



A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak

Journal:	<i>BMJ Open</i>
Manuscript ID:	bmjopen-2012-002175
Article Type:	Research
Date Submitted by the Author:	14-Oct-2012
Complete List of Authors:	Reuter, Sandra; Wellcome Trust Sanger Institute, Harrison, Tim; Health Protection Agency, Köser, Claudio; University of Cambridge, Ellington, Matthew; Health Protection Agency, Smith, Geoffrey; Illumina Inc, Parkhill, Julian; Wellcome Trust Sanger Institute, Peacock, Sharon; University of Cambridge, Bentley, Stephen; Wellcome Trust Sanger Institute, Torok, Estee; University of Cambridge, Department of Medicine
Primary Subject Heading:	Infectious diseases
Secondary Subject Heading:	Infectious diseases
Keywords:	INFECTIOUS DISEASES, Diagnostic microbiology < INFECTIOUS DISEASES, Molecular diagnostics < INFECTIOUS DISEASES, Public health < INFECTIOUS DISEASES

SCHOLARONE™
Manuscripts

only

Title

A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak

Authors

Sandra Reuter¹, Timothy G. Harrison², Claudio U. Köser^{3,4}, Matthew J. Ellington⁴, Geoffrey P. Smith⁵, Julian Parkhill¹, Sharon J. Peacock^{1,3,4,6}, Stephen D. Bentley^{1,3}, M. Estée Török^{3,4,6}

Affiliations

1. The Wellcome Trust Sanger Institute, Hinxton, United Kingdom
2. Respiratory and Systemic Infection Laboratory, Health Protection Agency Centre for Infections, London, United Kingdom
3. Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge, United Kingdom
4. Cambridge Public Health and Microbiology Laboratory, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom
5. Illumina Cambridge Ltd, Saffron Walden, United Kingdom
6. Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom

Corresponding author

Dr M. Estée Török

University of Cambridge, Department of Medicine,

Box 157, Addenbrooke's Hospital, Hills Road,

Cambridge CB2 0QQ, United Kingdom

Tel: +44 (0)1223 217520

Fax: +44 (0)1223 336846

Email: estee.torok@addenbrookes.nhs.uk

Keywords

Legionnaires' disease; *Legionella pneumophila*; outbreak; whole-genome sequencing; typing

Word count

Word count 2371; 1 table; 1 figure

ABSTRACT**Introduction**

Epidemiological investigations of Legionnaires' disease outbreaks rely on the rapid identification and typing of clinical and environmental *Legionella* isolates in order to identify and control the source of infection. Rapid bacterial whole-genome sequencing (WGS) is an emerging technology that has the potential to rapidly discriminate outbreak from non-outbreak isolates in a clinically relevant time frame.

Methods

We performed a pilot study to determine the feasibility of using bacterial WGS to differentiate outbreak from non-outbreak isolates collected during an outbreak of Legionnaires' disease. Seven *Legionella* isolates (three clinical and four environmental) were obtained from the reference laboratory and sequenced using the Illumina MiSeq platform at Addenbrooke's Hospital, Cambridge. Bioinformatic analysis was performed blinded to the epidemiological data at the Wellcome Trust Sanger Institute.

Results

We were able to distinguish outbreak from non-outbreak isolates using bacterial WGS, and to confirm the probable environmental source. Our analysis also highlighted constraints, which were the small number of *Legionella pneumophila* isolates available for sequencing, and the limited number of published genomes for comparison.

Conclusions

We have demonstrated the feasibility of using rapid WGS to investigate an outbreak of Legionnaires' disease. Future work includes building larger genomic databases of *Legionella pneumophila* from both clinical and environmental sources, developing automated data interpretation software, and conducting a cost benefit analysis of WGS versus current typing methods.

