized Viable Microbiome Profiles. *Cell Rep*. 2019;26(8):2227-2240.e5.

Newton HJ, Ang DKY, van Driel IR, Hartland EL. Molecular Pathogenesis of Infections Caused by *Legionella pneumophila*. *Clin Microbiol Rev.* 2010;23**:**274–298.

Nguyen TM, Ilef D, Jarraud S, Rouil L, Campese C, Che D, Haeghebaert S, Ganiayre F, Marcel F, Etienne J, Desenclos JC. A community-wide outbreak of legionnaires disease linked to industrial cooling towers - how far can contaminated aerosols spread? *J Infect Dis.* 2006;193(1):102-111.

NICE, 2018: National Institute for Health and Care Excellence (NICE). *Respiratory Tract Infections (Self-limiting): Prescribing Antibiotics* NICE Clinical Guideline 69 (Centre for Clinical Practice, 2008).

NIH HMP Working Group. The NIH Human Microbiome Project. Genome Res. 2009;19(12):2317-2323.

Nimmo C, Doyle R, Burgess C, William R, Gorton R, McHugh TD, Brown M, Morris-Jones S, Booth H, Breuer J. Rapid Identification of a *Mycobacterium tuberculosis* full genetic drug resistance profile through whole genome sequencing directly from sputum. *Int J Infect Dis.* 2017;62:44-46.

Nimmo C, Shaw LP, Doyle R, Williams R, Brien K, Burgess C, Breuer J, Balloux F, Pym AS. Whole genome sequencing *Mycobacterium tuberculosis* directly from sputum identifies more genetic diversity than sequencing from culture. *BMC Genomics*. 2019;20(1):389.

NNDSS 2015: Annual Report Working Group. Australia's Notifiable Disease Status, 2015: Annual Report of the National Notifiable Diseases Surveillance Sytem. *Commun Dis Intell. (2018)*. 2019;43.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome Res*. 2017;27(5):824–834.

O'Connor BA, Carman J, Eckert K, Tucker G, Givney R, Cameron S. Does using potting mix make you sick? Results from a *Legionella longbeachae* case-control study in South Australia. *Epidemiol Infect.* 2007;135(1):34-39.

Oh S, Caro-Quintero A, Tsementzi D, DeLeon-Rodriguez N, Luo C, Poretsky R, Konstantinidis KT. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol*. 2011;77(17):6000-6011.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733-745.

Oliva G, Sahr T, Buchrieser C. The life cycle of *L. pneumophila*: Cellular Differentiation is linked to virulence and metabolism. *Front Cell Infect Microbiol.* 2018;8:3.

O'Loughlin RE, Kightlinger L, Werpy MC, Brown E, Stevens V, Hepper C, Keane T, Benson RF, Fields BS, Moore MR. Restaurant outbreak of Legionnaires' disease associated with a decorative fountain: an environmental and case-control study. *BMC Infect. Dis.* 2007;7: 93.

Olsen CW, Elverdal P, Jørgensen CS, Uldum SA. Comparison of the sensitivity of the *Legionella* urinary antigen EIA kits from Binax and Biotest with urine from patients with infections caused by less common serogroups and subgroups of Legionella. *Eur J Clin Microbiol Infect Dis.* 2009;28:817–820.

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 2016;17(1):1–14.

Oppenheim BA, Sefton AM, Gill ON, Tyler JE, O'Mahony MC, Richards JM, Dennis PJ, Harrison TG. Widespread *Legionella pneumophila* contamination of dental stations in a dental school without apparent human infection. *Epidemiol Infect.* 1987;99(1):159-166.

Orrison LH, Cherry WB, Tyndall RL, Fliermans CB, Gough SB, Lambert MA, McDoughal LK, Bibb WF, Brenner DJ. *Legionella oakridgensis*: unusual new species isolated from cooling tower water. *Appl. Environ. Microbiol.* 1983;45:536-545.

Osterholm MT, Chin TD, Osborne DO, Dull HB, Dean AG, Fraser DW, Hayes PS, Hall WN. A 1957 outbreak of Legionnaires' disease associated with a meat packing plant. *Am. J. Epidemiol.* 1983;117:60-67.

PacificBiosciences, 2014. pbh5tools. Github: https://github.com/PacificBiosciences/pbh5tools.

Page AJ, Taylor B, Delaney AJ, Soares J, Seeman T, Keane JA, Harris SR. *SNP-sites*: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom.* 2016;2(4):e000056.

Palmer A, Painter J, Hassler H, Richards VP, Bruce T, Morrison S, Brown E, Kozak-Muiznieks NA, Lucas C, McNealy TL. *Legionella clemsonensis* sp. nov.: a green fluorescing *Legionella* strain from a patient with pneumonia. *Microbiol Immunol.* 2016;60(10):694-701.

Palutke WA, Crane LR, Wentworth BB, Geiger JG, Cardozo L, Singhakowinta A, Bartley J, Robinson BE. *Legionella feeleii*-associated pneumonia in humans. *Am J Clin Pathol.* 1986;86(3):348-51.

Park MY, Ko KS, Lee HK, Park HS, Kook YH. *Legionella busanensis*sp. nov., isolated from cooling tower water in Korea. *Int J Syst Evol Microbiol*. 2003;53:77-80.

Park M, Yun ST, Kim MS, Chun J, Ahn TI. Phylogenetic characterization of Legionella-like endosymbiotic X-bacteria in Amoeba proteus: a proposal for 'Candidatus *Legionella jeonii'* sp. nov. *Environ Microbiol.* 2004;6(12):1252-1263.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015;25(7):1043–1055.

Pasculle AW, Feeley JC, Gibson RJ, Cordes LG, Myerowitz RL, Patton CM, Gorman GW, Carmack CL, Ezzell JW, Dowling JN. Pittsburgh pneumonia agent: direct isolation from human lung tissue. *J Infect Dis.* 1980;141:727-732.

Pearce MM, Theodoropoulos N, Mandel MJ, Brown E, Reed KD, Cianciotto NP. *Legionella cardiaca* sp. nov., isolated from a case of native valve endocarditis in a human heart. *Int J Syst Evol Microbiol.* 2012;62:2946-2954.

Pendleton KM, Erb-Downward JR, Bao Y, Branton WR, Falkowski NR, Newton DW, Huffnagle GB, Dickson RP. Rapid Pathogen Identification in Bacterial Pneumonia Using Real-Time Metagenomics. *Am J Respir Crit Care Med*. 2017;196(12):1610-1612.

Phares CR, Wangroongsarb P, Chantra S, Paveenkitiporn W, Tondella ML, Benson RF, Thacker WL, Fields BS, Moore MR, Fischer J, Dowell SF, Olsen SJ. Epidemiology of severe pneumonia caused by *Legionella longbeachae, Mycoplasma pneumoniae,* and *Chlamydia pneumoniae*: 1-year, population-based surveillance for severe pneumonia in Thailand. *Clin Infect Dis.* 2007;45:e147–155.

PHE, 2016: PHE Legionnaires' disease in residents of England and Wales – 2015 Official Statistics, 2016 PHE publications gateway number: 2016332

PHE, 2017(a): Legionnaires' disease in residents of England and Wales: 2016. Public Health England, Official Statistics. PHE publications gateway number: 2017685.

PHE, 2017(b): Public Health England. (2017). Evaluations, validations and verifications of diagnostic tests. UK Standards for Microbiology Investigations. Q 1 Issue 5.

PHE, 2018: Implementing pathogen genomics: a case study, 2018. PHE publications gateway number: 2018254.

PHE, 2019: Guidance on Investigating Cases, Clusters and Outbreaks of Legionnaires' Disease for Public Health England Health Protection Teams. January 2019. PHE publications gateway number: 2018680.

Phin N, Cresswell T, Parry-Ford F; Incident Control Team. Case of Legionnaires disease in a neonate following a home birth in a heated birthing pool, England, June 2014. *Euro Surveill.* 2014;19(29).

Phin N, Parry-Ford F, Harrison T, Stagg HR, Zhang N, Kumar K, Lortholary O, Zumla A, Abubakar I. Epidemiology and clinical management of Legionnaires' disease. *Lancet Infect Dis*. 2014 Oct;14(10):1011-1021.

Picard Toolkit, 2019. Broad Institute. Github https://broadinstitute.github.io/picard/.

Pinto M, Borges V, Antelo M, Pinheiro M, Nunes A, Azevedo J, Borrego MJ, Mendonça J, Carpinteiro D, Vieira L, Gomes JP. Genome-scale analysis of the non-cultivable Treponema pallidum reveals extensive within-patient genetic variation. *Nat Microbiol*. 2016;2: 16190.

Pravinkumar SJ, Edwards G, Lindsay D, Redmond S, Stirling J, House R, Kerr J, Anderson E, Breen D, Blatchford O, McDonald E, Brown A. A cluster of Legionnaires' disease caused by *Legionella longbeachae* linked to potting compost in Scotland, 2008 – 2009. *Euro Surveill.* 2010;25:15(8):19496.

Qin X, Abe PM, Weissman SJ, Manning SC. Extrapulmonary Legionella micdadei infection in a previously healthy child. *Pediatr Infect Dis J.* 2002;21(12):1174-1176.

Qin T, Yan G, Ren H, Zhou H, Wang H, Xu Y, Zhao M, Guan H, Li M, Shao Z. High Prevalence, Genetic Diversity and Intracellular Growth Ability of *Legionella* in Hot Spring Environments. *PLoS One.* 2013;8(3):e59018.

Qin T, Zhang W, Liu W, Zhou H, Ren H, Shao Z, Lan R, Xu J. Population structure and minimum core genome typing of Legionella pneumophila. *Sci Rep.* 2016;6:21356.

Quero S, Párraga-Niño N, Sabria M, Barrabeig I, Sala MR, Jané M, Mateu L, Sopena N, Pedro-Botet ML, Garcia-Nuñez M. Legionella SBT applied directly to respiratory samples as a rapid molecular epidemiological tool. *Sci Rep.* 2019;24;9(1):623.

Quick J, Loman NJ, Duraffour S, et al., Real-time, portable genome sequencing for Ebola surveillance. *Nature.* 2016;530:228.

Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, Murat Eren A. DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* 2017;18(1):181.

Rambaut A, 2008. FigTree. Github https://github.com/rambaut/figtree

Raphael BH, Baker DJ, Nazarian E, Lapierre P, Bopp D, Kozak-Muiznieks NA, Morrison SS, Lucas CE, Mercante JW, Musser KA, Winchell JM. Genomic Resolution of Outbreak-Associated Legionella pneumophila Serogroup 1 Isolates from New York State. *Appl Environ Microbiol.* 2016;82(12):3582-3590.

Raphael BH, Huynh T, Brown E, Smith JC, Ruberto I, Getsinger L, White S, Winchell JM. Culture of Clinical Specimens Reveals Extensive Diversity of Legionella pneumophila Strains in Arizona. *mSphere*. 2019 Feb 27;4(1):e00649-18.

Ratcliff, RM, Lanser, JA, Manning PA, Heuzenroeder, MW. Sequence-based classification scheme for the genus *Legionella* targeting the mip gene. *J Clin Microbiol*. 1998; 36(6):1560-1567.

Ratzow S, Gaia V, Helbig JH, Fry NK, Lück PC. 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*. J Clin Microbiol.* 2007;45(6):1965-8.

Reuter S, Harrison TG, Köser CU, Ellington MJ, Smith GP, Parkhill J, Peacock SJ, Bentley SD, Török ME. A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak. *BMJ Open*. 2013;9;3(1).pii:e002175.

Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition*. Proc Natl Acad Sci U S A*. 2009;106(45):19126-31.

Rittig MG, Krause A, Häupl T, Schaible UE, Modolell M, Kramer MD, Lütjen-Drecoll E, Simon MM, Burmester GR. Coiling phagocytosis is the preferential phagocytic mechanism for *Borrelia burgdorferi*. *Infect Immun.* 1992;60:4205–4212.

Rizzardi K, Winiecka-Krusnell J, Ramliden M, Alm E, Andersson S, Byfors S. *Legionella norrlandica* sp. nov., isolated from the biopurification systems of wood processing plants. *Int J Syst Evol Microbiol.* 2015;65:598-603.

Rogers J. The origin and evolution of retroposons. *Int Review Cytology* 1985;93:187–279.

Rowbotham TJ. Current views on the relationship between amoebae, legionellae and man*. Isr J Med Sci.* 1986;22:678-689.

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12(1):87.

Samuel V, Bajwa AA, Curry JD. First case of *Legionella pneumophilia* native valve endocarditis*. Int J Infect Dis*. 2011;15:e576-e577.

Sanchez MC, Sebti R, Hassoun P, Mannion C, Goy AH, Feldman T, Mato A, Hong T. Osteomyelitis of the patella caused by *Legionella anisa*. *J Clin Microbiol.* 2013;51(8):2791-2793.

Sánchez-Busó L, Comas I, Jorques G, González-Candelas F. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. *Nat Genet*. 2014;46:1205–1211.

Sanger SF, Nicklen ARC. DNA sequencing with chain terminating. *Proc Natl Acad Sci.* 1977;74:5463–5467

Scaturro M, Fontana S, Ricci ML. Use of nested polymerase chain reaction based on sequence-based typing of clinical samples to determine the source of infection for hospital-acquired Legionnaires' disease. *Infect Control Hosp Epidemiol.* 2011;32(5):510-512.

Schaefer U, 2014. KmerID. Github https://github.com/phe-bioinformatics/kmerid

Schjørring S, Stegger M, Kjelsø C, Lilje B, Bangsborg JM, Petersen RF, David S, Uldum SA; ESCMID Study Group for Legionella Infections (ESGLI). Genomic investigation of a suspected outbreak of Legionella pneumophila ST82 reveals undetected heterogeneity by the present gold-standard methods, Denmark, July to November 2014. *Euro Surveill.* 2017;22;22(25).pii:30558.

Schlaberg *et al.,* 2017(a): Schlaberg R, Chiu CY, Miller S, Procop GW, Weinstock G. Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. *Arch Pathol Lab Med.* 2017;141(6):776-86.

Schlaberg *et al.,* 2017(b): Schlaberg R, Queen K, Simmon K, Tardif K, Stockmann C, Flygare S, Kennedy B, Voelkerding K, Bramley A, Zhang J, Eilbeck K, Yandell M, Jain S, Pavia AT, Tong S, Ampofo K. Viral Pathogen Detection by Metagenomics and Pan-Viral Group Polymerase Chain Reaction in Children With Pneumonia Lacking Identifiable Etiology. *J Infect Dis.* 2017;215(9):1407-1415.

Scholz M, Ward D V, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods.* 2016;13:435.

Seeman T, 2014. mlst. Github https://github.com/tseemann/mlst.

Seeman T, 2014. snippy. Github https://github.com/tseemann/snippy.

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods*. 2012;9(8):811–814.

Shachor-Meyouhas Y, Kassis I, Bamberger E, Nativ T, Sprecher H, Levy I, Srugo I. Fatal hospital-acquired Legionella pneumonia in a neonate. *Pediatr Infect Dis J.* 2010;29(3):280-281.

Sheehan KB, Henson JM, Ferris MH. *Legionella* species diversity in an acidic biofilm community in Yellowstone National Park. *Appl Environ Microbiol.* 2005;71:507-511.
Shelburne SA, Kielhofner MA, Tiwari PS. Cerebellar involvement in legionellosis. *South Med J.* 2004;97(1):61-64.

Shen H, Rogelj S, Kieft TL. Sensitive, real-time PCR detects low-levels of contamination by *Legionella pneumophila* in commercial reagents. *Mol Cell Probes.* 2006;20(3–4):147–153.

Singer M. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell.* 1982;28:433–434.

Sobkowiak B, Glynn JR, Houben RMGJ, Mallard K, Phelan JE, Guerra-Assunção JA, Banda L, Mzembe T, Viveiros M, McNerney R, Parkhill J, Crampin AC, Clark TG. Identifying mixed Mycobacterium tuberculosis infections from whole genome sequence data. *BMC Genomics*. 2018;19(1):613.

Sommese L, Scarfogliero P, Vitiello M, Catalanotti P, Galdiero E. Presence of *Legionella spp*. in thermal springs of the Campania region of south Italy. *New Microbiol.* 1996;19(4):315-320.

Stallworth C, Steed L, Fisher MA, Nolte FS. Legionnaires' disease caused by *Legionella londiniensis*. *J Clin Microbiol.* 2012;50(12):4178-4179.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312-1313.

Steele TW, Lanser J, Sangster N. Isolation of *Legionella longbeachae* serogroup 1 from potting mixes. *Appl Environ Microbiol.* 1990;56:49-53.

Stewart CR, Muthye V, Cianciotto NP. *Legionella pneumophila* persists within biofilms formed by *Klebsiella pneumoniae*, *Flavobacterium* sp., and *Pseudomonas fluorescens* under dynamic flow conditions. *PLoS One.* 2012;7:e50560.

Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, Liachko I, Snelling TJ, Dewhurst RJ, Walker AW, Roehe R, Watson M. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun.* 2018;28;9(1):870.

St-Martin G, Uldum S, Molbak K. 2013. Incidence and prognostic factors for Legionnaires' disease in Denmark 1993-2006. *ISRN Epidemiol.* 2013:8.

Stone BJ, Abu Kwaik Y. Expression of multiple pili by *Legionella pneumophila*: Identification and characterization of a type IV pilin gene and its role in adherence to mammalian and protozoan cells. *Infect Immun.* 1998;66:1768–1775.

Stone BJ, Abu Kwaik Y. Natural competence for DNA transformation by *Legionella pneumophila* and its association with expression of type IV pili. *J Bacteriol.* 1999;181(5):1395–1402.

Strong MJ, Xu G, Morici L, Splinter Bon-Durant S, Baddoo M, Lin Z, Fewell C, Taylor C, Flemington, E. K. Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples. *PLoS Pathogens.* 2014;10(11).

Swanson MS, Isberg RR. Association of Legionella pneumophila with the macrophage endoplasmic reticulum. *Infect Immun.* 1995;63**:**3609–3620.

Tachado SD, Samrakandi MM, Cirillo JD. Non-opsonic phagocytosis of Legionella pneumophila by macrophages is mediated by phosphatidylinositol 3-kinase. *PLoS One.* 2018;3.

Tanabe M, Nakajima H, Nakamura A, Ito T, Nakamura M, Shimono T, Wada H, Shimpo H, Nobori T, Ito M. Mycotic aortic aneurysm associated with *Legionella anisa*. *J Clin Microbiol.* 2009;47(7):2340-2343.

Tang PW, Toma S, Moss CW, Steigerwalt AG, Cooligan TG, Brenner DJ. *Legionella bozemanii* serogroup 2: a new etiological agent*. J Clin Microbiol.* 1984;19(1):30-33.

Tang PW, Toma S, MacMillan LG. *Legionella oakridgensis*: laboratory diagnosis of a human infection. *J Clin Microbiol.* 1985;21(3):462-463.

Tatlock HA Rickettsia-like organism recovered from guinea pigs. *Proc Soc Exp Biol. Med.* 1944;57:95-99.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*. 2000;28(1):33–36.

Teeling H, Glöckner FO. Current opportunities and challenges in microbial metagenome analysis--a bioinformatic perspective. *Brief Bioinform.* 2012;13(6):728-742.

Terranova W, Cohen ML, Fraser DW. 1974 outbreak of Legionnaires' Disease diagnosed in 1977. Clinical and epidemiological features. *Lancet.* 1978;2:122-124.

Thacker SB, Bennett JV, Tsai TF, Fraser DW, McDade JE, Shepard CC, Williams KH Jr, Stuart WH, Dull HB, Eickhoff TC. An outbreak in 1965 of severe respiratory illness caused by the Legionnaires' disease bacterium. *J Infect Dis.* 1978;138:512-519.

Thacker WL, Wilkinson HW, Plikaytis BB, Steigerwalt AG, Mayberry WR, Moss CW, Brenner DJ. Second serogroup of *Legionella feeleii* strains isolated from humans. *J Clin Microbiol.* 1985;22(1):1-4.

Thacker WL, Benson RF, Staneck JL, Vincent SR, Mayberry WR, Brenner DJ, Wilkinson HW. *Legionella cincinnatiensis* sp. nov. isolated from a patient with pneumonia. *J. Clin. Microbiol.* 1988;26:418-420.

Thacker WL, Benson RF, Schifman RB, Pugh E, Steigerwalt AG, Mayberry WR, Brenner DJ, Wilkinson HW. *Legionella tucsonensis* sp. nov. isolated from a renal transplant recipient. *J Clin Microbiol.* 1989;27:1831-1834.

Thacker WL, Benson RF, Hawes L, Gidding H, Dwyer B, Mayberry WR, Brenner DJ. *Legionella fairfieldensis* sp. nov. isolated from cooling tower waters in Australia. *J Clin Microbiol.* 1991;29:475-478.

Thacker WL, Dyke JW, Benson RF, Havlichek Jr DJ, Robinson-Dunn B, Stiefel H, Mayberry WR, Brenner DJ. *Legionella lansingensis* sp. nov. isolated from a patient with pneumonia and underlying chronic lymphocytic leukemia. *J Clin. Microbiol.* 1992;30:2398-2401.

Thoendel M, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, Abdel MP, Patel R. Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. *J Microbiol Methods.* 2016;127:141–145.

Tilney LG, Harb OS, Connelly PS, Robinson CG, Roy CR. 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. *J Cell Sci.* 2001;114:4637–4650.

Timms VJ, Rockett R, Bachmann NL, Martinez E, Wang Q, Chen SC, Jeoffreys N, Howard PJ, Smith A, Adamson S, Gilmour R, Sheppeard V, Sintchenko V. Genome Sequencing Links Persistent Outbreak of Legionellosis in Sydney (New South Wales, Australia) to an Emerging Clone of *Legionella pneumophila* Sequence Type 211*. Appl Environ Microbiol.* 2018;84(5).pii:e02020-17.

Tijet N, Tang P, Romilowych M, Duncan C, Ng V, Fisman DN, Jamieson F, Low DE, Guyard C. 2010. New endemic *Legionella pneumophila* serogroup I clones, Ontario, Canada. *Emerg Infect Dis.* 2010;16:447–454.

Tobin JO, Beare J, Dunnill MS, Fisher-Hoch S, French M, Mitchell RG, Morris PJ, Muers MF. Legionnaires' disease in a transplant unit: isolation of the causative agent from shower baths. *Lancet.* 1980;19;2(8186):118-21.

Travis TC, Brown EW, Peruski LF, Siludjai D, Jorakate P, Salika P, Yang G, Kozak NA, Kodani M, Warner AK, Lucas CE, Thurman KA, Winchell JM, Thamthitiwat S, Fields BS. Survey of *Legionella* species found in Thai soil*. Int J Microbiol.* 2012;1:218791.

Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. 2011;13(1):36-46.

The UniProt Consortium, UniProt: a worldwide hub of protein knowledge, *Nucleic Acids Res*. 2019;47(D1):D506–D515.

Valsangiacomo C, Baggi F, Gaia V, Balmelli T, Peduzzi R, Piffaretti JC. Use of amplified fragment length polymorphism in molecular typing of Legionella pneumophila and application to epidemiological studies. *J Clin Microbiol*. 1995;33(7):1716-1719.

van der Zee A, Peeters M, de Jong C, Verbakel H, Crielaard JW, Claas EC, Templeton KE. Qiagen DNA extraction kits for sample preparation for *Legionella* PCR are not suitable for diagnostic purposes. *J Clin Microbiol*. 2002;40(3):1126.

Verma UK, Brenner DJ, Thacker WL, Benson RF, Vesey G, Kurtz JB, Dennis PJL, Steigerwalt AG, Robinson JS, Moss CW. *Legionella shakespearei* sp. nov., isolated from cooling tower water. *Int J Syst Bacteriol.* 1992;42:404-407.

Viasus D, Di Yacovo S, Garcia-Vidal C, Verdaguer R, Manresa F, Dorca J, Gudiol F, Carratalà J. Community-acquired *Legionella pneumophila* pneumonia: a single-center experience with 214 hospitalized sporadic cases over 15 years. *Medicine (Baltimore).* 2013;92(1):51-60.

von Baum H, Ewig S, Marre R, Suttorp N, Gonschior S, Welte T, Lück C. Competence Network for Community Acquired Pneumonia Study Group. Community-acquired Legionella pneumonia: new insights from the German competence network for community acquired pneumonia. *Clin Infect Dis.* 2008;46(9):1356-1364.

Waldor MK, Wilson B, Swartz M Cellulitis caused by *Legionella pneumophila. Clin Infect Dis.* 1993;16:51–53.

Walter P, Blobel G. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature*. 1982;299(5885):691-698.

Waring M & Britten RJ. Nucleotide Sequence Repetition: A Rapidly Reassociating Fraction of Mouse DNA. *Science*. 1966;154(3750):791-794

Wartha F, Beiter K, Normark S, Henriques-Normark B. Neutrophil extracellular traps: casting the NET over pathogenesis. *Curr Opin Microbiol.* 2007;10:52–56.

Weiner AM, Deininger PL & Efstratiadis A. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem*. 1986;55:631–661.

Wewalka, G., Schmid, D., Harrison, T. G., Uldum, S. a. & Lück, C. Dual infections with different *Legionella* strains. *Clin. Microbiol. Infect.* 2014;20:1–7.

Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: The unseen majority. *Proc Natl Acad Sci.* 1998;95(12):6578.

WHO, 2007: Legionella and the prevention of legionellosis. World Health Organisation 2007. ISBN 92 4 156297 8.

Wick *et al.,* 2017(a): Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom.* 2017;3(10):e000132.

Wick *et al.,* 2017(b): Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol.* 2017;13(6):e1005595.

Wilkinson HW, Thacker WL, Steigerwalt AG, Brenner DJ, Ampel NM, Wing EJ. Second serogroup of *Legionella hackeliae* isolated from a patient with pneumonia. *J Clin Microbiol.* 1985;22(4):488-9.

Wilkinson HW, Thacker WL, Brenner DJ, Ryan KJ. Fatal *Legionella maceachernii* pneumonia. *J Clin Microbiol.* 1985;22(6):1055.

Wilkinson HW, Thacker WL, Benson RF, Polt SS, Brookings E, Mayberry WR, Brenner DJ, Gilley RG, Kirklin JK. *Legionella birminghamensis* sp. nov. isolated from a cardiac transplant recipient. *J Clin Microbiol.* 1987;25:2120-2122.

Wilkinson HW, Drasar V, Thacker WL, Benson RF, Schindler J, Ptuznikova B, Mayberry WR, Brenner DJ. *Legionella moravica* sp. nov. and *Legionella brunensis* sp. nov. isolated from cooling-tower water. *Ann Inst Pasteur Microbiol.* 1988;139:393-402.

Wilson DA, Yen-Lieberman B, Reischl U, Gordon SM, Procop GW. Detection of Legionella pneumophila by real-time PCR for the *mip* gene. *J Clin Microbiol.* 2003;41(7):3327-3330.

Wilson DA, Reischl U, Hall GS, Procop GW. Use of partial 16S rRNA gene sequencing for identification of *Legionella pneumophila* and non-pneumophila *Legionella spp*. *J Clin Microbiol.* 2007;45(1):257-8.

Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, Salamat SM, Somasekar S, Federman S, Miller S, Sokolic R, Garabedian E, Candotti F, Buckley RH, Reed KD, Meyer TL, Seroogy CM, Galloway R, Henderson SL, Gern JE, DeRisi JL, Chiu CY. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med*. 2014;370(25):2408-2417.

Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 2014;15(3):R46.

Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. 2016;15;32(4):605-607.

Yan Q, Cui S, Chen C, Li S, Sha S, Wan X, Yang R, Xin Y, Ma Y. Metagenomic Analysis of Sputum Microbiome as a Tool toward Culture-Independent Pathogen Detection of Patients with Ventilator-associated Pneumonia. Am J Respir Crit Care Med. 2016;194(5):636-639.

Yang G, Benson RF, Ratcliff RM, Brown EW, Steigerwalt AG, Thacker WL. Daneshvar MI, Morey RE, Saito A, Fields BS. *Legionella nagasakiensis*sp. nov., isolated from water samples and from a patient with pneumonia. *Int J Syst Evol Microbiol.* 2012;62:284-288.

Yiallouros, P. K. *et al.,* First outbreak of nosocomial legionella infection in term neonates caused by a cold mist ultrasonic humidifier. *Clin Infect Dis.* 2013;57**:**48–56.

Yu H, Higa F, Koide M, Haranaga S, Yara S, Tateyama M, Li H, Fujita J. Lung abscess caused by *Legionella* species: implication of the immune status of hosts. *Intern Med.* 2009;48(23):1997-2002.

9. Appendix

9.1 Code

9.1.1 Assembly of PacBio genomes

```
# convert PacBio bas.h5 files to FASTQ format

$ bash5tools.py input.bas.h5 \
      --outFilePrefix prefixname
      --outType fastq \
      --readType Raw

# assemble genome

$ BugBuilder \
      --longfastq input.PacBio.fastq \
      --platform pacbio
      --assembler SPAdes \
      --scaffolder sspace
```

9.1.2 De-multiplex Paired-End Data Files

```
# demultiplex

$ paste -d '' <(echo; sed -n '1,${n;p;}' <barcodes.fastq> | sed G) <R1.fq>
| sed '/^$/d' | fastx_barcode_splitter.pl --bol --bcfile <mappingfile.txt>
--prefix <prefix>

$ paste -d '' <(echo; sed -n '1,${n;p;}' <barcodes.fastq> | sed G) <R2.fq>
| sed '/^$/d' | fastx_barcode_splitter.pl --bol --bcfile <mappingfile.txt>
--prefix <prefix>

# remove appended header

$ bbduk.sh \
      in1=%_R1.fq \
      in2=%_R2.fq \
      out1=%_NOBARCODE_R1.fq \
      out2=%_NOBARCODE_R2.fq \
      ftl=16
```

9.1.3 Data QC (all instances of % refer to sample number)

```
$ fastqc \
      %.fq \
      -o <path_to_output_directory>
```

9.1.4 Adapter Trimming, Quality Trimming and Quality Filtering

```
$ bbduk.sh \
      in1=%_NOBARCODE_R1.fq \
      in2=%_NOBARCODE_R2.fq \
      out1=%_AQF_R1.fq \
      out2=%_AQF_R2.fq \
      ref=<adapter_sequence.fa \
      ktrim=r k=<15/23> \
      tbo tpe \
      qtrim=r trimq=20 \
      ftr=291 \
      minlen=50 \
      stats=%_stats.txt bhist=%_bhist.txt \
      qhist=%_qhist.txt aqhist=%_aqhist.txt \
      lhist=%_lhist.txt
```

9.1.5 Remove PhiX Reads

```
$ bbduk.sh \
      in1=%_AQF_R1.fq \
      in2=%_AQF_R2.fq \
      out1=%_NOPHIX_R1.fq \
      out2=%_NOPHIX_R2.fq \
      ref=phix.fa k=31 hdist=1 \
      stats=%_NOPHIX_stats.txt
```

9.1.6 Remove Human Genome Reads

```
$ bbmap.sh \
      path=hg38_indexed \
      in1=%_NOPHIX_R1.fq \
      in2=%_NOPHIX_R2.fq \
      outu1=%_NOHUMAN_R1.fq \
      outu2=%_NOHUMAN_R2.fq \
      outm=%_HUMAN.fq \
      covstats=%_NOHUMAN_stats.txt
```

9.1.7 Taxonomic Classification

(All instances of % refer to sample number)

```
$ centrifuge
     -x <centrifugedatabase> \
     -p 16 \
     -k 1 \
     -1 %_NOHUMAN_S1_L001_R1_001.fq \
     -2 %_NOHUMAN_S1_L001_R2_001.fq \
     --report-file %_report.centrifuge \
     -S %_results.centrifuge
```

```
# generate kraken-style report

$ centrifuge-kreport \
     -x <centrifugedatabase> \
     %_results.centrifuge \
     > %_kreport.centrifuge
```

9.1.8 *L. pneumophila* 7-loci Sequence Type Analysis

```
$ srst2 \
      --input_pe \
      -forward %_NOHUMAN_S1_L001_R1_001.fq \
      -reverse %_NOHUMAN_S1_L001_R2_001.fq \
      --read_type q \
      --output <outputdir>
      --log \
      --mlst_db allele_sequences.fasta \
      --mlst_definitions allelic_profile.txt
```

9.1.9 Identification of mixed *L. pneumophila* strains

```
# download the list of available bacteria genomes from RefSeq ftp server

$ wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/assembly_summary.txt

# extract all directory-names from file assembly_summary.txt related to
Legionella pneumophila

$ grep -E 'Legionella.*pneumophila' assembly_summary.txt | cut -f 20 >
ftpdirpaths

# create and run the download script

$ awk
'BEGIN{FS=OFS="/";filesuffix="genomic.fna.gz"}{ftpdir=$0;asm=$10;file=asm"_
"filesuffix;print "wget "ftpdir,file}' ftpdirpaths > download_fna_files.sh

$ source download_fna_files.sh

# assign sequence type to RefSeq genomes

$ mlst --legacy –scheme lpneumophila *.fna > lpneuST.tsv

# strainest analysis

# map genomes to reference sequence

$ strainest mapgenomes *.fna Phil1.fasta alignment.fasta

# record variable positions in snp matrix

$ strainest map2snp Phil1.fasta alignment.fasta lp_snpmatric.dgrp
```

```
# calculate site different between each pair

$ strainest snpdist lp_snpmatrix.dgrp lp_snpdist.txt lp_snphist.pdf

# perform hierarchical clustering to produce a reduced list of
representative genomes

$ strainest snpclust lp_snpmatrix.dgrp lp_snpdist.txt lp_snpclust.dgrp
lp_clust.txt

# the representative genomes database was indexed and metagenomes were mapped
to the database

$ bowtie2-build lp_reps.fasta lp_reps

$ bowtie2 \
      --very-fast \
      --no-unal \
      -x lp_reps \
      -1 METAGENOME_R1.fastq \
      -2 METAGENOME_R2.fastq \
      -S METAGENOME.sam

# convert SAM to BAM

$ samtools view -b METAGENOME.sam > METAGENOME.bam

# sort and index

$ samtools sort METAGENOME.bam -o METAGENOME.sorted.bam
$ samtools index METAGENOME.sorted.bam

# infer L. pneumophila strains and relative abundance of strains

$ strainest est lp_snpclust.dgrp METAGENOME.sorted.bam -t 4 -a <> -d <>
<outdirectory>
```

9.1.10 Alignment-based analysis of Sensitivity Tests

```
# align reads to reference genome

$ bbmap.sh \
      in1=%_ NOHUMAN_S1_L001_R1_001.fq \
      in2=$i\_ NOHUMAN_S1_L001_R2_001.fq \
      ref=<reference_genome. \
      minid=0.95 \
      outm1=$i\_out_R1.fq \
      outm2=$i\_out_R2.fq

# remove duplicates

$ dedupe.sh \
      in1=%_out_R1.fq \
      in2==%_out_R2.fq \
      out1=%_outdedupe_R1.fq \
      out2==%_outdedupe_R2.fq \
      stats=%_LEGDEDUPE_stats.txt
```

```
# align deduplicated reads back to reference genome to generate mapping and
coverage statistics

$ bbmap.sh \
      in1=%_outdedupe_R1.fq \
      in2=%_outdedupe_R2.fq \
      ref=<reference_genome.fasta \
      outm=%_LEG-REMAP.fq \
      covstats=%_covstats.txt \
      covhist=%_covhist.txt \
      basecov=%_basecov.txt \
      bincov=%_bincov.txt \
      mhist=%_mhist.txt
```

9.1.11 Mash Genome Distance Calculation

```
# distance calculation between genomes

$ mash sketch -o reference.msh <reference genome>
$ mash sketch *.fna > query.msh
$ mash dist reference.msh query.msh
$ mash sketch *.fna -o fna.msh
```

9.1.12 Alignment-based analysis of Target Capture Data

```
# assign a closely related reference sequence

$ kmerid.py -f <some_file_to_be_analysed -c config/config.cnf

# align metagenomes to the reference sequence

$ bowtie2 \
      -x <indexed reference> --no-unal \
      -1 sample_R1.fq -2 sample_R2.fq -S sample.sam

# convert SAM to BAM and sort

$ picard SortSam \
      INPUT=sample.sam OUTPUT=sample_sorted.bam \
      SORT_ORDER=coordinate

# mark and remove duplicate reads

$ picard MarkDuplicates \
      INPUT=sample_sorted.bam OUTPUT=sample_dedupe.bam \
      METRICS_FILE=sample.txt REMOVE_DUPLICATES=true

# generate metrics before and after duplicate removal

$ pileup.sh in=sample_sorted.bam out=sample_withduplicates.txt
$ pileup.sh in=sample_dedupe.bam out=sample_withoutduplicates.txt
```

9.1.13 Identification of the 50/1455 Core Genes that constitute the Multi-Locus Sequence Based Typing Scheme

```
$ srst2 --input_pe SAMPLE_S1_L001_R1_001.fq SAMPLE _S1_L001_R2_001.fq \
      --gene_db CG.fasta --log --output SAMPLE
```

9.1.14 Metagenome assembly

```
$ spades.py –meta -1 <R1.fq> -2 <R2.fq> \
      --only-assembler -k 27, 47, 67, 87, 107, 127 \
      -o <output_directory>
```

9.1.15 Decontaminate assemblies

```
# classify assembly contigs using Centrifuge

$ centrifuge
     -x <centrifugedatabase> \
     -p 16 \
     -k 1 \
     -U contigs.fasta
     --report-file contigs_report.centrifuge \
     -S contigs_results.centrifuge

# find and print taxid for all reads classified as the genus Legionella
from Centrifuge report file

$ awk '{ if ($1 ~ /Legionella/ || $1 ~ /Tatlockia/) { print } }'
contigs_report.centrifuge

# extract all positive read IDS and print to file

$ awk '{ if ($3 == 446 || $3 == 450) { print } }'
contigs_results.centrifuge > legionella_contigs_results.centrifuge

$ awk '{ if ($3 == 446 || $3 == 450) { print } }'
contigs_results.centrifuge | awk '{print $1}' > legionella_IDS.txt

# extract contigs with corresponding IDS using BBTools filterbyname script

$ filterbyname.sh -Xmx24g \
      in=contigs.fasta \
      out=legionella_only.fasta \
      names=legionella_IDS.txt include=t
```

9.1.16 Phylogenetic analysis of concatenated partial protein sequences

```
$ raxml -m PROTGAMMAAUTO -sample.fasta \
-n sampletree -f a -N autoMRE -x 12345 -p 12345 -T 8
```

9.1.17 Phylogenetic analysis of SNP sites from partial 1455 core genes

```
# for each set of reads from each sample, perform SNP calling

$ snippy --outdir <> --ref CG1455.fasta --R1 R1.fastq --R2 R2.fastq

# create a core alignment

$ snippy-core --prefix <> --ref CG1455.fasta <sample1> <sample2>
<sample..n>

# carry out a phylogenetic analysis based on snippy-core output

$ ./raxml-ng --msa <>_phylo.aln --prefix <> --model GTR+G --bs-trees
autoMRE --bs-metric fbp

# generate best tree containing bootstrap support labels

$ ./raxml-ng --support --tree <>.raxml.bestTree --bs-trees <>.raxml.mlTrees
```

9.1.18 Heterozygous SNP Analysis

```
# for each set of read from each sample, perform SNP calling

snippy --outdir <> --ref CG1455.fasta --R1 R1.fastq --R2 R2.fastq

# from snippy raw output, extract heterozygous alleles of high quality only

$ bcftools view -g het --include 'QUAL>=100 && FMT/DP>=10 &&
(FMT/AO)/(FMT/DP)>=0' snps.raw.vcf > het.snps.vcf

# generate vcf report of heterozygous SNPs

snippy-vcf_report --cpus 8 het_snps.raw.vcf > het_snps.report.txt

# clean the file for plotting

$ sed -n '/#/!p' het_snps.raw.vcf > nohash_het.csv
cut -f 10 nohash_het.csv | sed 's/[\t]/,/g' > het_tab.csv
$ sed 's/:/\t/g' het_tab.csv > het_col.csv
$ cut -f 3 het_col.csv | sed 's/[\t]/,/g' > hetfinal_col.csv
$ awk '{ if(($2 > 4) || ($1 == 0)) { print }}' hetfinal_col.csv >
hetfinal_filtered.csv
```

9.1.19 Guppy Basecalling for Oxford Nanopore (ONT) data

```
$ guppy_basecaller \
      --flowcell FLO-MIN106 --kit SQK-LSK108 \
      -i fast5/ -s basecalled/ \
      -x auto -t 4 --recursive \
      --qscore_filtering --min_qscore 7
```

9.1.20 Remove human DNA reads from ONT data

```
$ minimap2 -ax map-ont hg38.fa ont.fq.gz | samtools fastq -n -f 4 - >
ont_nohuman.fastq.gz
```

9.1.21 Pipeline for the Processing 16S rRNA Illumina data

```
# Combine the forward and reverse reads barcodes

$ module load qiime-1.9.0
$ extract_barcodes.py --input_type barcode_paired_end \
      -f clusters_S1_L001_I1_001.fastq \
      -r clusters_S1_L001_I2_001.fastq --bc1_len 8 --bc2_len 8 \
      -o clusterRun/parsed_barcodes/

# Adapter and Quality Trimming

$ trim_galore -a GGATTAGATACCCNNGTA -a2 CNCTTTANNCCCANT \
      --length 150 --paired \
      clusters_S1_L001_R1_001.fastq \
      clusters_S1_L001_R2_001.fastq \
      -o clusterRun/

# Join the forward and reverse reads

$ join_paired_ends.py \
      -f clusterRun/clusters_S1_L001_R1_001_val_1.fastq \
      -r clusterRun/clusters_S1_L001_R2_001_val_2.fastq \
      -o clusterRun/joined_seqs/ \
      -b clusterRun/parsed_barcodes/barcodes.fastq -j 200 -p 10

# Examine Data Quality

mkdir run_quality_statsnohup

fastx_quality_stats \
-i clusterRun/joined_seqs/fastqjoin.join.fastq \
-o run_quality_stats/clusterRun_quality_stats -Q33

# Demultiplex

nohup split_libraries_fastq.py \
-i clusterRun/joined_seqs/fastqjoin.join.fastq \
-b clusterRun/joined_seqs/fastqjoin.join_barcodes.fastq \
-m clusterRun/max_plate_1_map.txt \
-o clusterRun/split_lib_all \
-q 29 --barcode_type 16 -r 10 -p 0.70

# Remove PhiX contamination

mkdir phix_removed

bwa aln -n 5 phix_genome.fasta split_lib_all/seqs.fna >
phix_removed/clusters.sai

bwa samse phix_genome.fasta phix_removed/clusters.sai
split_lib_all/seqs.fna > phix_removed/clusters_bwa.sam
```

```
samtools view -F 4 -Sbh phix_removed/clusters_bwa.sam
>phix_removed/clusters_phix_only.bam

bamtools convert -in phix_removed/clusters_phix_only.bam-format fasta >
phix_removed/clusters_phix_only.fasta

samtools view -f 4 -Sbh phix_removed/clusters_bwa.sam >
phix_removed/clusters_no_phix.bam

bamtools convert -in phix_removed/clusters_no_phix.bam -format fasta >
phix_removed/clusters_no_phix.fasta

# Count number of PhiX reads removed

grep -c ">" phix_removed/clusters_phix_only.fasta
grep -c ">" phix_removed/clusters_no_phix.fasta

# OTU (Operational Taxonomic Unit) Picking

pick_open_reference_otus.py \
-i phix_removed/clusters_no_phix.fasta \
-r
/16S_reference_databases/silva_115/silva_115_database_final/sequences.fna \
-o /clusterRun/uclust_open_ref_picked_otus_prefilter \
--prefilter_percent_id 0.6  -m uclust -a -O 8 -s 0.1 \
--suppress_taxonomy_assignment --suppress_align_and_tree

# Pick representative sequences for each OTU

mkdir rep_set_uclust_prefilter

pick_rep_set.py \
-i uclust_open_ref_picked_otus_prefilter/final_otu_map_mc2.txt \
-f phix_removed/max_habibi_no_phix.fasta \
-o rep_set_uclust_prefilter/rep_set.fasta -m most_abundant \
-l rep_set_uclust_prefilter/log.txt \

# Align sequences

nohup align_seqs.py \
-i rep_set_uclust_prefilter/rep_set.fasta \
-t
/16S_reference_databases/silva_115/silva_115_database_final/test_ref_align3
.txt \

# Identify and remove chimeric sequences

mkdir chimeric_seqs

nohup identify_chimeric_seqs.py \
-m ChimeraSlayer \
-i clusterRun/pynast_aligned/rep_set_aligned.fasta \
-a
/16S_reference_databases/silva_115/silva_115_database_final/test_ref_align3
.txt \
-o chimeric_seqs/chimeric_seqs.txt

filterfasta.py \
-f pynast_aligned/rep_set_aligned.fasta \
-o pynast_aligned/non-chimeric_rep_set_aligned.fasta \
-s chimeric_seqs/chimeric_seqs.txt -n
```

```
# Filter the chimeric screened sequences for entropic regions

filter_alignement.py \
-i pynast_aligned/non_chimeric_rep_set_aligned.fasta \
-e 0.10 -f 0.80 \
-o pynast_aligned/

# Make phylogenetic tree

mkdir phylogenetic_tree

nohup make_phylogeny.py \
-i pynast_aligned/non_chimeric_rep_set_aligned_pfiltered.fasta \
-o phylogenetic_tree/phylogenetic.tree

# Assign taxonomy

nohup assign_taxonomy.py \
-i rep_set_uclust_prefilter/rep_set.fasta \
-t /16S_reference_databases/silva_115/silva_115_database_final/taxonomy.txt
\
-r
/16S_reference_databases/silva_115/silva_115_database_final/sequences.fna \
-m uclust

# Make the OTU table

mkdir otu_table
make_otu_table.py \
-i uclust_open_ref_picked_otus_prefilter/final_otu_map_mc2.txt \
-o otu_table/otu_table.biom \
-e chimeric_seqs/chimeric_seqs.txt \
-t uclust_assigned_taxonomy/rep_set_tax_assignments.txt

# tidy up

mkdir final_files_for_r
python
/data/lungen/microbiome_data/seq_process_scripts/parse_rep_set_for_r.pyrep_
set_uclust_prefilter/rep_set.fasta final_files_for_r/rep_set_for_r.fasta

cp phylogenetic_tree final_files_for_r/biom convert \
-i otu_table/otu_table.biom \
-o final_files_for_r/otu_table_hdf5.biom \
--table -type="OTU table" --to-hdf5

#16S rRNA Data Analysis in R

#import data into phyloseq

library(phyloseq) # load phyloseq

# import the files

all_data <- import_biom(BIOMfilename = "otu_table_hdf5.biom",
                        treefilename = "phylogenetic.tree",
                        refseqfilename = "rep_set_for_r.fasta")

# import the mapping file as .csv
```

```
map <- read.csv("map.csv", row.names =1)

# turn that into sample data

map <- sample_data(map)

# merge these objects into 1

all_data <- merge_phyloseq(all_data, map)

# rename OTUs and taxonomic ranks

library(Biostrings)
library(ggplot2)
library(plyr)
library(stringr)
library(phyloseq)
library(gridExtra)
library(reshape2)
library(knitr)
library(vegan)

all_data = subset_taxa(all_data, Rank1 != "Unclassified")
all_data = subset_taxa(all_data, Rank3 != "o__Rhodobacterales")
all_data = subset_taxa(all_data, Rank3 != "o__Rhizobiales")
all_data = subset_taxa(all_data, Rank4 != "f__Oxalobacteraceae")
all_data = subset_taxa(all_data, Rank3 != "o__Methylophilales")
all_data = subset_taxa(all_data, Rank5 != "g__Derxia")
all_data = subset_taxa(all_data, Rank1 != "p__Cyanobacteria")
all_data = subset_taxa(all_data, Rank5 != "g__Rhodococcus")

# Makes a string label using the lowest informative tax level

makeTaxLabel <- function(OTU, mydata){
    OTU <- as.character(OTU) # the OTU numbers are stored as character not
integer!
    taxstrings <- as.character(tax_table(mydata)[OTU])
    empty_strings <- c("k__", "p__", "c__", "o__", "f__", "g__", "s__")
    tax_name <- NA
    tax_level <- length(taxstrings) # start at lowest tax level

    while(is.na(tax_name) |
        (tax_name %in% empty_strings)){
    tax_name <- taxstrings[tax_level]
    tax_level <- tax_level -1
    }
    tax_name
}

tax_table(all_data) =gsub("s__uncultured_bacterium",
                           as.character(NA),
                           tax_table(all_data))

tax_table(all_data) =gsub("s__uncultured_organism",
                           as.character(NA),
                           tax_table(all_data))

tax_table(all_data) =gsub("g__uncultured",
                           as.character(NA),
                           tax_table(all_data))
```

```
mynames = NULL
for (i in 1:length(taxa_names(all_data))){
mynames <- rbind(mynames,
c(makeTaxLabel(taxa_names(all_data)[i],all_data)))
}

mynames = gsub("s__", "", mynames)
mynames = gsub("g__", "", mynames)
mynames = gsub("f__", "", mynames)
mynames = gsub("o__", "", mynames)
mynames = gsub("c__", "", mynames)
mynames = gsub("p__", "", mynames)

OTUID = str_c(mynames[,1],"_",seq(1,length(taxa_names(all_data)),by=1))
tax_table(all_data) <- cbind(tax_table(all_data), mynames=OTUID)
#tax_table(all_data)

# Rename tax table headings

colnames(tax_table(all_data)) = c("Kingdom",
                                  "Phylum",
                                   "Class",
                                   "Order",
                                   "Family",
                                   "Genus",
                                   "Species",
                                   "OTUID")

# Rename the taxa_names (frequently used as plot labels by phyloseq)
# to be the same as the new unique informative names

taxa_names(all_data) <- tax_table(all_data)[,8]
save(all_data, file = "LegionellaOutbreaks.Rdata")

# Pick the top 40 dominant OTUs then rarefy to an even depth

pos_top_40 <- names(sort(taxa_sums(positive_control), TRUE)[1:40])
pos_top_40 <- prune_taxa(pos_top_40, positive_control)
pos_top_40_rare <- rarefy_even_depth(pos_top_40, 2000)
```

9.2 Complete Whole Genome Sequences Used for *Legionella pneumophila* Bait Design.

Strain Designation	Serogroup	Sequence Type	Database Source/Accession Number
Lorraine	1	47	NCBI-RefSeq/NC_018139
Pontiac (NCTC 11191)	1	62	NCBI-RefSeq/NZ_CP016029
OLDA (NCTC 12008)	1	1	NCBI-RefSeq/NZ_CP016030
Paris	1	1	NCBI-RefSeq/NC_006368
Lens	1	15	NCBI-RefSeq/NC_006369
Corby	1	51	NCBI-RefSeq/NC_009494
Alcoy	1	578 or 678?	NCBI-RefSeq/NC_014125
Concorde 3 (NCTC11985)	8	8	NCBI-RefSeq/NZ_LT906452
Knoxville-1 (NCTC11286)	1		NCBI-RefSeq/NZ_LT906476
Philadelphia-1(NCTC 11192)	1	36	NCBI-RefSeq/NC_002942
Philadelphia_1 ATCC	1	36	NCBI-RefSeq/NZ_CP015927
Philadelphia_2 (NCTC 11193)	1	36	NCBI-RefSeq/NZ_CP015929
Philadelphia_3	1	36	NCBI-RefSeq/NZ_CP015930
Philadelphia_4	1	36	NCBI-RefSeq/NZ_CP015931
570-CO-H (ATCC43290)	12	187	NCBI-RefSeq/NC_016811
Thunder Bay	6	187	NCBI-RefSeq/CP003730
Toronto-2005	1	222	NCBI-RefSeq/NZ_CP012019
HL06041035	1	734	NCBI-RefSeq/NC_018140
Lpm7613	1	30	NCBI-RefSeq/NZ_LT598657
LPE509		unknown	NCBI-RefSeq/NC_020521
L10-023	1	62	NCBI-RefSeq/NZ_CP011105
F-4185 (subsp. pascullei)	1	1395	NCBI-RefSeq/Nz_CP014255
D-7158 (subsp. pascullei)	5	1335	NCBI-RefSeq/Nz_CP014256
D-7119 (subsp. pascullei)	1	1395	NCBI-RefSeq/NZ_CP014257
D-7630	1	731	NCBI-RefSeq/NZ_CP015344
D-7631	1	731	NCBI-RefSeq/NZ_CP015343
D-7632	1	731	NCBI-RefSeq/NZ_CP015342
F4468	1	731	NCBI-RefSeq/NZ_CP014759
F4469	1	731	NCBI-RefSeq/NZ_CP014760
FFI102	1	15	NCBI-RefSeq/NZ_CP016868
FFI103	1	15	NCBI-RefSeq/NZ_CP016870
FFI104	1	462	NCBI-RefSeq/NZ_CP016872
FFI105	1	462	NCBI-RefSeq/NZ_CP016873
FFI329	1	15	NCBI-RefSeq/NZ_CP016874
FFI337	1	462	NCBI-RefSeq/NZ_CP016876
ST62		62	NCBI-RefSeq/NZ_LT632614
ST23	1	23	NCBI-RefSeq/NZ_LT632615
ST37		37	NCBI-RefSeq/NZ_LT632616
ST42	1	42	NCBI-RefSeq/NZ_LT632617
Detroit-1 (subsp fraseri)	1	Unknown	NCBI-RefSeq/NZ_CP017457
Dallas-1E (subsp fraseri)	5	Unknown	NCBI-RefSeq/NZ_CP017458
C1-S	1	36	NCBI-RefSeq/NZ_CP015932
C2-S	1	36	NCBI-RefSeq/NZ_CP015933
C3-O	1	36	NCBI-RefSeq/NZ_CP015934
C4-S	1	36	NCBI-RefSeq/NZ_CP015935
C5-P	1	36	NCBI-RefSeq/NZ_CP015936
C6-S	1	36	NCBI-RefSeq/NZ_CP015937
C7-O	1	36	NCBI-RefSeq/NZ_CP015938
C8-S	1	36	NCBI-RefSeq/NZ_CP015939
C9_S	1	36	NCBI-RefSeq/NZ_CP015941
C10-S	1	36	NCBI-RefSeq/NZ_CP015944

C11-O	1	36	NCBI-RefSeq/NZ_CP015945
E1-P	1	36	NCBI-RefSeq/NZ_CP015946
E2_N	1	36	NCBI-RefSeq/NZ_CP015947
E3_N	1	36	NCBI-RefSeq/NZ_CP015949
E4_N	1	36	NCBI-RefSeq/NZ_CP015950
E5_N	1	36	NCBI-RefSeq/NZ_CP015951
E6_N	1	36	NCBI-RefSeq/NZ_CP015953
E7_O	1	36	NCBI-RefSeq/NZ_CP015954
E8_O	1	36	NCBI-RefSeq/NZ_CP015955
E9_O	1	36	NCBI-RefSeq/NZ_CP015956
E10_P	1	36	NCBI-RefSeq/NZ_CP015925
E11_U	1	36	NCBI-RefSeq/NZ_CP015926
Cambridge-2/Atkinson (NCTC 11417)	Unknown	Unknown	Sanger NCTC 3000/ERS1080589
Chicago-8 (NCTC 11984)	7	Unknown	Sanger NCTC 3000/ERS1080591
IN-23-GI-C2 (NTCT 11986)	9	390	Sanger NCTC 3000/ERS1080592
Leiden 1 (NCTC 12000)	10	17	Sanger NCTC 3000/ERS1080593
Allentown 1 (NCTC 12024)	1	47	Sanger NCTC 3000/ERS1080594
Heysham 1 (NCTC 12025)	Unknown	Unknown	Sanger NCTC 3000/ERS1080595
1169-MN-H (NCTC 12174)	Unknown	Unknown	Sanger NCTC 3000/ERS1080596
12181 (NCTC 12181)	Unknown	Unknown	Sanger NCTC 3000/ERS1092523
UFW (NCTC 12272)	Unknown	Unknown	Sanger NCTC 3000/ERS1092524
Philadelphia-2 (NCTC 11193)	1	36	Sanger NCTC 3000/ERS1110721
Togus-1 (NCTC 11230)	2	39	Sanger NCTC 3000/ERS1110722
Bellingham-1/77-091436 (NCTC 11404)	Unknown	Unknown	Sanger NCTC 3000/ERS1110726
OLDA (NCTC 12008)	1	1	Sanger NCTC 3000/ERS1110727
Philadelphia 1 (NCTC 11192)	1	36	Sanger NCTC 3000 /ERS579195
France 5811 (NCTC 12007)	Unknown	Unknown	Sanger NCTC 3000 /ERS579199
12180 (NCTC 12180)	Unknown	Unknown	Sanger NCTC 3000 /ERS579201
Knoville-1 (NCTC 11286)	1	Unknown	Sanger NCTC 3000/ERS1211137
Concorde 3 (NCTC 11985)	8	8	Sanger NCTC 3000/ERS1211138

9.3 Sequence Types of *L. pneumophila* Genome Assemblies Deposited in RefSeq ftp Server

Assembly accession number for complete and draft *L. pneumophila* genome assemblies deposited in the RefSeq ftp server (accessed on 5th of October 2018), sequence type (ST) and ESGLI database allele numbers.

U = undetermined sequence type.

~ = allele profile not reported in the ESGLI database.

? = allele with a partial match to a known allele in the ESGLI database.

– = allele absent from assembly.

RefSeq Assembly Accession No.	ST	*flaA*	*pilE*	*asd*	*mip*	*mompS*	*proA*	*neuA/h*
GCF_000048645	1	1	4	3	1	1	1	1
GCF_000694995	1	1	4	3	1	1	1	1
GCF_000953915	1	1	4	3	1	1	1	1
GCF_001582235	1	1	4	3	1	1	1	1
GCF_001582245	1	1	4	3	1	1	1	1
GCF_001582325	1	1	4	3	1	1	1	1
GCF_001582395	1	1	4	3	1	1	1	1
GCF_001582615	1	1	4	3	1	1	1	1
GCF_001582715	1	1	4	3	1	1	1	1
GCF_001582785	1	1	4	3	1	1	1	1
GCF_001582855	1	1	4	3	1	1	1	1
GCF_001583585	1	1	4	3	1	1	1	1
GCF_001601165	1	1	4	3	1	1	1	1
GCF_001601215	1	1	4	3	1	1	1	1
GCF_001601235	1	1	4	3	1	1	1	1
GCF_001601245	1	1	4	3	1	1	1	1
GCF_001601375	1	1	4	3	1	1	1	1
GCF_001601395	1	1	4	3	1	1	1	1
GCF_001601415	1	1	4	3	1	1	1	1
GCF_001601425	1	1	4	3	1	1	1	1
GCF_001601475	1	1	4	3	1	1	1	1
GCF_001601485	1	1	4	3	1	1	1	1
GCF_001677075	1	1	4	3	1	1	1	1
GCF_002934165	1	1	4	3	1	1	1	1
GCF_003004275	1	1	4	3	1	1	1	1
GCF_003004315	1	1	4	3	1	1	1	1

GCF_003205135	1	1	4	3	1	1	1	1
GCF_003205155	1	1	4	3	1	1	1	1
GCF_900048925	1	1	4	3	1	1	1	1
GCF_900049255	1	1	4	3	1	1	1	1
GCF_900050205	1	1	4	3	1	1	1	1
GCF_900050235	1	1	4	3	1	1	1	1
GCF_900051655	1	1	4	3	1	1	1	1
GCF_900051695	1	1	4	3	1	1	1	1
GCF_900052065	1	1	4	3	1	1	1	1
GCF_900052075	1	1	4	3	1	1	1	1
GCF_900052095	1	1	4	3	1	1	1	1
GCF_900052275	1	1	4	3	1	1	1	1
GCF_900052915	1	1	4	3	1	1	1	1
GCF_900052935	1	1	4	3	1	1	1	1
GCF_900053675	1	1	4	3	1	1	1	1
GCF_900057545	1	1	4	3	1	1	1	1
GCF_900057555	1	1	4	3	1	1	1	1
GCF_900057735	1	1	4	3	1	1	1	1
GCF_900058565	1	1	4	3	1	1	1	1
GCF_900058575	1	1	4	3	1	1	1	1
GCF_900058805	1	1	4	3	1	1	1	1
GCF_900060715	1	1	4	3	1	1	1	1
GCF_900061585	1	1	4	3	1	1	1	1
GCF_900062335	1	1	4	3	1	1	1	1
GCF_900062855	1	1	4	3	1	1	1	1
GCF_900063065	1	1	4	3	1	1	1	1
GCF_900063785	1	1	4	3	1	1	1	1
GCF_900064715	1	1	4	3	1	1	1	1
GCF_900452705	1	1	4	3	1	1	1	1
GCF_900053335	2	6	10	19	3	19	4	9
GCF_900461545	3	1	4	3	1	14	1	9
GCF_900061555	4	1	10	19	1	9	4	1
GCF_900050175	5	1	4	3	1	1	1	14
GCF_900053695	5	1	4	3	1	1	1	14
GCF_900060205	5	1	4	3	1	1	1	14
GCF_900061605	6	1	4	3	1	1	1	15
GCF_900051715	7	1	4	3	1	1	1	6
GCF_900061505	7	1	4	3	1	1	1	6
GCF_900052055	9	3	10	1	3	14	9	11
GCF_900052255	9	3	10	1	3	14	9	11
GCF_900053705	9	3	10	1	3	14	9	11
GCF_900059965	9	3	10	1	3	14	9	11

GCF_900060385	9	3	10	1	3	14	9	11
GCF_900061495	9	3	10	1	3	14	9	11
GCF_900061595	9	3	10	1	3	14	9	11
GCF_900063045	9	3	10	1	3	14	9	11
GCF_900051705	10	1	6	3	1	1	1	1
GCF_000048665	15	12	9	26	5	3,26	17	15
GCF_900048915	21	2	3	3	15	2	6	6
GCF_900052905	23	2	3	9	10	2	1	6
GCF_900059575	25	2	6	17	15	12	8	6
GCF_900070125	25	2	6	17	15	12	8	6
GCF_900052085	26	2	6	21	12	12	8	11
GCF_001582635	27	3	10	1	10	14	9	6
GCF_900053345	28	3	10	1	3	14	9	1
GCF_900062315	29	1	4	1	1	14	9	1
GCF_000586295	30	3	10	1	3	14	9	9
GCF_900092465	30	3	10	1	3	14,14	9	9
GCF_900049245	34	3	13	1	25	14	9	6
GCF_000008485	36	3	4	1	1	14	9	1
GCF_000586375	36	3	4	1	1	14	9	1
GCF_001582475	36	3	4	1	1	14	9	1
GCF_001600915	36	3	4	1	1	14	9	1
GCF_001600925	36	3	4	1	1	14	9	1
GCF_001601055	36	3	4	1	1	14	9	1
GCF_001601085	36	3	4	1	1	14	9	1
GCF_001601115	36	3	4	1	1	14	9	1
GCF_001601135	36	3	4	1	1	14	9	1
GCF_001685545	36	3	4	1	1	14	9	1
GCF_001685575	36	3	4	1	1	14	9	1
GCF_001752705	36	3	4	1	1	14	9	1
GCF_001752725	36	3	4	1	1	14	9	1
GCF_001752745	36	3	4	1	1	14	9	1
GCF_001752765	36	3	4	1	1	14	9	1
GCF_001752785	36	3	4	1	1	14	9	1
GCF_001752805	36	3	4	1	1	14	9	1
GCF_001752825	36	3	4	1	1	14	9	1
GCF_001752845	36	3	4	1	1	14	9	1
GCF_001752865	36	3	4	1	1	14	9	1
GCF_001752885	36	3	4	1	1	14	9	1
GCF_001752905	36	3	4	1	1	14	9	1
GCF_001752925	36	3	4	1	1	14	9	1
GCF_001752945	36	3	4	1	1	14	9	1
GCF_001752965	36	3	4	1	1	14	9	1

GCF_001753065	36	3	4	1	1	14	9	1
GCF_001753085	36	3	4	1	1	14	9	1
GCF_001753105	36	3	4	1	1	14	9	1
GCF_001753125	36	3	4	1	1	14	9	1
GCF_001753145	36	3	4	1	1	14	9	1
GCF_001753265	36	3	4	1	1	14	9	1
GCF_001753285	36	3	4	1	1	14	9	1
GCF_001753305	36	3	4	1	1	14	9	1
GCF_001753325	36	3	4	1	1	14	9	1
GCF_001753345	36	3	4	1	1	14	9	1
GCF_001753365	36	3	4	1	1	14	9	1
GCF_001753385	36	3	4	1	1	14	9	1
GCF_001753405	36	3	4	1	1	14	9	1
GCF_001941585	36	3	4	1	1	14	9	1
GCF_002082955	36	3	4	1	1	14	9	1
GCF_002934185	36	3	4	1	1	14	9	11
GCF_003004155	36	3	4	1	1	14	9	1
GCF_900065305	36	3	4	1	1	14	9	1
GCF_900452735	36	3	4	1	1	14	9	1
GCF_000823645	37	3	4	1	1	14	9	11
GCF_900048945	37	3	4	1	1	14	9	11
GCF_900050565	37	3	4	1	1	14	9	11
GCF_900050985	37	3	4	1	1	14	9	11
GCF_900050995	37	3	4	1	1	14	9	11
GCF_900051735	37	3	4	1	1	14	9	11
GCF_900052105	37	3	4	1	1	14	9	11
GCF_900052285	37	3	4	1	1	14	9	11
GCF_900052295	37	3	4	1	1	14	9	11
GCF_900052305	37	3	4	1	1	14	9	11
GCF_900052315	37	3	4	1	1	14	9	11
GCF_900053385	37	3	4	1	1	14	9	11
GCF_900053395	37	3	4	1	1	14	9	11
GCF_900053405	37	3	4	1	1	14	9	11
GCF_900053715	37	3	4	1	1	14	9	11
GCF_900053725	37	3	4	1	1	14	9	11
GCF_900053735	37	3	4	1	1	14	9	11
GCF_900054415	37	3	4	1	1	14	9	11
GCF_900054425	37	3	4	1	1	14	9	11
GCF_900055195	37	3	4	1	1	14	9	11
GCF_900055205	37	3	4	1	1	14	9	11
GCF_900055575	37	3	4	1	1	14	9	11
GCF_900055595	37	3	4	1	1	14	9	11

GCF_900056185	37	3	4	1	1	14	9	11
GCF_900057765	37	3	4	1	1	14	9	11
GCF_900057775	37	3	4	1	1	14	9	11
GCF_900058335	37	3	4	1	1	14	9	11
GCF_900058345	37	3	4	1	1	14	9	11
GCF_900058585	37	3	4	1	1	14	9	11
GCF_900058595	37	3	4	1	1	14	9	11
GCF_900058815	37	3	4	1	1	14	9	11
GCF_900059975	37	3	4	1	1	14	9	11
GCF_900059985	37	3	4	1	1	14	9	11
GCF_900059995	37	3	4	1	1	14	9	11
GCF_900060395	37	3	4	1	1	14	9	11
GCF_900060405	37	3	4	1	1	14	9	11
GCF_900060695	37	3	4	1	1	14	9	11
GCF_900060705	37	3	4	1	1	14	9	11
GCF_900060755	37	3	4	1	1	14	9	11
GCF_900060765	37	3	4	1	1	14	9	11
GCF_900061935	37	3	4	1	1	14	9	11
GCF_900061945	37	3	4	1	1	14	9	11
GCF_900062325	37	3	4	1	1	14	9	11
GCF_900062345	37	3	4	1	1	14	9	11
GCF_900062355	37	3	4	1	1	14	9	11
GCF_900062365	37	3	4	1	1	14	9	11
GCF_900062375	37	3	4	1	1	14	9	11
GCF_900062385	37	3	4	1	1	14	9	11
GCF_900062395	37	3	4	1	1	14	9	11
GCF_900062405	37	3	4	1	1	14	9	11
GCF_900062415	37	3	4	1	1	14	9	11
GCF_900062425	37	3	4	1	1	14	9	11
GCF_900062435	37	3	4	1	1	14	9	11
GCF_900062445	37	3	4	1	1	14	9	11
GCF_900062455	37	3	4	1	1	14	9	11
GCF_900062465	37	3	4	1	1	14	9	11
GCF_900062475	37	3	4	1	1	14	9	11
GCF_900062485	37	3	4	1	1	14	9	11
GCF_900062865	37	3	4	1	1	14	9	11
GCF_900063075	37	3	4	1	1	14	9	11
GCF_900063085	37	3	4	1	1	14	9	11
GCF_900063095	37	3	4	1	1	14	9	11
GCF_900063105	37	3	4	1	1	14	9	11
GCF_900063825	37	3	4	1	1	14	9	11
GCF_900064175	37	3	4	1	1	14	9	11

GCF_900064185	37	3	4	1	1	14	9	11
GCF_900064195	37	3	4	1	1	14	9	11
GCF_900064485	37	3	4	1	1	14	9	11
GCF_900064735	37	3	4	1	1	14	9	11
GCF_900064745	37	3	4	1	1	14	9	11
GCF_900065895	37	3	4	1	1	14	9	11
GCF_900070155	37	3	4	1	1	14	9	11
GCF_900073025	37	3	4	1	1	14	9	11
GCF_900073045	37	3	4	1	1	14	9	11
GCF_900073055	37	3	4	1	1	14	9	11
GCF_900119775	37	3	4	1	1	14	9	11
GCF_900452825	37	3	6	1	14	14	9	11
GCF_900048905	38	3	4	1	14	14	9	11
GCF_000586355	39	3	5	1	7	14	9	8
GCF_001582565	39	3	5	1	7	14	9	8
GCF_900452655	39	3	5	1	7	14	9	8
GCF_900056915	40	3	6	1	14	14	9	11
GCF_900057235	40	3	6	1	14	14	9	11
GCF_900061455	40	3	6	1	14	14	9	11
GCF_900070145	40	3	6	1	14	14	9	11
GCF_000211115	42	4	7	11	3	11	12	9
GCF_000823305	42	4	7	11	3	11	12	9
GCF_900049265	42	4	7	11	3	11	12	9
GCF_900053365	42	4	7	11	3	11	12	9
GCF_900053415	42	4	7	11	3	11	12	9
GCF_900054405	42	4	7	11	3	11	12	9
GCF_900056655	42	4	7	11	3	11	12	9
GCF_900057245	42	4	7	11	3	11	12	9
GCF_900061465	42	4	7	11	3	11	12	9
GCF_900062495	42	4	7	11	3	11	12	9
GCF_900062505	42	4	7	11	3	11	12	9
GCF_900062515	42	4	7	11	3	11	12	9
GCF_900063055	42	4	7	11	3	11	12	9
GCF_900064725	42	4	7	11	3	11	12	9
GCF_900065315	42	4	7	11	3	11	12	9
GCF_900070135	42	4	7	11	3	11	12	9
GCF_900070185	42	4	7	11	3	11	12	9
GCF_900119785	42	4	7	11	3	11	12	9
GCF_900452675	42	4	7	11	3	11	12	9
GCF_003004295	44	4	8	11	10	10	12	2
GCF_900050185	44	4	8	11	10	10	12	2
GCF_900452765	44	4	8	11	10	10	12	2

GCF_900056925	45	5	1	22	26	6	10	12
GCF_900059945	46	5	1	22	5	6	10	15
GCF_900061525	46	5	1	22	5	6	10	15
GCF_000306865	47	5	10	22	15	6	2	6
GCF_900048955	47	5	10	22	15	6	2	6
GCF_900048965	47	5	10	22	15	6	2	6
GCF_900049285	47	5	10	22	15	6	2	6
GCF_900049295	47	5	10	22	15	6	2	6
GCF_900049305	47	5	10	22	15	6	2	6
GCF_900049375	47	5	10	22	15	6	2	6
GCF_900050245	47	5	10	22	15	6	2	6
GCF_900051005	47	5	10	22	15	6	2	6
GCF_900051015	47	5	10	22	15	6	2	6
GCF_900051745	47	5	10	22	15	6	2	6
GCF_900052325	47	5	10	22	15	6	2	6
GCF_900052335	47	5	10	22	15	6	2	6
GCF_900052945	47	5	10	22	15	6	2	6
GCF_900053425	47	5	10	22	15	6	2	6
GCF_900053435	47	5	10	22	15	6	2	6
GCF_900053445	47	5	10	22	15	6	2	6
GCF_900054165	47	5	10	22	15	6	2	6
GCF_900054385	47	5	10	22	15	6	2	6
GCF_900054645	47	5	10	22	15	6	2	6
GCF_900054655	47	5	10	22	15	6	2	6
GCF_900054665	47	5	10	22	15	6	2	6
GCF_900054675	47	5	10	22	15	6	2	6
GCF_900054685	47	5	10	22	15	6	2	6
GCF_900055215	47	5	10	22	15	6	2	6
GCF_900055605	47	5	10	22	15	6	2	6
GCF_900055615	47	5	10	22	15	6	2	6
GCF_900055625	47	5	10	22	15	6	2	6
GCF_900055635	47	5	10	22	15	6	2	6
GCF_900056195	47	5	10	22	15	6	2	6
GCF_900056205	47	5	10	22	15	6	2	6
GCF_900056215	47	5	10	22	15	6	2	6
GCF_900056935	47	5	10	22	15	6	2	6
GCF_900056945	47	5	10	22	15	6	2	6
GCF_900057265	47	5	10	22	15	6	2	6
GCF_900057575	47	5	10	22	15	6	2	6
GCF_900057585	47	5	10	22	15	6	2	6
GCF_900057595	47	5	10	22	15	6	2	6
GCF_900057605	47	5	10	22	15	6	2	6

GCF_900057615	47	5	10	22	15	6	2	6
GCF_900057785	47	5	10	22	15	6	2	6
GCF_900058605	47	5	10	22	15	6	2	6
GCF_900058825	47	5	10	22	15	6	2	6
GCF_900058835	47	5	10	22	15	6	2	6
GCF_900059595	47	5	10	22	15	6	2	6
GCF_900059605	47	5	10	22	15	6	2	6
GCF_900060005	47	5	10	22	15	6	2	6
GCF_900060015	47	5	10	22	15	6	2	6
GCF_900060415	47	5	10	22	15	6	2	6
GCF_900060425	47	5	10	22	15	6	2	6
GCF_900060435	47	5	10	22	15	6	2	6
GCF_900060445	47	5	10	22	15	6	2	6
GCF_900060455	47	5	10	22	15	6	2	6
GCF_900060775	47	5	10	22	15	6	2	6
GCF_900060785	47	5	10	22	15	6	2	6
GCF_900061535	47	5	10	22	15	6	2	6
GCF_900062525	47	5	10	22	15	6	2	6
GCF_900062535	47	5	10	22	15	6	2	6
GCF_900062545	47	5	10	22	15	6	2	6
GCF_900062555	47	5	10	22	15	6	2	6
GCF_900062565	47	5	10	22	15	6	2	6
GCF_900062575	47	5	10	22	15	6	2	6
GCF_900062585	47	5	10	22	15	6	2	6
GCF_900062595	47	5	10	22	15	6	2	6
GCF_900062605	47	5	10	22	15	6	2	6
GCF_900062615	47	5	10	22	15	6	2	6
GCF_900062625	47	5	10	22	15	6	2	6
GCF_900062635	47	5	10	22	15	6	2	6
GCF_900062645	47	5	10	22	15	6	2	6
GCF_900062655	47	5	10	22	15	6	2	6
GCF_900062665	47	5	10	22	15	6	2	6
GCF_900062675	47	5	10	22	15	6	2	6
GCF_900062685	47	5	10	22	15	6	2	6
GCF_900062695	47	5	10	22	15	6	2	6
GCF_900062705	47	5	10	22	15	6	2	6
GCF_900062715	47	5	10	22	15	6	2	6
GCF_900062725	47	5	10	22	15	6	2	6
GCF_900062735	47	5	10	22	15	6	2	6
GCF_900062745	47	5	10	22	15	6	2	6
GCF_900062755	47	5	10	22	15	6	2	6
GCF_900063115	47	5	10	22	15	6	2	6

GCF_900063125	47	5	10	22	15	6	2	6
GCF_900063135	47	5	10	22	15	6	2	6
GCF_900063145	47	5	10	22	15	6	2	6
GCF_900063835	47	5	10	22	15	6	2	6
GCF_900063845	47	5	10	22	15	6	2	6
GCF_900063855	47	5	10	22	15	6	2	6
GCF_900063865	47	5	10	22	15	6	2	6
GCF_900063875	47	5	10	22	15	6	2	6
GCF_900063885	47	5	10	22	15	6	2	6
GCF_900063895	47	5	10	22	15	6	2	6
GCF_900064205	47	5	10	22	15	6	2	6
GCF_900064315	47	5	10	22	15	6	2	6
GCF_900064325	47	5	10	22	15	6	2	6
GCF_900064495	47	5	10	22	15	6	2	6
GCF_900064755	47	5	10	22	15	6	2	6
GCF_900064765	47	5	10	22	15	6	2	6
GCF_900073005	47	5	10	22	15	6	2	6
GCF_900048935	48	5	2	22	27	6	10	12
GCF_900053355	48	5	2	22	27	6	10	12
GCF_000092545	51	6	10	15	28	9	14	6
GCF_003004135	58	7	10	17	10	5,5	4	13
GCF_900186855	58	7	10	17	10	5	4	13
GCF_001582315	59	7	6	17	3	13	11	11
GCF_001582455	59	7	6	17	3	13	11	11
GCF_001582485	59	7	6	17	3	13	11	11
GCF_001582555	59	7	6	17	3	13	11	11
GCF_900057755	59	7	6	17	3	13	11	11
GCF_900060735	59	7	6	17	3	13	11	11
GCF_900060745	60	7	6	17	3	13	11	9
GCF_003003815	61	7	6	17	3	24	11	11
GCF_900475735	61	7	6	17	3	24	11	11
GCF_900050195	62	8	10	3	15	18	1	6
GCF_900054435	62	8	10	3	15	18	1	6
GCF_900056225	62	8	10	3	15	18	1	6
GCF_900057745	62	8	10	3	15	18	1	6
GCF_900058355	62	8	10	3	15	18	1	6
GCF_900058865	62	8	10	3	15	18	1	6
GCF_900060025	62	8	10	3	15	18	1	6
GCF_900063915	62	8	10	3	15	18	1	6
GCF_900063935	62	8	10	3	15	18	1	6
GCF_900064375	62	8	10	3	15	18	1	6
GCF_900065905	62	8	10	3	15	18	1	6

GCF_900060375	63	9	6	3	10	22	15	11
GCF_000785905	68	3	13	1	28	14	9	3
GCF_900053375	68	3	13	1	28	14	9	3
GCF_900061615	68	3	13	1	28	14	9	3
GCF_900053685	72	1	4	3	1	1	1	16
GCF_900051685	77	6	10	14	10	2	3	6
GCF_900061515	77	6	10	14	10	2	3	6
GCF_900049275	78	2	3	6	25	2	1	15
GCF_900063815	78	2	3	6	25	2	1	15
GCF_900452645	80	7	6	3	8	13	11	3
GCF_001600995	94	12	8	11	5	20	12	2
GCF_001601005	94	12	8	11	5	20	12	2
GCF_001601075	94	12	8	11	5	20	12	2
GCF_900065435	107	3	6	1	3	14	9	11
GCF_001583575	150	11	14	16	1	15	13	1
GCF_001582295	154	11	14	16	16	15	13	2
GCF_001582535	159	11	14	16	1	15	13	2
GCF_000239175	187	3	10	1	28	14	9	3
GCF_000586115	187	3	10	1	28	14	9	3
GCF_001997245	187	3	10	1	28	14	9	3
GCF_003345615	187	3	10	1	28	14	9	3
GCF_003345635	187	3	10	1	28	14	9	3
GCF_000823325	191	6	10	19	28	19,78	4	6
GCF_000823345	191	6	10	19	28	19,78	4	6
GCF_000823365	191	6	10	19	28	19,78	4	6
GCF_000823385	191	6	10	19	28	19,78	4	6
GCF_000823405	191	6	10	19	28	78,78	4	6
GCF_000823445	191	6	10	19	28	19,78	4	6
GCF_000823465	191	6	10	19	28	19,78	4	6
GCF_000823505	191	6	10	19	28	78?	4	6
GCF_000823525	191	6	10	19	28	19,78	4	6
GCF_000823745	191	6	10	19	28	19	4	6
GCF_003004195	259	21	27	28	2	15	29	6
GCF_003004235	259	21	27	28	2	15	29	6
GCF_000586195	336	11	14	16	25	7	13	24
GCF_003003865	336	11	14	16	25	7	13	24
GCF_001582645	337	10	22	7	28	16	18	6
GCF_000586175	390	1	4	3	28	1	1	6
GCF_001582135	390	1	4	3	28	1	1	6
GCF_001582155	390	1	4	3	28	1	1	6
GCF_900452775	390	1	4	3	28	1	1	6
GCF_000586335	395	7	6	17	20	13	11	3

GCF_001582545	395	7	6	17	20	13	11	3
GCF_001582325	578	6	10	15	13	9	14	6
GCF_001582735	583	7	6	17	28	13	11	3
GCF_000823805	591	5	2	22	15	6	10	6
GCF_000823485	616	2	10	3	10	9	4	28
GCF_001582145	630	1	4	3	1	1	1	10
GCF_001600985	731	7	10	17	12	29	11	9
GCF_001601155	731	7	10	17	12	29	11	9
GCF_001601175	731	7	10	17	12	29	11	9
GCF_001652645	731	7	10	17	12	29	11	9
GCF_001652665	731	7	10	17	12	29	11	9
GCF_001652685	731	7	10	17	12	29	11	9
GCF_001969405	731	7	10	17	12	29	11	9
GCF_001989475	731	7	10	17	12	29	11	9
GCF_000306845	734	2	6	17	1	1	8	11
GCF_001582695	752	22	4	3	1	1	1	1
GCF_001582705	752	22	4	3	1	1	1	1
GCF_900058555	762	2	3	9	10	1	1	6
GCF_900061475	860	2	3	18	15	1	1	6
GCF_001582405	1101	6	6	15	3	9	14	11
GCF_001583565	1119	2	10	14	10	21	4	3
GCF_000695015	1151	7	43	31	3	48	15	40
GCF_900050215	1156	6	10	15	3	21	7	9
GCF_900056645	1156	6	10	15	3	21	7	9
GCF_001600905	1204	34	27	56	57	72	29	44
GCF_001600975	1204	34	27	56	57	72	29	44
GCF_003003535	1204	34	27	56	57	72	29	44
GCF_003003555	1204	34	27	56	57	72	29	44
GCF_000950675	1288	7	6	17	28	13	11	207
GCF_000950695	1288	7	6	17	28	13	11	207
GCF_000586275	1300	11	14	16	18	15	13	201
GCF_000586075	1318	6	10	5	10	9	1	209
GCF_900452805	1318	6	10	5	10	9	1	209
GCF_000586235	1319	2	6	17	14	12	8	211
GCF_001582305	1319	2	6	17	14	12	8	211
GCF_900461605	1319	2	6	17	14	12	8	211
GCF_000586215	1320	8	1	22	30	6	10	203
GCF_001582225	1320	8	1	22	30	6	10	203
GCF_900186925	1320	8	1	22	30	6	10	203
GCF_000699225	1323	6	10	3	28	9	4	207
GCF_900461535	1324	5	1	22	30	6	10	203
GCF_900057565	1326	3	10	1	28	14	9	207

GCF_900061625	1326	3	10	1	28	14	9	207
GCF_000586315	1334	11	14	16	25	7	13	206
GCF_003003755	1334	11	14	16	25	7	13	206
GCF_000586255	1335	14	18	8	18	28	19	201
GCF_001590645	1335	14	18	8	18	28	19	201
GCF_003004115	1335	14	18	8	18	28	19	201
GCF_900475745	1335	14	18	8	18	28	19	201
GCF_000586095	1362	2	10	3	28	9	4	207
GCF_000586135	1362	2	10	3	28	9	4	207
GCF_001582165	1362	2	10	3	28	9	4	207
GCF_900050225	1362	2	10	3	28	9	4	207
GCF_900057255	1362	2	10	3	28	9	4	207
GCF_900452685	1362	2	10	3	28	9	4	207
GCF_001590615	1395	14	18	8	10	28	19	2
GCF_001590695	1395	14	18	8	10	28	19	2
GCF_900054395	1834	2	3	18	15	63	1	6
GCF_900051025	1983	8	10	3	15	33	1	6
GCF_900051675	1983	8	10	3	15	33	1	6
GCF_900051755	1983	8	10	3	15	33	1	6
GCF_900053745	1983	8	10	3	15	33	1	6
GCF_900054695	1983	8	10	3	15	33	1	6
GCF_900056235	1983	8	10	3	15	33	1	6
GCF_900056245	1983	8	10	3	15	33	1	6
GCF_900057275	1983	8	10	3	15	33	1	6
GCF_900057625	1983	8	10	3	15	33	1	6
GCF_900058845	1983	8	10	3	15	33	1	6
GCF_900058855	1983	8	10	3	15	33	1	6
GCF_900060795	1983	8	10	3	15	33	1	6
GCF_900061545	1983	8	10	3	15	33	1	6
GCF_900061565	1983	8	10	3	15	33	1	6
GCF_900062875	1983	8	10	3	15	33	1	6
GCF_900063905	1983	8	10	3	15	33	1	6
GCF_900064345	1983	8	10	3	15	33	1	6
GCF_900064365	1983	8	10	3	15	33	1	6
GCF_900064385	1983	8	10	3	15	33	1	6
GCF_900065915	1983	8	10	3	15	33	1	6
GCF_900073035	1983	8	10	3	15	33	1	6
GCF_001582375	1999	3	6	1	6	1	11	9
GCF_900060725	2122	2	10	3	10	9	4	9
GCF_003004065	2186	30	18	44	77	61	13	217
GCF_003004335	2258	21	27	29	80	15	29	230
GCF_900052925	2439	2	3	9	10	93	1	6

GCF_002813735	2512	11	14	16	1	7	13	207
GCF_003205055	2581	3	12	1	14	14	8	3
GCF_003205065	2581	3	12	~1	14	14	8	3
GCF_003205175	2581	3	12	1	14	14	8	3
GCF_001583595	U	1	4	3	1	~21	1	1
GCF_000950745	U	2	10	17	6	14	14	207
GCF_000953935	U	2	10	17	14	9,21	14	221
GCF_001592705	U	2	19	5	10	18,63	1	10
GCF_002967055	U	2	19	5	10	18,63	1	10
GCF_003205035	U	2	10	5	5	35,63	5	6
GCF_003205045	U	2	10	5	5	35,63	5	6
GCF_003205115	U	2	10	5	5	35,63	5	6
GCF_900050555	U	2	3	6	25	63	1	15
GCF_900050975	U	2	3	6	25	63	1	15
GCF_900051725	U	2	3	6	25	93	1	15
GCF_900052265	U	2	3	9	10	1,93	1	6
GCF_900053665	U	2	10	9	13	2,63	5	6
GCF_900054155	U	2	3	6	25	1	1	15
GCF_900055585	U	2	3	6	25	1	1	15
GCF_900060195	U	2	3	6	10	98?	1	6
GCF_900061485	U	2	3	9	10	2,63	1	6
GCF_900061635	U	2	3	6	25	2,63	1	15
GCF_900061645	U	2	3	6	25	2,63	1	15
GCF_900061655	U	2	3	6	25	2,63	1	15
GCF_900061665	U	2	3	6	25	93	1	15
GCF_900061675	U	2	3	6	25	2,63	1	15
GCF_900061685	U	2	3	6	25	1	1	15
GCF_900061695	U	2	3	6	25	63	1	15
GCF_900061705	U	2	3	6	25	1,93	1	15
GCF_900061715	U	2	3	6	25	93	1	15
GCF_900061725	U	2	3	6	25	1,93	1	15
GCF_900061975	U	2	3	9	10	63,93	1	6
GCF_900063795	U	2	10	18	10	~63	1	9
GCF_900063805	U	2	3	6	25	1	1	15
GCF_900119755	U	2	3	9	10	2,93	1	6
GCF_000347615	U	3	10	1	1	47?	9	1
GCF_002002645	U	3	13	1	28	-	9	3
GCF_900059585	U	3	4	1	1	7,14	9	11
GCF_000823425	U	6	10	14	28	4,9	3	207
GCF_000823545	U	6	10	19	28	19	4	6
GCF_000823565	U	6	10	19	28	19,78	4	6
GCF_000823585	U	6	10	19	28	19,78	4	6

GCF_000823605	U	6	10	19	28	19,78	4	6
GCF_000823625	U	6	10	19	28	19,78	4	6
GCF_000823665	U	6	10	19	28	19,78	4	6
GCF_000823685	U	6	10	19	28	78	4	6
GCF_000823705	U	6	10	19	28	78	4	6
GCF_000823725	U	6	10	19	28	78	4	6
GCF_000823765	U	6	10	19	28	78	4	6
GCF_000823785	U	6	10	19	28	78	4	6
GCF_000823825	U	6	10	19	28	78	4	6
GCF_000823845	U	6	10	19	28	78	4	6
GCF_000823865	U	6	10	19	28	78	4	6
GCF_000823885	U	6	10	19	28	78	4	6
GCF_001601325	U	6	10	19	28	78?	4	11
GCF_001601355	U	6	10	19	28	78?	4	11
GCF_001582385	U	6	10	17	28	9	14	11
GCF_001582795	U	6	10	15	28	88?	7	207
GCF_001582865	U	6	10	20	12	88?	4	3
GCF_001601505	U	6	10	19	28	79?	4	11
GCF_001601535	U	6	10	19	28	77?	4	11
GCF_002002625	U	6	10	15	3	~21	14	9
GCF_900059935	U	6	10	15	24	17,98	14	6
GCF_900059955	U	6	10	21	12	98	4	11
GCF_900060365	U	6	10	21	12	9,98	4	11
GCF_900061575	U	6	10	21	12	9,98,98	4	11
GCF_900070165	U	6	10	21	12	98	4	11
GCF_001600895	U	8	10	3	15	18,33	1	6
GCF_001601455	U	8	10	3	15	60?	1	6
GCF_001610735	U	8	10	3	15	18,33	1	6
GCF_001677115	U	8	10	3	15	18,33	1	6
GCF_003004175	U	8	3	3	15	9,21	1	6
GCF_900063925	U	8	10	3	15	18,33	1	6
GCF_900064335	U	8	10	3	15	18,33	1	6
GCF_900064355	U	8	10	3	15	18,33	1	6
GCF_900065445	U	8	10	3	15	18,33	1	6
GCF_900119765	U	8	10	3	15	18,33	1	6
GCF_001549915	U	11	14	16	10	7	13	2
GCF_001549925	U	11	14	16	10	7	13	2
GCF_001582625	U	11	14	16	31	7	13	210
GCF_001582775	U	11	14	16	16	7	13	2
GCF_001583645	U	11	14	16	1	7	13	1
GCF_001639045	U	11	14	16	10	7	13	2
GCF_001886795	U	11	6	16	16	7,15	13	2

GCF_001886835	U	11	14	16	18	7,15	13	201
GCF_002813715	U	11	14	16	1	7,15	13	207
GCF_002934205	U	11	14	16	16	7,15	13	2
GCF_003003595	U	11	14	16	3	7,15	13	9
GCF_003003675	U	11	14	16	16	7,15	13	2
GCF_003003955	U	11	14	16	30	7,15	13	213
GCF_003004215	U	11	14	16	1	7,15	13	1
GCF_003004255	U	11	14	16	16	7,15	13	2
GCF_900051665	U	11	14	16	1	7	13	6
GCF_900063035	U	11	14	16	1	7	13	6
GCF_000586155	U	12	17	11	10	5	12	-
GCF_001582215	U	12	17	11	10	5	12	-
GCF_001766275	U	12	9	26	5	3,26	17	15
GCF_001766295	U	12	9	26	5	3,26	17	15
GCF_001766315	U	12	9	2	5	3,50	17	15
GCF_001766335	U	12	9	2	5	3,50	17	15
GCF_001766355	U	12	9	26	5	3,26	17	15
GCF_001766375	U	12	9	2	5	3,50	17	15
GCF_003003515	U	21	27	28	83	15,15	29	-
GCF_001583655	U	28	21	12	37	41?	1	215
GCF_001582465	U	31	10	20	10	88?	4	11

9.4 StrainEst Predictions for Mock Communities and Sensitivity Tests - Chapter 3

ST = sequence type

Min/Max Depth = minimum and maximum depth of coverage of Single Nucleotide Polymorphisms

Std = standard deviation

MSEAve = Average mean squared error = statistical measure of the quality of the estimator (values closer to zero are better)

Community	Strain Predictions	Abundance	ST	Min/Max Depth	MSEAve (Std)
Mock1-1	**GCF_000953915**	**0.62062**	**1**	67/150	0.002 (4.99E-05)
	GCF_001601485	**0.18784**	**1**		
	GCF_001752885	**0.06658**	**36**		
	GCF_001601245	**0.05593**	**1**		
	GCF_900050175	0.02691	5		
	GCF_900053675	0.01506	1		
	GCF_001582785	0.0116	1		
	GCF_900461545	0.00773	3		
	GCF_001582695	0.0044	752		
	GCF_900061605	0.00261	6		
	GCF_001753105	0.00064	36		
	GCF_900452775	8.10E-05	390		

Community	Strain Predictions	Abundance	ST	Min/Max Depth	MSEAve (Std)
Mock1-2	**GCF_001752885**	**0.41589**	**36**	57/126	0.0018 (2.02E-05)
	GCF_000953915	**0.34729**	**1**		
	GCF_001601485	**0.12891**	**1**		
	GCF_001601245	**0.04421**	**1**		
	GCF_900050175	0.02003	5		
	GCF_900461545	0.0152	3		
	GCF_900053675	0.01114	1		
	GCF_900073025	0.00542	37		
	GCF_001582695	0.00407	752		
	GCF_001582785	0.0032	1		
	GCF_900061605	0.00282	6		
	GCF_900452775	0.00153	390		
	GCF_900452655	0.00029	39		

Community	Strain Predictions	Abundance	ST	Min/Max Depth	MSEAve (Std)
Mock1-3	**GCF_001752885**	**0.83961**	**36**	56/127	0.000845 (2.60E-05)
	GCF_000953915	**0.08166**	**1**		
	GCF_001601485	**0.02294**	**1**		
	GCF_001601245	**0.01103**	**1**		
	GCF_001753105	**0.00782**	**36**		
	GCF_001601085	**0.00733**	**36**		

	GCF_900056185	0.00575	37		
	GCF_900061605	0.00459	6		
	GCF_900050175	0.00425	5		
	GCF_900062315	0.00405	29		
	GCF_001752965	0.00353	36		
	GCF_900053675	0.00241	1		
	GCF_900073025	0.00179	37		
	GCF_900461545	0.0014	3		
	GCF_003004175	0.00109	U		
	GCF_001582695	0.00045	752		
	GCF_003205045	0.0003	U		

Community	Strain Predictions	Abundance	ST	Min/Max Depth	MSEAve (Std)
Mock2-1	**GCF_000953915**	**0.565351**	**1**	58/135	0.0017 (2.65E-05)
	GCF_001601485	**0.155375**	**1**		
	GCF_001752885	**0.147011**	**36**		
	GCF_001601245	**0.063739**	**1**		
	GCF_900053675	**0.0214**	**1**		
	GCF_900050175	0.017079	5		
	GCF_900461545	0.011198	3		
	GCF_001582785	0.009807	1		
	GCF_001582695	0.004647	752		
	GCF_900061605	0.002712	6		
	GCF_900073025	0.000867	37		
	GCF_001753105	0.000659	36		
	GCF_001601085	0.000155	36		

Community	Strain Predictions	Abundance	ST	Min/Max Depth	MSEAve (Std)
Mock2-2	**GCF_001752885**	**0.41603**	**36**	42/96	0.0021 (2.45E-05)
	GCF_000953915	**0.3605**	**1**		
	GCF_001601485	**0.11771**	**1**		
	GCF_001601245	**0.03982**	**1**		
	GCF_900050175	0.02061	5		
	GCF_900461545	0.01371	3		
	GCF_900053675	0.01239	1		
	GCF_001582695	0.00731	752		
	GCF_900073025	0.00689	37		
	GCF_900061605	0.00308	6		
	GCF_900452775	0.00147	390		
	GCF_900452655	0.00049	39		

Community	Strain Predictions	Abundance	ST	Min/Max Depth	MSEAve (Std)
Mock2-3	**GCF_001752885**	**0.910805**	**36**	38/89	0.00084 (2.89E-05)
	GCF_000953915	**0.024031**	**1**		
	GCF_001601245	**0.021302**	**1**		
	GCF_001601485	**0.015231**	**1**		
	GCF_900050175	0.007755	5		
	GCF_001601085	0.004866	36		

	GCF_900452775	0.003838	390		
	GCF_900061605	0.003389	6		
	GCF_001753105	0.003316	36		
	GCF_003004175	0.001838	U		
	GCF_900461545	0.001787	3		
	GCF_001582785	0.001541	1		
	GCF_900062315	0.000301	29		

Community	Strain Predictions	Abundance	ST	Min/Max Depth	MSEAve (Std)
Mock3-1	**GCF_000953915**	**0.423394**	**1**	11/31	0.0047 (3.35e-05)
	GCF_001752885	**0.31465**	**36**		
	GCF_001601485	**0.145111**	**1**		
	GCF_001601245	**0.048384**	**1**		
	GCF_900050175	0.02215	5		
	GCF_900461545	0.021597	3		
	GCF_900053675	0.012915	1		
	GCF_001582695	0.00447	752		
	GCF_900452775	0.003235	390		
	GCF_900061605	0.001637	6		
	GCF_001582785	0.001471	1		
	GCF_900073025	0.000985	37		

Community	Strain Predictions	Abundance	ST	Min/Max Depth	MSEAve (Std)
Mock3-2	**GCF_001752885**	**0.790711**	**36**	35/82	0.00133 (2.11e-05)
	GCF_000953915	**0.125847**	**1**		
	GCF_001601485	**0.034854**	**1**		
	GCF_900056185	0.008326	37		
	GCF_900073025	0.006888	37		
	GCF_001752965	0.0055	36		
	GCF_001753105	0.005209	36		
	GCF_900461545	0.004504	3		
	GCF_900050175	0.003917	5		
	GCF_001601245	0.003788	1		
	GCF_001582695	0.002936	752		
	GCF_900062315	0.002717	29		
	GCF_900053675	0.002116	1		
	GCF_900061605	0.001365	6		
	GCF_003004175	0.001272	U		
	GCF_003205045	5.10E-05	U		

Community	Strain Predictions	Abundance	ST	Min/Max Depth	MSEAve (Std)
Mock3-3	**GCF_001752885**	**0.967426**	**36**	50/114	0.000439 (2.13e-05)
	GCF_001601245	**0.011815**	**1**		
	GCF_000953915	**0.006312**	**1**		
	GCF_001601485	**0.005845**	**1**		
	GCF_900050175	0.002454	5		
	GCF_900461545	0.002239	3		

	GCF_900452775	0.00128	390		
	GCF_001582695	0.000799	752		
	GCF_001582785	0.000603	1		
	GCF_900061605	0.00057	6		
	GCF_900073025	0.000538	37		
	GCF_001601085	0.000116	36		
	GCF_001753105	2.00E-06	36		
	GCF_001752885	0.967426	36		

Single Strain Control Tests

Community	Strain Predictions	Abundance	ST	Min/Max Depth	MSEAve (Std)
S8	**GCF_001752885**	**0.999562**	**36**	21/56	0.000445 (3.16e-05)
	GCF_001752965	**0.000427**	**36**		
	GCF_900056185	1.1E-05	37		

Community	Strain Predictions	Abundance	ST	Min/Max Depth	MSEAve (Std)
S9	**GCF_001752885**	**0.99612**	**36**	7/24	0.000514 (3.49e-05)
	GCF_001752965	**0.000248**	**36**		
	GCF_900062315	0.00014	29		

Community	Strain Predictions	Abundance	ST	Min/Max Depth	MSEAve (Std)
S10	**GCF_001752885**	**0.99379**	**36**	2/12	0.00035 (4.46E-05)
	GCF_001753105	**0.005123**	**36**		
	GCF_900062315	0.000709	29		
	GCF_900056185	0.000378	37		

9.5 Completed Legionella Genomes in KmerID Database.

Species/Strain Designation	Database Source/Accession Number
L. pneumophila Lorraine	NCBI-RefSeq/NC_018139
L. pneumophila Corby	NCBI-RefSeq/NC_009494
L. pneumophila ST62	NCBI-RefSeq/NZ_LT632614
L. pneumophila ST23	NCBI-RefSeq/NZ_LT632615
L. pneumophila ST42	NCBI-RefSeq/NZ_LT632617
L. pneumophila C8-S	NCBI-RefSeq/NZ_CP015939
L. pneumophila Toronto-2005	NCBI-RefSeq/NZ_CP012019
L. pneumophila FFI329	NCBI-RefSeq/NZ_CP016874
L. pneumophila D-7631	NCBI-RefSeq/NZ_CP015343
L. pneumophila D-7158 (subsp. pascullei)	NCBI-RefSeq/Nz_CP014256
L. pneumophila Dallas-1E (subsp fraseri)	NCBI-RefSeq/NZ_CP017458
L. pneumophila Concorde 3 (NCTC11985)	NCBI-RefSeq/NZ_LT906452
L. pneumophila HL06041035	NCBI-RefSeq/NC_018140
L. pneumophila Allentown 1 (NCTC 12024)	Sanger NCTC 3000/ERS1080594
L. pneumophila Togus-1 (NCTC 11230)	Sanger NCTC 3000 /ERS1110722
L. pneumophila 1169-MN-H (NCTC 12174)	Sanger NCTC 3000/ERS1080596
L. pneumophila OLDA (NCTC 12008)	Sanger NCTC 3000 /ERS1110727
L. pneumophila 12181 (NCTC 12181)	Sanger NCTC 3000/ERS1092523
L. pneumophila Leiden 1 (NCTC 12000)	Sanger NCTC 3000/ERS1080593
L. pneumophila Cambridge-2/Atkinson (NCTC 11417)	Sanger NCTC 3000/ERS1080589
L. pneumophila Chicago-8 (NCTC 11984)	Sanger NCTC 3000/ERS1080591
L. waltersii (NCTC13017)	NCBI-RefSeq/NZ_LT906442
L. fallonii LLAP-10	NCBI-RefSeq/NZ_LN614827
L. birminghamensis	Sanger NCTC 3000/ERS1497499
L. clemsonensis CDC-D5610	NCBI-RefSeq/NZ_CP016397
L. taurinensis	Sanger NCTC 3000/ERS1324129
L. spiritensis	Sanger NCTC 3000/ERS1324117
L. micdadei NZ2015	NCBI-RefSeq/NZ_CP020614
L. donaldsonii	Sanger NCTC 3000/ERS1110725
L. feeleii	Sanger NCTC 3000/ERS579197
L. lansingensis (NCTC12830)	NCBI-RefSeq/NZ_LT906451
L. longbeachae	Sanger NCTC 3000/ERS950475
L. oakridgensis	Sanger NCTC 3000/ERS950476
L. sainthelensi	Sanger NCTC 3000/ERS579202
L. steigerwaltii	Sanger NCTC 3000/ERS1324118
L. cherrii	Sanger NCTC 3000/ERS579196
L. anisa FDAARGOS 200 AMERTCC 13	NCBI-RefSeq/NBTX01000001
L. wadsworthii	Sanger NCTC 3000/ERS956173
L. hackeliae	Sanger NCTC 3000/ERS579198

9.6 Sankey Diagrams for the Taxonomic Classification of Bacterial Reads – Chapter 5

Taxonomic classification of bacterial reads from the dilution series (D1, D2, D3, D4 [for further details see Table 5.4]), the mock sample (for further details see Table 5.5), clinical samples (H1 to H10) and environmental samples (E1 to E9).

D = Domain, P = Phylum, F = Family, G = Genus and S = Species.

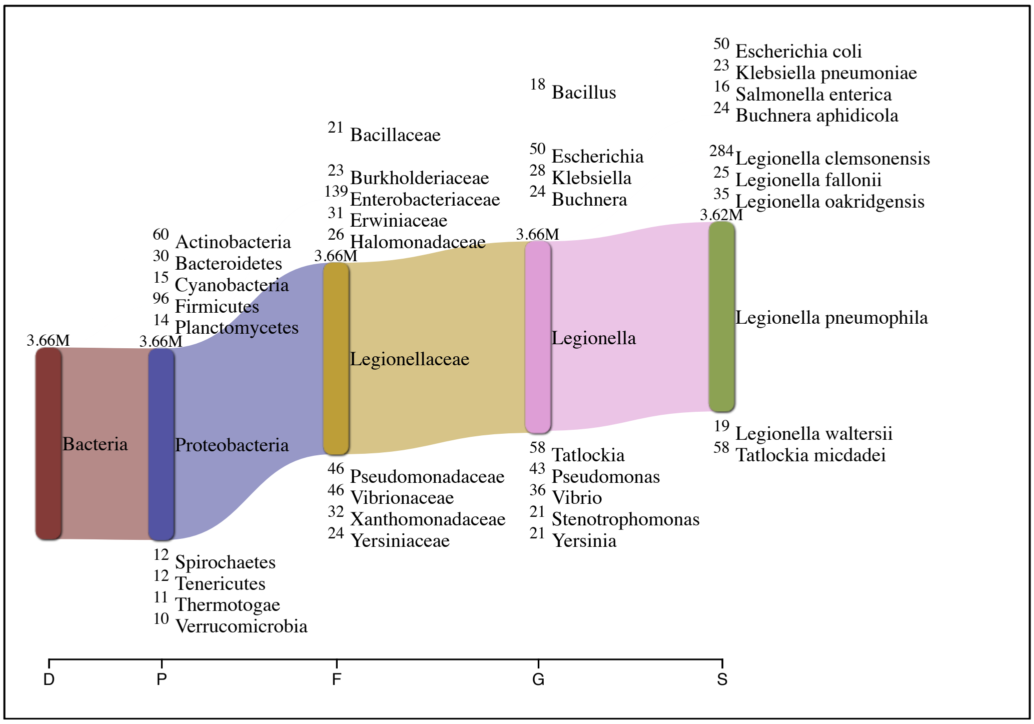


Figure 1. Sample D1

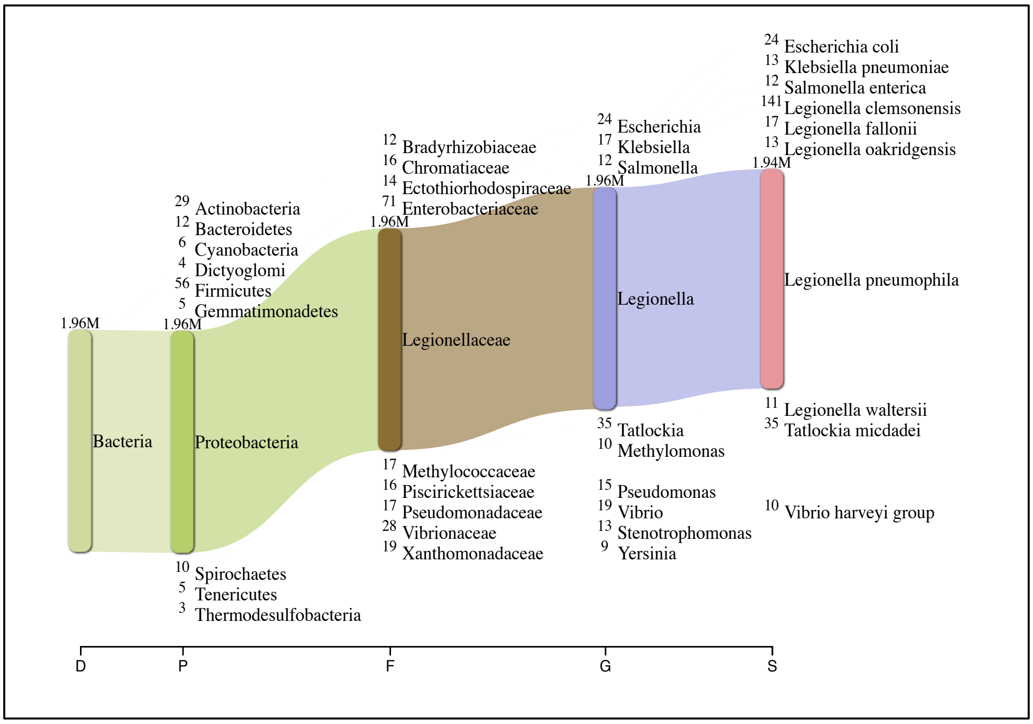


Figure 2. Sample D2

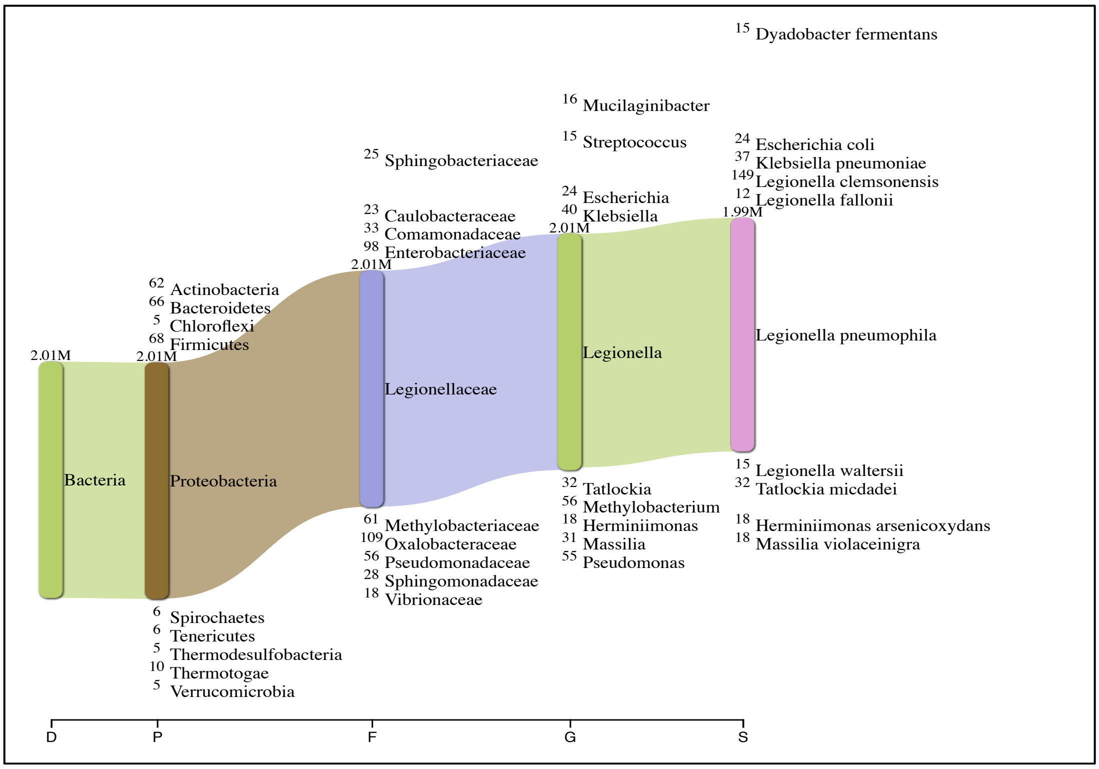


Figure 3. Sample D3

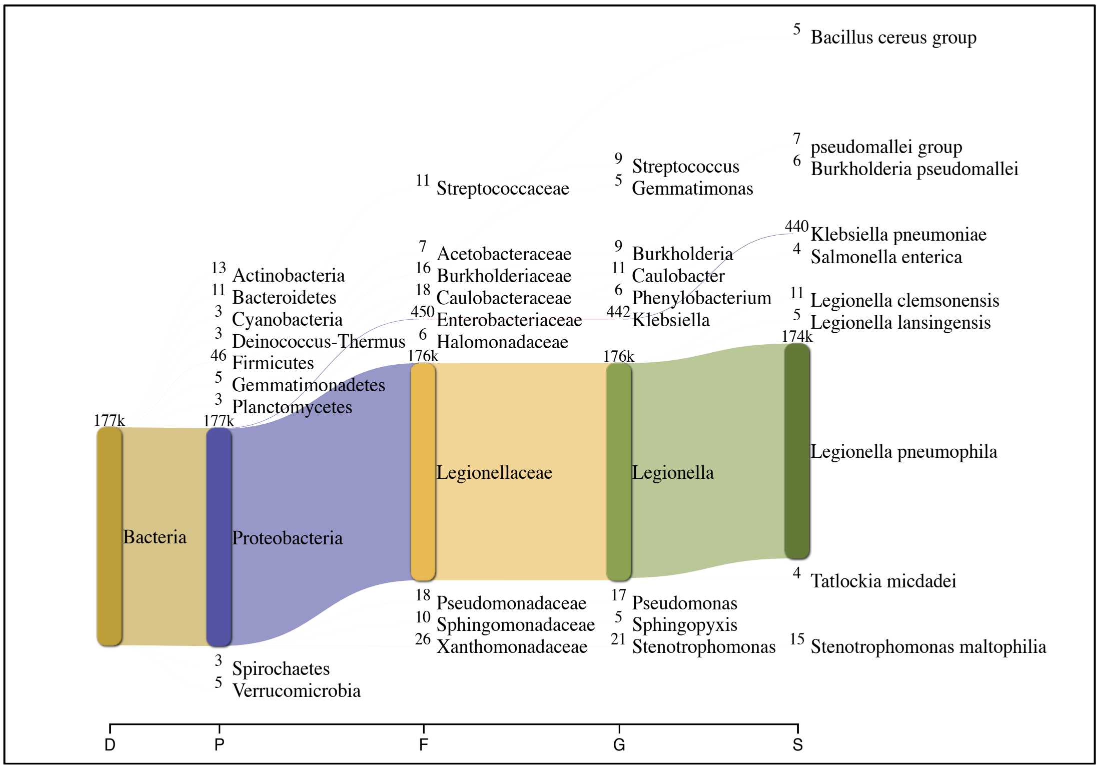


Figure 4. Sample D4

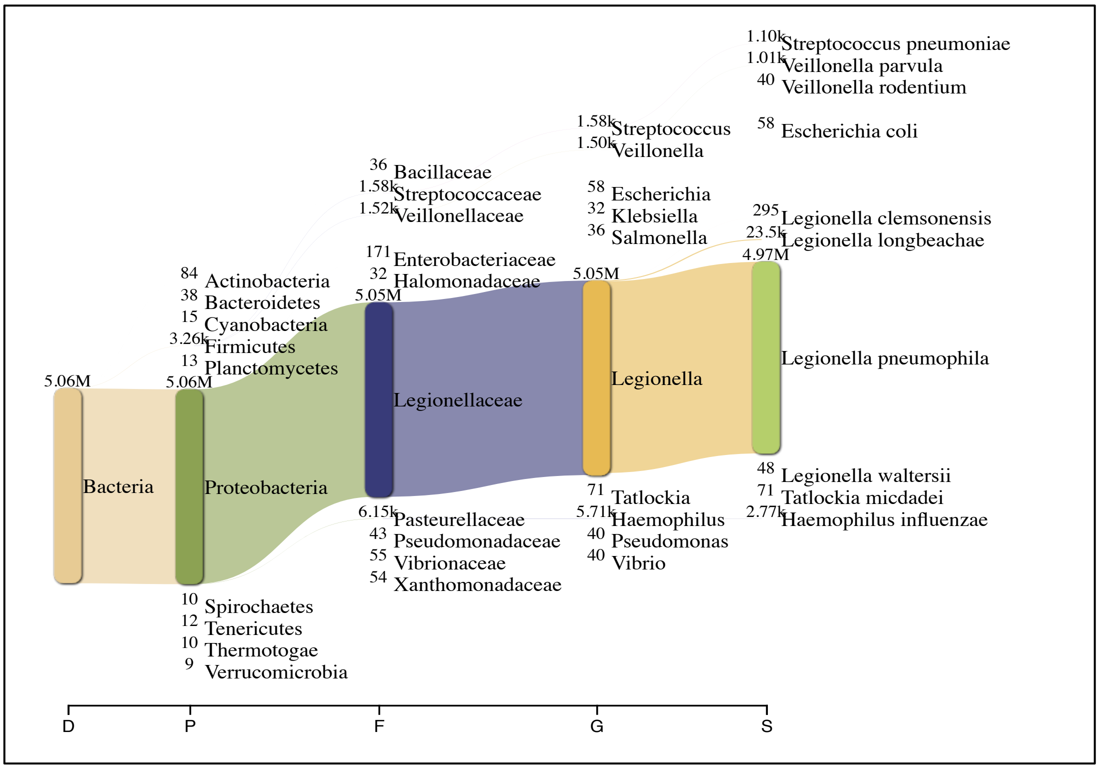


Figure 5. Sample Mock

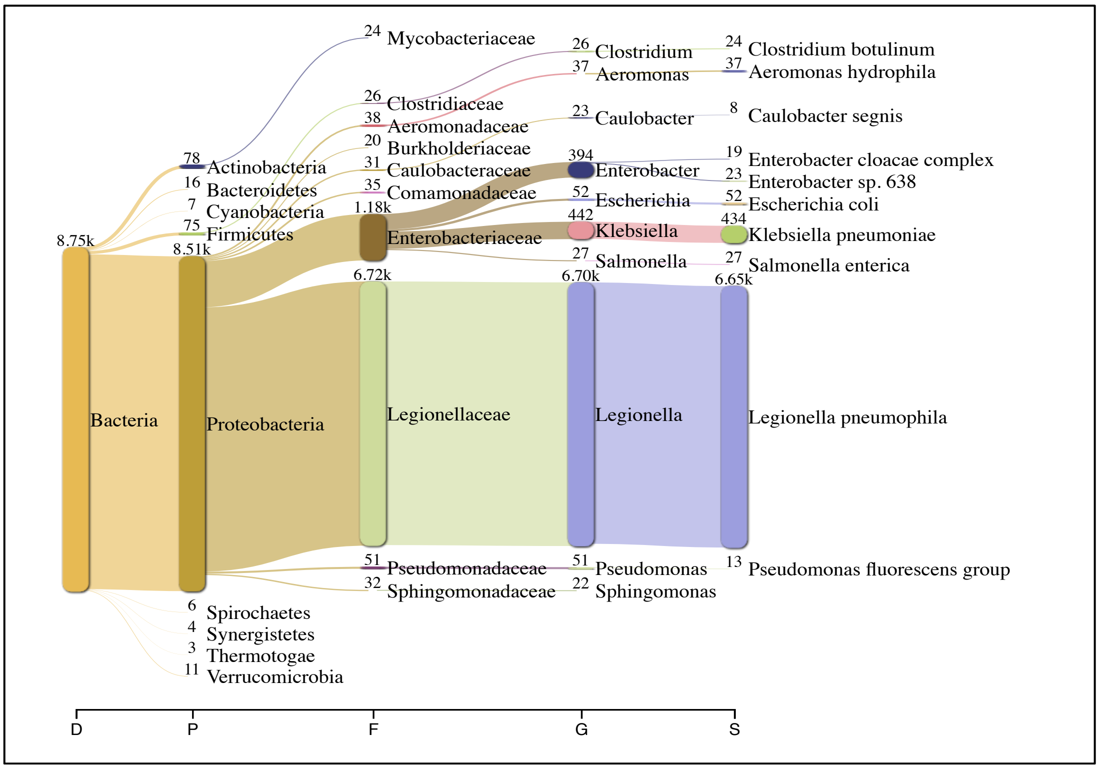


Figure 6. Sample S1

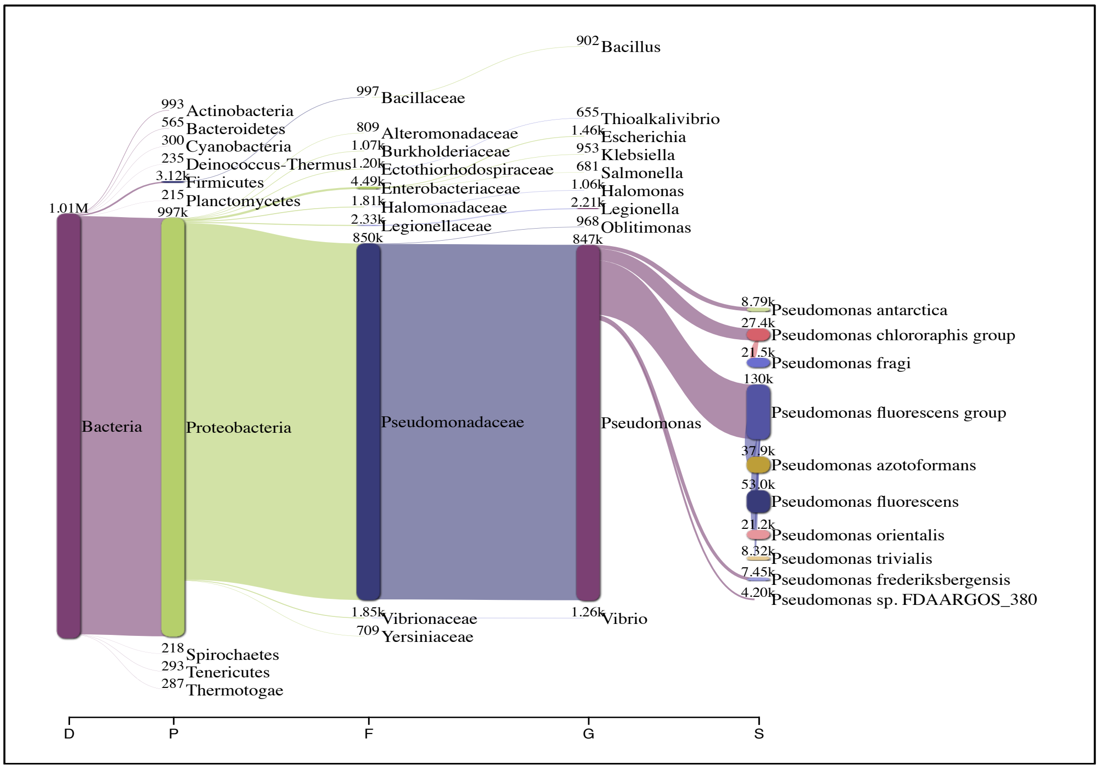


Figure 7. Sample S2

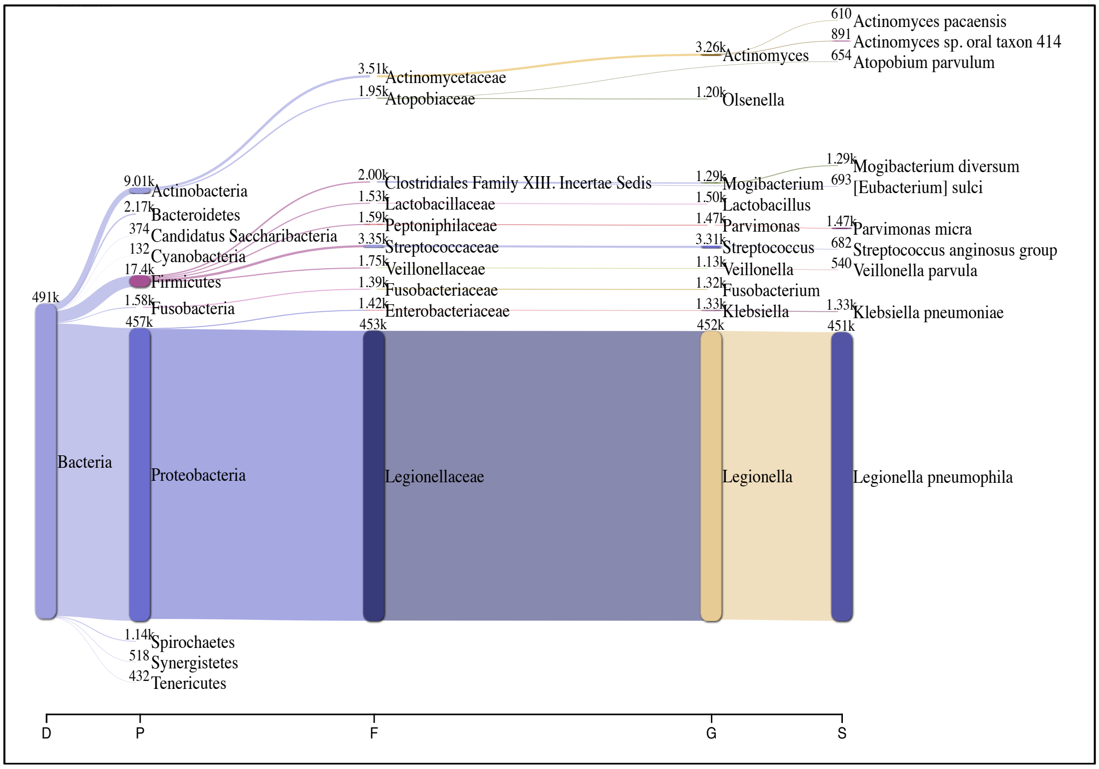


Figure 8. Sample S3

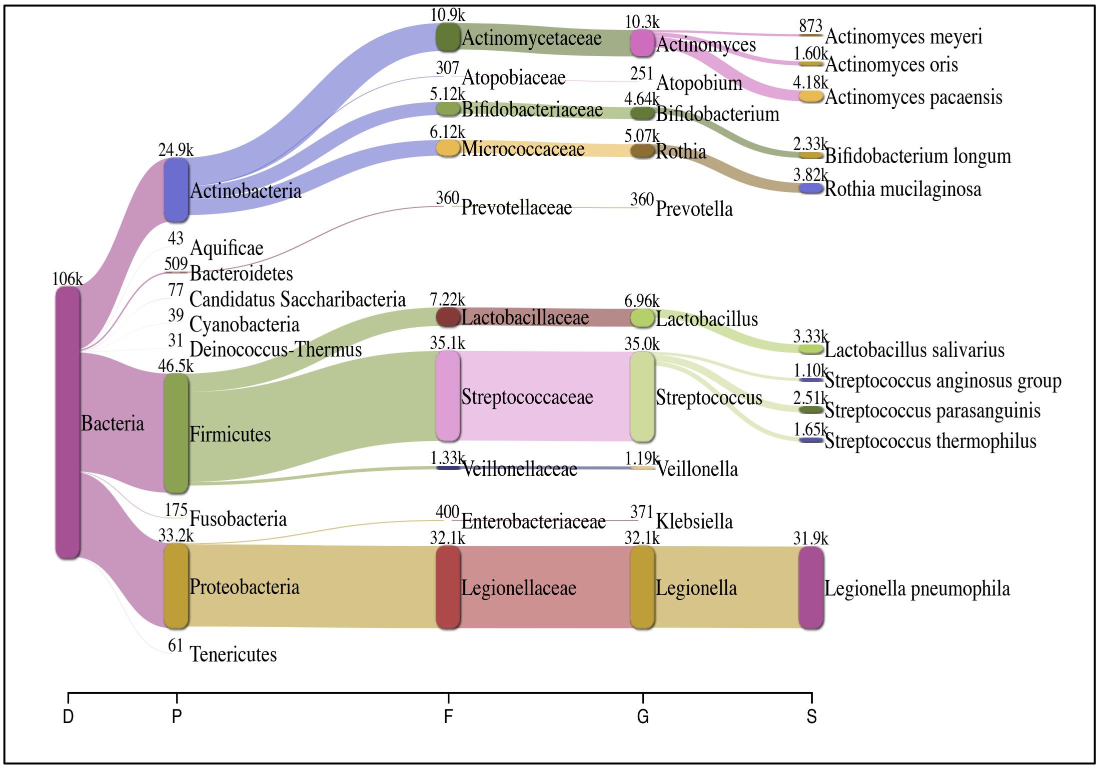


Figure 9. Sample S4

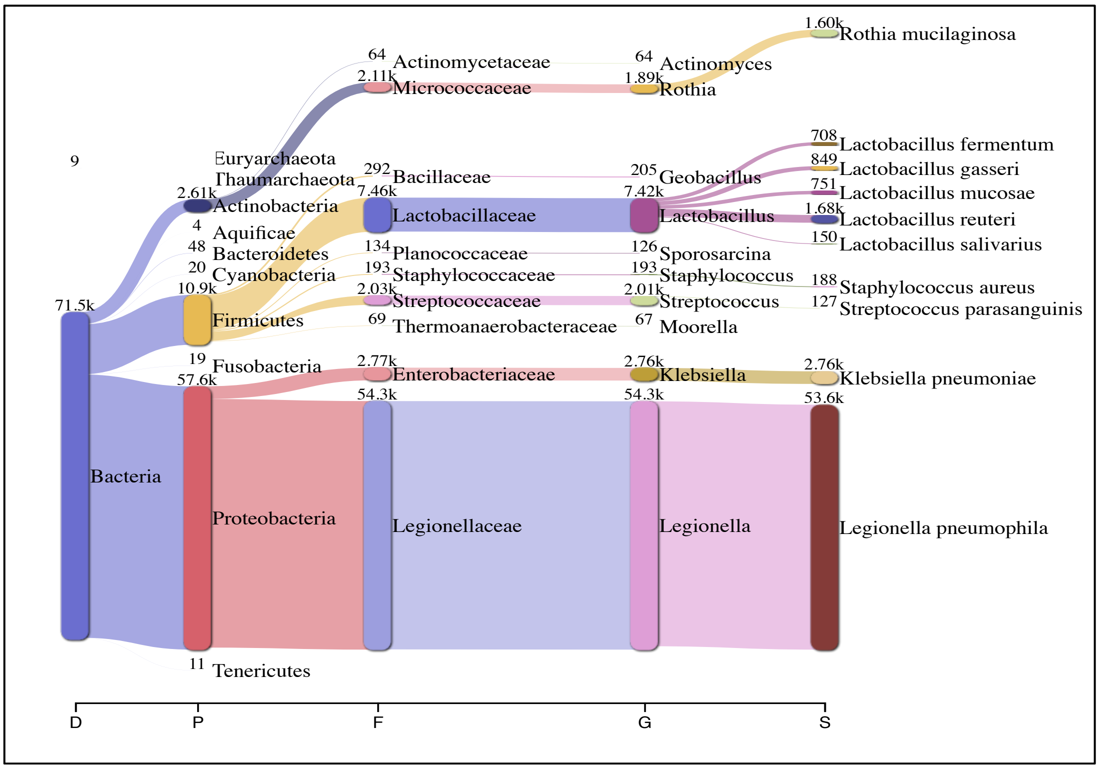


Figure 10. Sample S5

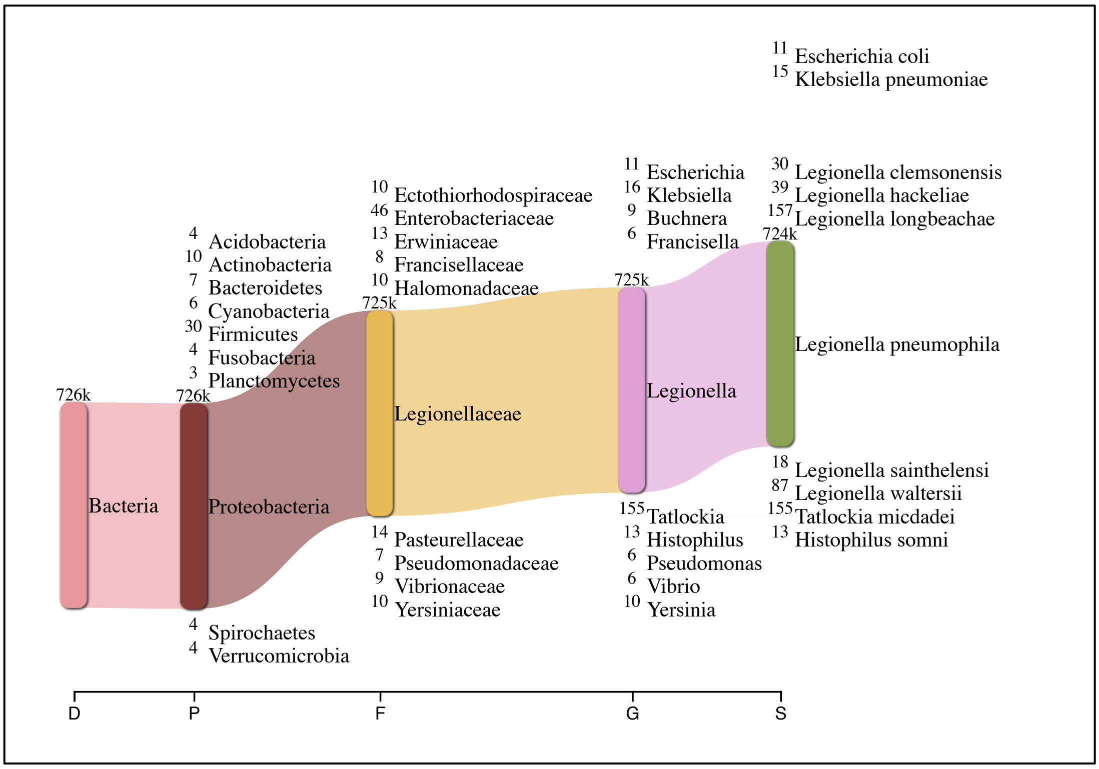


Figure 11. Sample S6

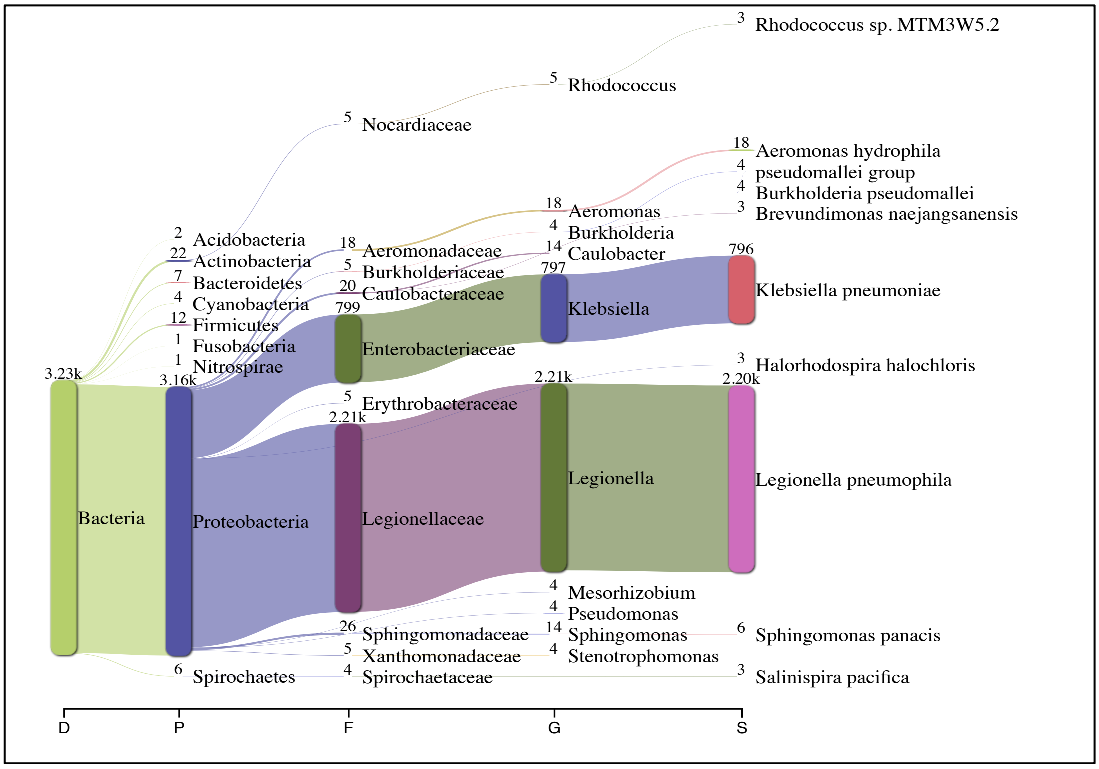


Figure 12. Sample S7

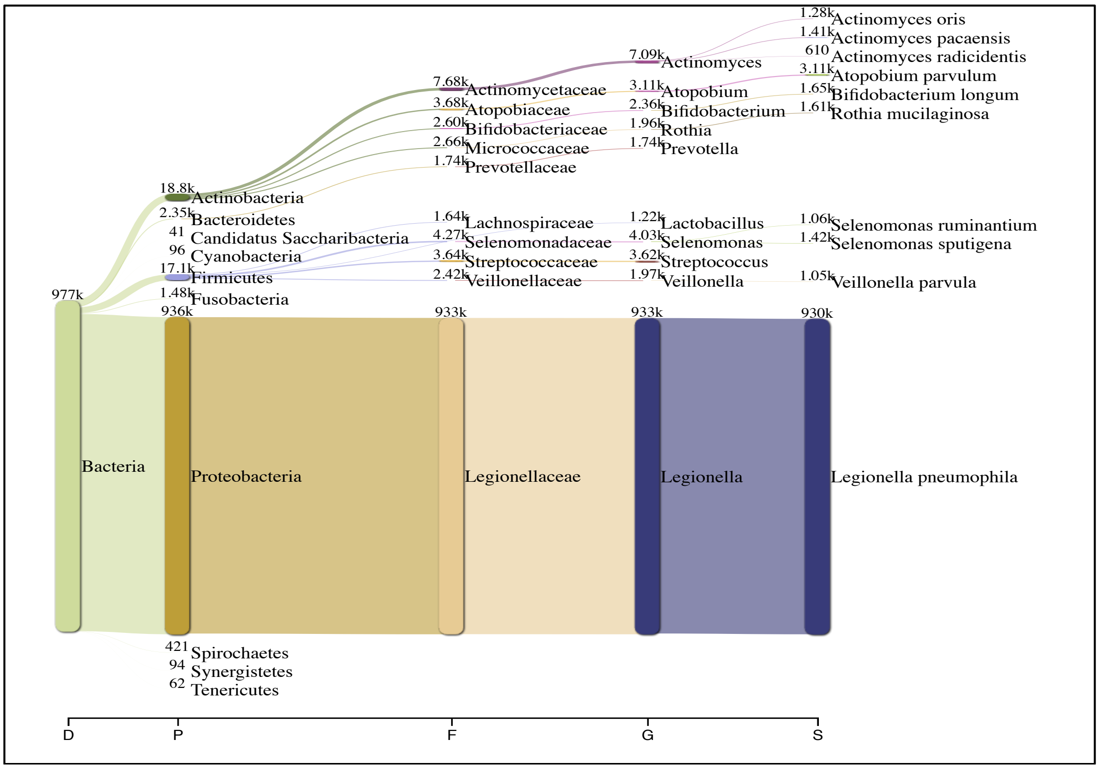


Figure 13. Sample S8

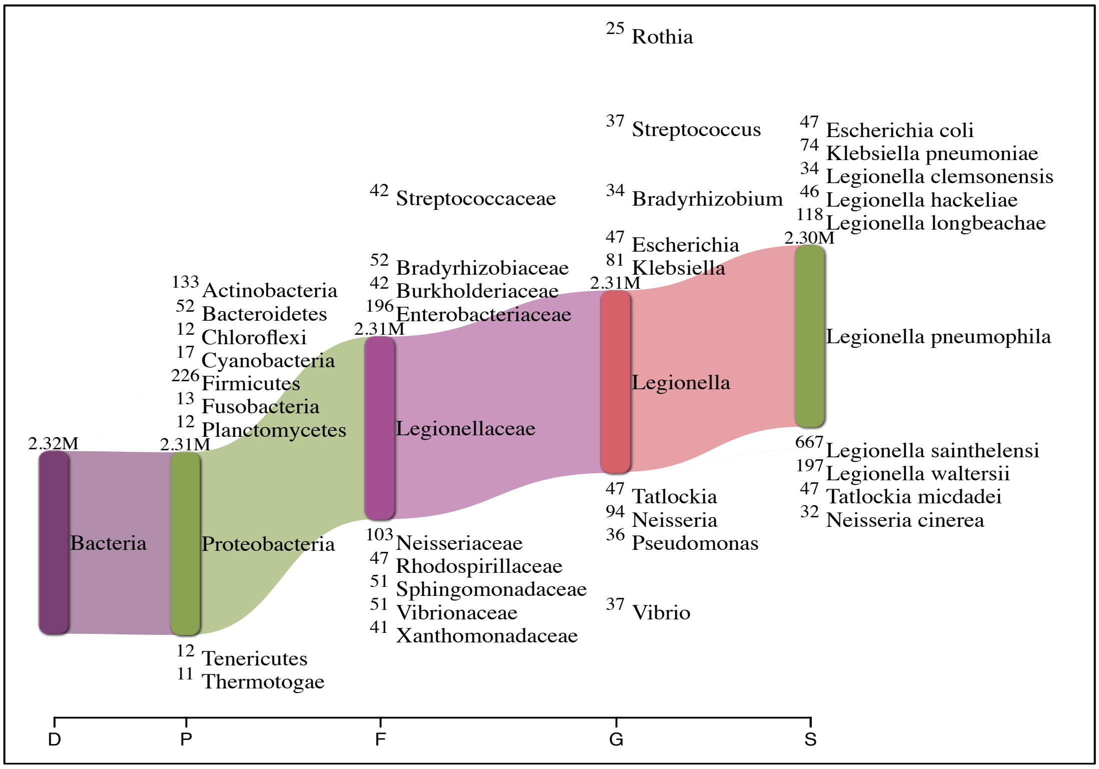


Figure 14. Sample S9

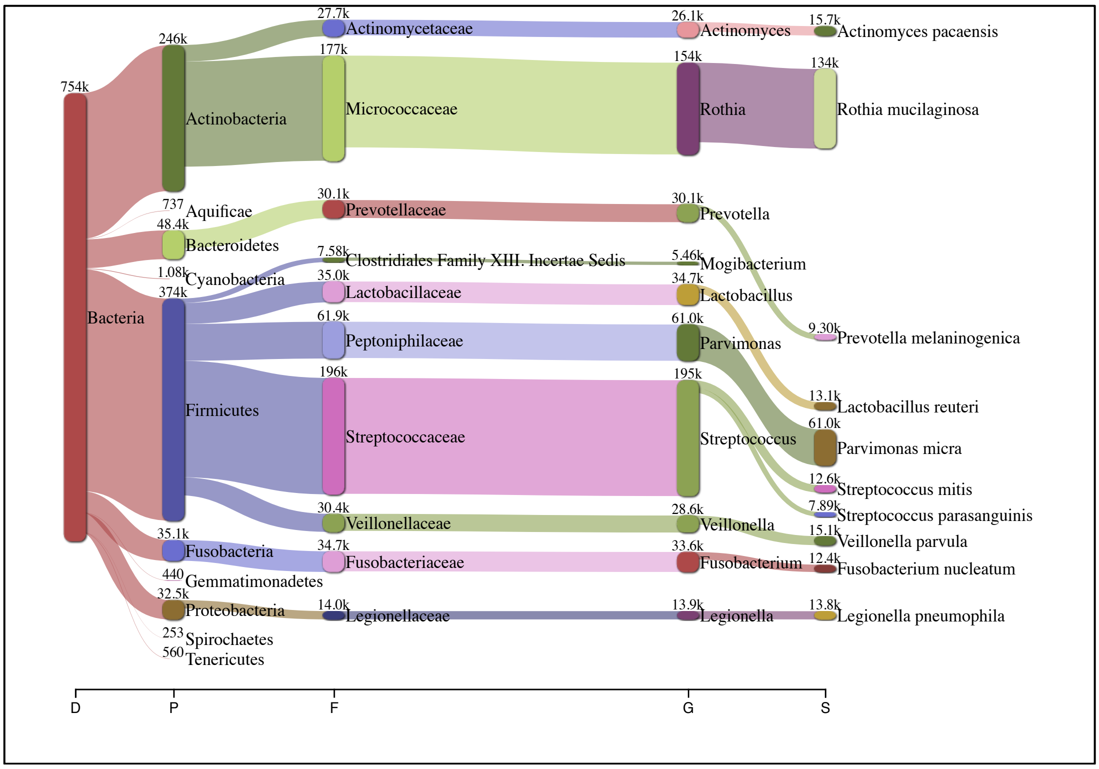


Figure 15. Sample S10

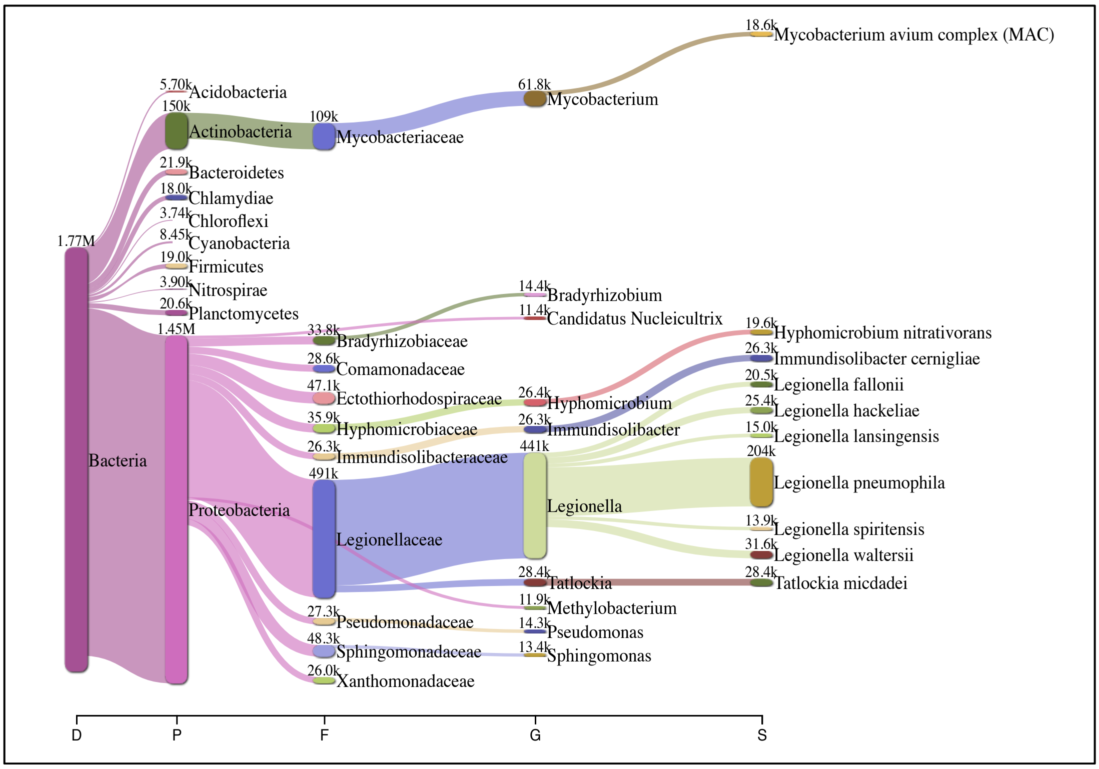


Figure 16. Sample E1

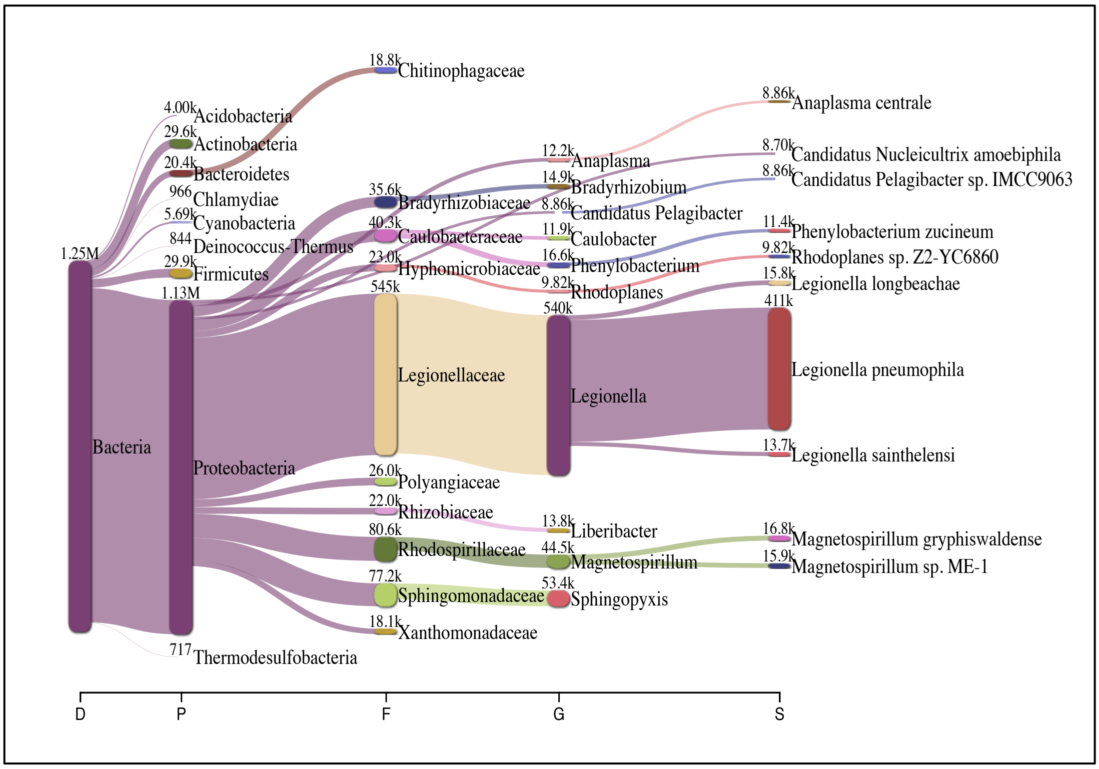


Figure 17. Sample E2

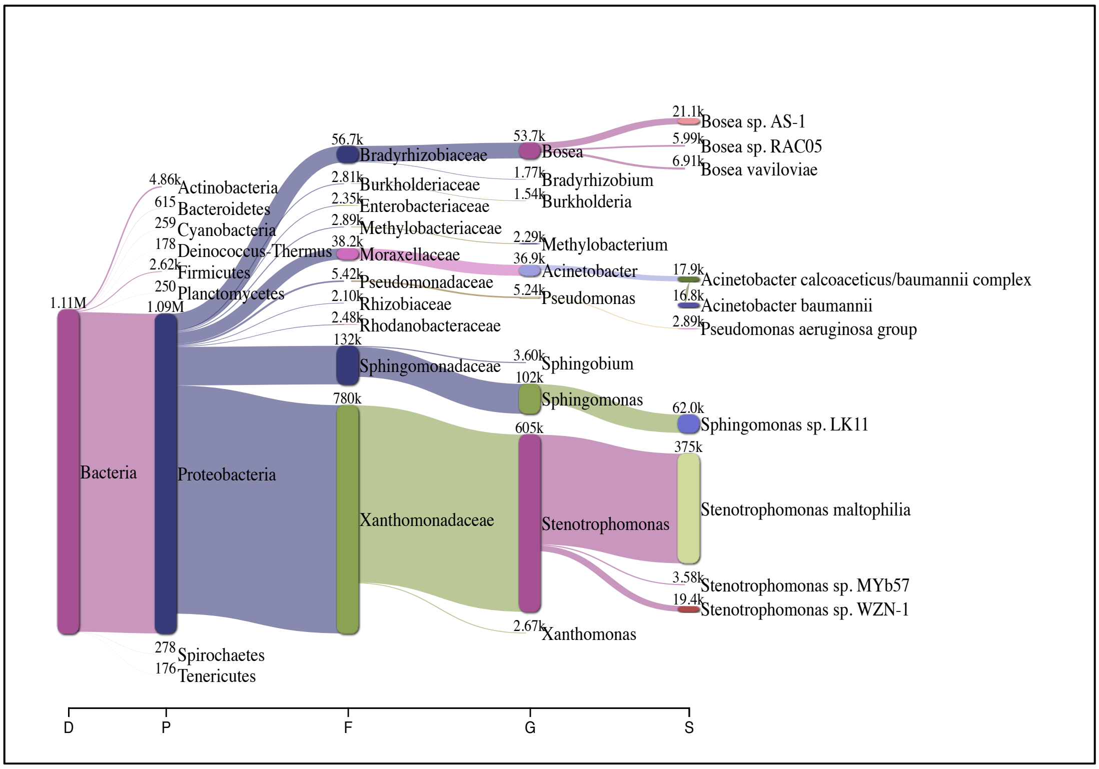


Figure 18. Sample E3

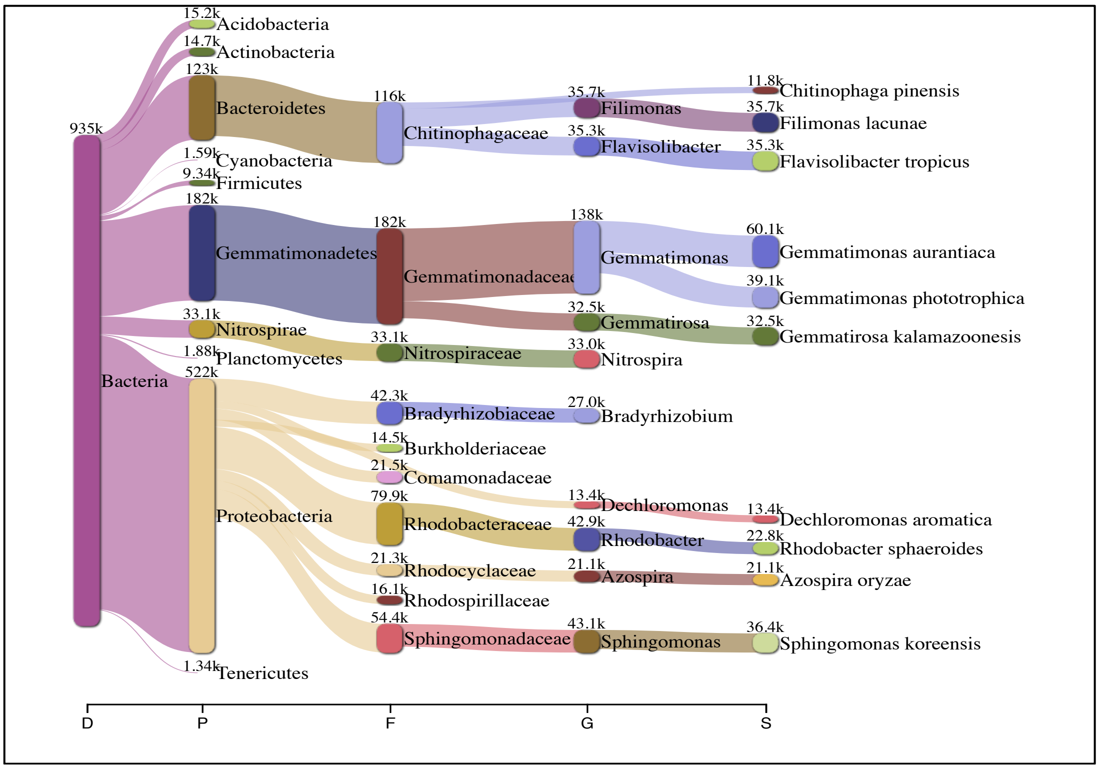


Figure 19. Sample E4

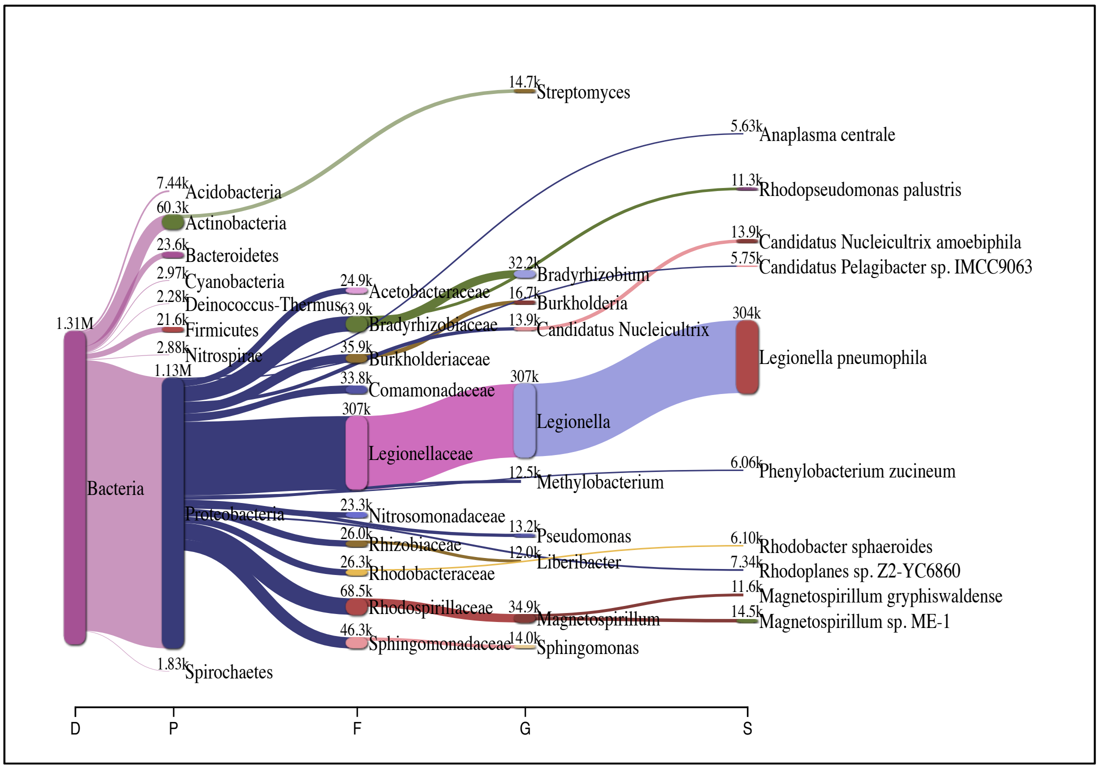


Figure 20. Sample E5

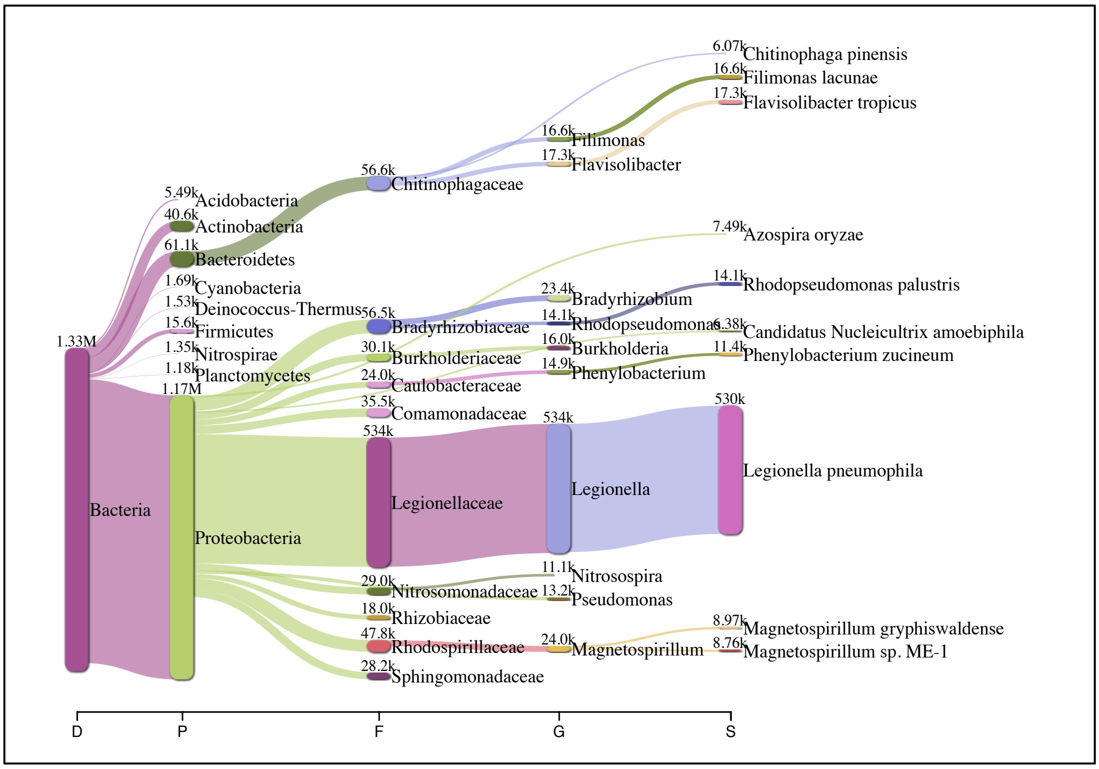


Figure 21. Sample E6

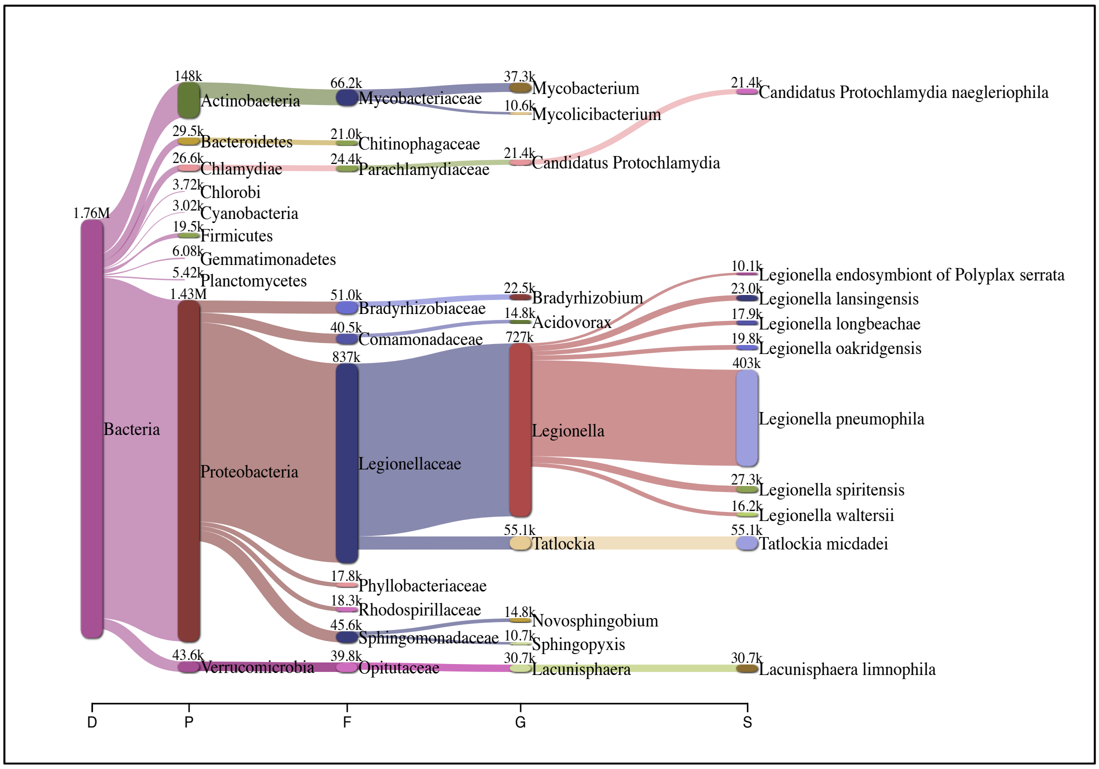


Figure 22. Sample E7

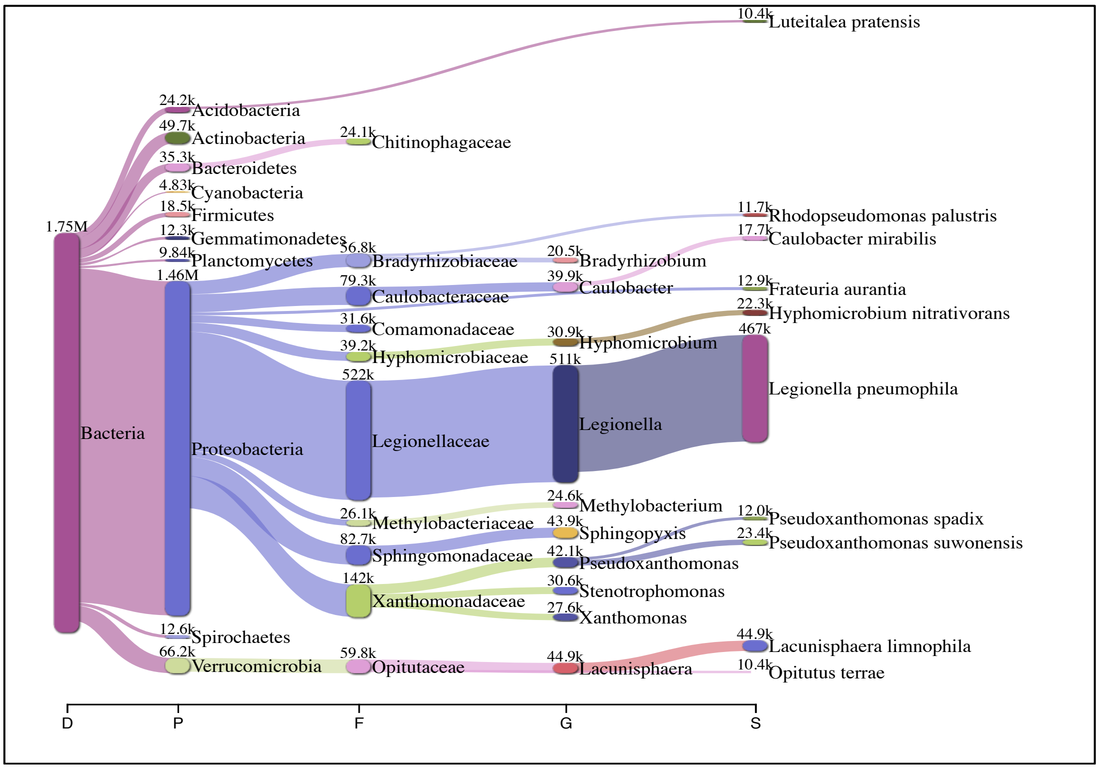


Figure 23. Sample E8

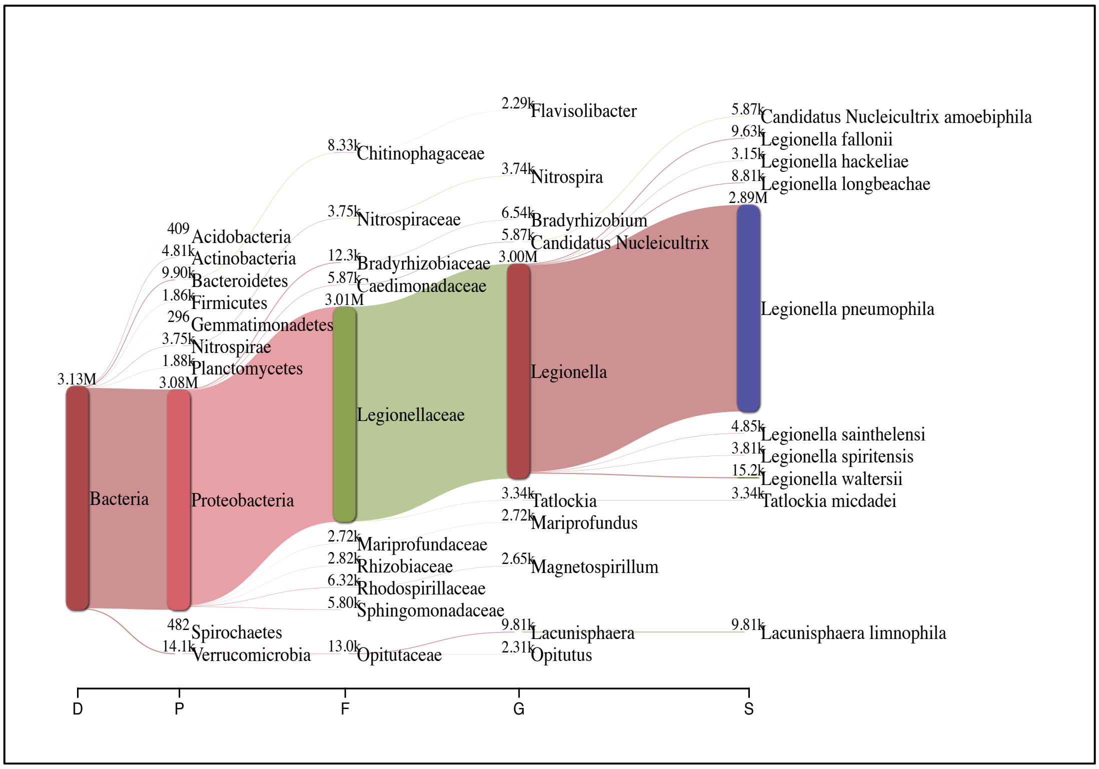


Figure 24. Sample E9

9.7 *L. pneumophila* 50-Gene MLST – Presence/Absence Analysis – Chapter 5

GENE	PRODUCT	D1	D2	D3	D4	Mock	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	E1	E2	E3	E4	E5	E6	E7	E8	E9
lpg0085	hypothetical protein	•	•	•	•	•			•			•		•	•			•			•	•			•
lpg0104	peptide methionine sulfoxide reductase	•	•	•	•	•			•			•		•	•							•			•
lpg0131	dihydropicolinate reductase	•	•	•	•	•			•			•		•	•						•	•			•
lpg0136	pyruvate kinase II	•	•	•	•	•			•			•		•	•		•	•			•	•		•	•
lpg0189	hypothetical protein	•	•	•	•	•			•						•									•	•
lpg0245	NAD-glutamate dehydrogenase	•	•	•	•	•			•			•		•	•						•	•		•	•
lpg0329	50S ribosomal protein L3	•	•	•	•	•						•		•	•		•				•	•			•
lpg0331	50S ribosomal protein L23	•	•	•		•			•			•		•	•		•				•	•		•	•
lpg0409	hypothetical, SURF1 family	•	•	•	•	•			•			•		•	•						•	•		•	•
lpg0419	glucokinase	•	•	•	•	•			•			•		•	•						•	•			•
lpg0525	hypothetical virulence protein	•	•	•	•	•			•			•		•	•						•	•		•	•
lpg0596	hypothetical protein	•	•	•		•			•			•		•	•						•	•		•	•
lpg0601	ABC transporter, permease	•	•	•	•	•			•			•		•	•		•	•			•	•		•	•
lpg0607	lysyl tRNA synthetase	•	•	•	•	•			•			•		•	•		•	•			•	•		•	•
lpg0622	transmembrane protein	•	•	•	•	•						•		•	•			•							•
lpg0664	D-ribulose-5-phosphate-3-epimerase	•	•	•	•	•			•			•		•	•						•	•		•	•
lpg0689	DNA binding stress protein	•	•	•	•	•			•			•		•	•						•	•		•	•
lpg0700	protein-L-isoaspartate-O-methyltransferase	•	•	•	•	•						•		•	•						•	•			•
lpg0812	rod shape determining protein MreC	•	•	•	•	•						•		•	•		•	•			•	•		•	•
lpg0866	3-methyladenine DNA glycosylase	•	•	•	•	•			•			•		•	•			•			•	•			•
lpg0871	hypothetical protein	•	•	•	•	•						•		•	•							•		•	•
lpg0890	cystathionine beta-lyase	•	•	•	•	•			•			•		•	•						•	•		•	•
lpg0957	hypothetical protein	•	•	•	•	•			•			•		•	•		•	•			•	•		•	•

lpg1323	drug resistance transporter, Bcr/CflA	•	•	•	•	•			•			•		•	•		•				•	•		•	•
lpg1503	pyruvate dehydrogenase E2 component	•	•	•	•	•						•		•	•							•			•
lpg1534	glutamate-1-semialdehyde-2,1-aminomutase	•	•	•	•	•			•			•		•	•						•	•			•
lpg1543	transmembrane protein	•	•	•		•			•			•		•	•			•			•	•		•	•
lpg1586	hypothetical protein	•	•	•	•	•			•			•		•	•		•	•			•	•			•
lpg1737	glutamyl/tRNA (Gln) amidotransferase, B subunit	•	•	•		•						•		•	•						•	•		•	•
lpg1744	HesB family protein	•	•	•	•	•						•		•	•		•					•			•
lpg1759	flagellar motor switch protein FliG	•	•	•	•	•			•			•		•	•						•	•		•	•
lpg1811	aspartokinase	•	•	•	•	•			•			•		•	•			•			•	•	•	•	•
lpg1869	ribonuclease III	•	•	•	•	•			•			•		•	•		•	•			•	•			•
lpg1909	hypothetical protein	•	•	•	•	•						•		•	•							•		•	•
lpg2229	saframycin Mx1 synthetase B	•	•	•	•	•			•			•		•	•			•			•	•		•	•
lpg2264	hypothetical protein	•	•	•	•	•						•		•	•		•				•	•	•	•	•
lpg2331	biotin synthase BioC	•	•	•	•	•			•			•		•	•		•	•			•	•			•
lpg2349	alkylhydroperoxidase AhpD family core domain protein	•	•	•	•	•			•			•		•	•						•	•	•		•
lpg2387	plasminogen activator	•	•	•	•	•			•			•		•							•	•		•	•
lpg2494	hypothetical protein	•	•	•	•	•						•		•	•							•			•
lpg2528	alpha-amylase, putative	•	•	•	•	•			•			•		•	•			•			•	•			•
lpg2597	DNA processing enzyme DprA (SMF family)	•	•	•	•	•			•			•		•	•		•	•			•	•			•
lpg2633	hypothetical protein	•	•	•	•	•			•		•	•		•	•		•	•			•	•		•	•
lpg2654	GTP binding protein	•	•	•	•	•			•			•		•	•		•	•			•	•		•	•
lpg2691	cation transporting ATPase PacS	•	•	•		•						•		•	•							•			•
lpg2699	ATPase or kinase	•	•	•		•			•			•		•	•		•	•			•	•			•
lpg2864	hypothetical protein	•	•	•	•	•			•			•		•	•		•				•	•		•	•
lpg2878	cobalt/magnesium uptake transporter	•	•	•	•	•			•			•		•	•			•			•	•		•	•
lpg2882	methionyl tRNA synthetase	•	•	•	•	•			•			•		•	•		•				•	•		•	•
lpg2902	hypothetical protein	•	•	•	•	•			•			•		•	•			•			•	•			•

• = presence

9.8 Core Genes Included in Phylogenetic Trees - Chapter 6

Case Study 1

Phylogenetic Tree1

lpg0004, , lpg0047, lpg0084, lpg0101, lpg0102, lpg0116, lpg0129, lpg0130, lpg0136, lpg0138, lpg0213, lpg0217, lpg0238, lpg0239, lpg0271, lpg0322, lpg0323, lpg0325, lpg0362, lpg0384, lpg0461, lpg0477, lpg0499, lpg0506, lpg0510, lpg0583, lpg0603, lpg0641, lpg0651, lpg0657, lpg0670, lpg0672, lpg0719, lpg0726, lpg0729, lpg0738, lpg0805, lpg0816, lpg0851, lpg0873, lpg0874, lpg0882, lpg0887, lpg0891, lpg0924, lpg0932, lpg0946, lpg0957, lpg0958, lpg0962, lpg0970, lpg1137, lpg1212, lpg1283, lpg1284, lpg1286, lpg1320, lpg1336, lpg1348, lpg1352, lpg1375, lpg1397, lpg1415, lpg1417, lpg1547, lpg1582, lpg1666, lpg1669, lpg1674, lpg1707, lpg1734, lpg1753, lpg1755, lpg1810, lpg1812, lpg1842, lpg1846, lpg1871, lpg1888, lpg1893, lpg2004, lpg2012, lpg2039, lpg2051, lpg2186, lpg2200, lpg2242, lpg2256, lpg2263, lpg2276, lpg2302, lpg2312, lpg2313, lpg2340, lpg2347, lpg2389, lpg2515, lpg2538, lpg2608, lpg2614, lpg2630, lpg2633, lpg2635, lpg2645, lpg2652, lpg2671, lpg2698, lpg2714, lpg2772, lpg2794, lpg2808, lpg2842, lpg2858, lpg2873, lpg2924, lpg2925, lpg2927, lpg2933, lpg2937, lpg2965, lpg2974, lpg2982

Phylogenetic Tree 2

lpg0119, lpg0458, lpg0601, lpg0652, lpg0686, lpg0720, lpg0872, lpg0951, lpg1190, lpg1457, lpg1462, lpg1484, lpg1576, lpg1805, lpg1854, lpg1894, lpg1904, lpg2176, lpg2513, lpg2622, lpg2796, lpg2847, lpg2864, lpg2879, lpg2971

Phylogenetic Tree 3

lpg0024, lpg0027, lpg0047, lpg0079, lpg0101, lpg0118, lpg0119, lpg0140, lpg0175, lpg0294, lpg0384, lpg0404, lpg0421, lpg0449, lpg0456, lpg0461, lpg0469, lpg0483, lpg0530, lpg0532, lpg0559, lpg0577, lpg0601, lpg0608, lpg0611, lpg0624, lpg0626, lpg0627, lpg0652, lpg0660, lpg0678, lpg0679, lpg0692, lpg0704, lpg0748, lpg0800, lpg0802, lpg0803, lpg0804, lpg0818, lpg0822, lpg0826, lpg0829, lpg0833, lpg0838, lpg0851, lpg0872, lpg0924, lpg0936, lpg0937, lpg0941, lpg0954, lpg0966, lpg1143, lpg1164, lpg1166, lpg1189, lpg1202, lpg1214, lpg1221, lpg1225, lpg1278, lpg1285, lpg1291, lpg1302, lpg1304, lpg1306, lpg1320, lpg1349, lpg1352, lpg1358, lpg1363, *lpg1372, lpg1394, lpg1401, lpg1417, lpg1463, lpg1566, lpg1659, lpg1763, lpg1814, lpg1816, lpg1821, lpg1842, lpg1855, lpg1873, lpg1993, lpg2001, lpg2004, lpg2009, lpg2028, lpg2048, lpg2186, lpg2203, lpg2220, lpg2231, lpg2243, lpg2248, lpg2262, lpg2273, lpg2331, lpg2336, lpg2347, lpg2469, lpg2476, lpg2495, lpg2616, lpg2620, lpg2624, lpg2629, lpg2655, lpg2674, lpg2698, lpg2711, lpg2714, lpg2727, lpg2740, lpg2755, lpg2772, lpg2782, lpg2796, lpg2822, lpg2823, lpg2836, lpg2842, lpg2843, lpg2859, lpg2898, lpg2924, lpg2925, lpg2957, lpg2960, lpg2968, lpg2971, lpg2995*

Phylogenetic Tree 4

lpg0010, lpg0116, lpg0129, lpg0212, lpg0248, lpg0323, lpg0410, lpg0456, lpg0497, lpg0557, lpg0583, lpg0612, lpg0626, lpg0630, lpg0686, lpg0745, lpg0905, lpg1225, lpg1346, lpg1397, lpg1419, lpg1504, lpg1582, lpg1597, lpg1701, lpg1830, lpg1910, lpg2260, lpg2280, lpg2735, lpg2861

Phylogenetic Tree 5

lpg0479, lpg1586

Case Study 2

lpg0116, lpg0293, lpg0322, lpg0547, lpg0616, lpg0640, lpg0641, lpg0643, lpg0651, lpg0652 lpg0654, lpg0656, lpg0657, lpg0658, lpg0659, lpg0660, lpg0662, lpg0663, lpg0664, lpg0665 lpg0667, lpg0670, lpg0672, lpg0673, lpg0674, lpg0679, lpg0680, lpg0685, lpg0686, lpg0688 lpg0692, lpg0697, lpg0698, lpg0699, lpg0700, lpg0701, lpg0704, lpg0716, lpg0719, lpg0720 lpg0721, lpg0723, lpg0724, lpg0725, lpg0726, lpg0729, lpg0730, lpg0732, lpg0734, lpg0737 lpg0738, lpg0739, lpg0740, lpg0742, lpg0747, lpg0748, lpg0749, lpg0752, lpg0753, lpg0754 lpg0785, lpg0786, lpg0958, lpg0960, lpg0961, lpg0962, lpg0963, lpg0966, lpg0970, lpg0971 lpg1190, lpg1214, lpg2625, lpg2627, lpg2628, lpg2629, lpg2630, lpg2631, lpg2633, lpg2634 lpg2635, lpg2636, lpg2641, lpg2643, lpg2645, lpg2651, lpg2652

Figure Permission Requests

Figure Number	Source Citation	Copyright holder and contact	License Number from RightsLink	Permission request date	I have permission
Figure 1.1(a)	PHIL 6640, CDC/Dr Barry S. Fields, PhD	Public Domain	NA	Courtesy request to CDC on 15th July, 2019	Yes
Figure 1.1(b)	Abdel-Nour M, Duncan C, Low DE, Guyard C. Biofilms: the stronghold of *Legionella pneumophila*. *Int J Mol Sci*. 2013;14(11):21660–21675.	All IJMS papers are fully open access and distributed under the terms and conditions of the Creative Commons Attribution License (CC BY), i.e., authors retain the copyright of their own paper.	NA	Permission requested by personal communication with Dr. Mena Abdel-Nour and Dr Cyril Guyard on 17th July, 2019	Yes
Figure 1.2	Mercante JW, Winchell JW. Current and emerged *Legionella* diagnostics for laboratory and outbreak investigations. *Clin Microbiol Rev.* 2015;28:95-133	American Society for Microbiology Copyright 2015	NA	15th July, 2019	Yes
Figure 1.3	Hoffmann C, Harrison CF, Hilbi H. The natural alternative: protozoa as cellular models for Legionella infection. *Cell Microbiol.* 2014;16(1):15-26.	John Wiley and Sons Ltd	4631370737170	15th July, 2019	Yes
Figure 1.4	European Centre for Disease Prevention and Control. Legionnaires' disease. In: ECDC. Annual epidemiological report for 2017. Stockholm: ECDC; 2019	European Centre for Disease Control (ECDC)	NA	15th July, 2019	Yes
Figure 1.5	European Centre for Disease Prevention and Control. Legionnaires' disease. In: ECDC. Annual	European Centre for Disease Control (ECDC)	NA	15th July, 2019	Yes

	epidemiological report for 2017. Stockholm: ECDC; 2019				
Figure 1.6	European Centre for Disease Prevention and Control. Legionnaires' disease. In: ECDC. Annual epidemiological report for 2017. Stockholm: ECDC; 2019	European Centre for Disease Control (ECDC)	NA	15th July, 2019	Yes
Figure 1.7	PHE, 2017(a): Legionnaires' disease in residents of England and Wales: 2016. Public Health England, Official Statistics. PHE publications gateway number: 2017685.	Public Health England (PHE)	Open Government License v3.0	NA	Yes
Figure 1.8	Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome*. 2018;6(1):42.	Springer Nature (Microbiome)	Awaiting Response	19th July, 2019	Yes
Figure 1.9	Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome*. 2018;6(1):42.	Spring Nature (Microbiome)	Awaiting Response	19th July, 2019	Yes
Figure 4.1	Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. 2011;13(1):36-46.	Springer Nature (Nature Reviews Genetics)	4630180719524	15th July, 2019	Yes

Figure 4.2	Deininger P. *Alu* elements: know the SINEs. *Genome Biol*. 2011;12(12):236.	Springer Nature (Genome Biology)	4630780586117	15th July, 2019	Yes
Figure 4.3	Britten RJ & Kohne DE. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science.* 1968;161:529–540.	The American Association for the Advancement of Science, Science Journal	4630210537706	15th July, 2019	Yes
Figure 4.4	No reference	NA	NA	NA	NA