ARTICLE SUMMARY**Article focus**

- Epidemiological investigations of Legionnaires' disease outbreaks rely on the rapid identification and typing of clinical and environmental *Legionella pneumophila* isolates in order to identify and control the source of infection
- Rapid bacterial whole genome sequencing (WGS) is an emerging technology that has the ability to identify and discriminate bacterial isolates
- We hypothesised that WGS could be used to discriminate outbreak from non-outbreak *Legionella* isolates in a clinically relevant time frame

Key messages

- We retrospectively applied bacterial WGS to isolates cultured during a previous outbreak investigation, and were able to rapidly distinguish outbreak from non-outbreak isolates, and to identify the probable environmental source
- Our findings were consistent with those of previous epidemiological and microbiological investigations of the same outbreak
- This raises the possibility of conducting combined epidemiological and genomic outbreak investigations in real time

Strengths and limitations of this study

- We have demonstrated the feasibility of using rapid WGS to investigate an outbreak of Legionnaires' disease
- Our study was limited by the small number of *Legionella pneumophila* genomes available for comparison
- Future work includes the development of automated data interpretation software and a cost benefit analysis of current typing methods compared with WGS

MAIN ARTICLE

Introduction

Legionella pneumophila causes outbreaks of respiratory infection in community settings and results in significant morbidity and mortality.¹ The organism is common in aquatic environments and is spread by aerosol from a contaminated source, often cooling towers and other aerosol-producing devices. Nosocomial outbreaks that are related to contaminated water supplies have also been widely reported.²⁻⁴ The diagnosis of Legionnaires' disease (LD) is based on a compatible clinical syndrome and detection of *L. pneumophila* urinary antigen⁵ or isolation of the organism from respiratory specimens, which requires culture on selective media.⁶ Most cases of human infection are caused by *L. pneumophila* serogroup 1. During *Legionella* outbreaks, clinical and environmental isolates are collected and sent to the reference laboratory for typing.⁷ Epidemiological investigations are dependent on the rapid identification and typing of the associated organisms in order to identify and control the source of infection. Current typing methods include phenotypic (monoclonal antibody subgrouping⁸) and genotypic (sequence-based typing⁹) methods, which typically take one to two days. High-throughput sequencing technology has the potential to rapidly provide information on organism identity and genetic relatedness, and has been shown to provide a high degree of discrimination for a range of other bacteria such as methicillin-resistant *Staphylococcus aureus*,¹⁰ *Mycobacterium tuberculosis*,¹¹ *Escherichia coli* O104:H4¹² and *Klebsiella pneumoniae*.¹³ We hypothesised that WGS could be used to discriminate outbreak from non-outbreak isolates of *L. pneumophila* in a comparable time frame, and with a higher level of discrimination, when compared with current typing methods. We therefore conducted a pilot study to determine the feasibility of using a rapid bench-top sequencing platform (Illumina MiSeq) to retrospectively investigate a *Legionella* outbreak.

Objectives

The aim of this pilot study was to determine the feasibility of using bacterial WGS for the investigation of a previous *Legionella* outbreak.

Epidemiological and microbiological investigation

In 2003, an outbreak of LD occurred in Hereford, United Kingdom.⁵ The outbreak started with two community cases that presented with clinical features of infection within a few days of each other, one of whom died. Active case finding identified two further cases in the

1
2
3 local hospital and a formal outbreak investigation was conducted. Twenty-four further cases
4 of LD were identified over the next three weeks. All cases had a positive *L. pneumophila*
5 urinary antigen test, and three patients' samples were culture-positive for *L. pneumophila*
6 serogroup 1. Epidemiological and environmental investigations were undertaken to
7 determine possible sources. One hundred and forty-two environmental samples were
8 collected from potential sources, which included 50 cooling towers on 11 premises.
9 *L. pneumophila* serogroup 1 was isolated from samples collected at three cooling towers at
10 two different locations (sites A and B) and a domestic spa pool. Clinical and environmental
11 isolates were referred to the Respiratory and Systemic Infection Laboratory, Health
12 Protection Agency, London, for *L. pneumophila* monoclonal antibody (mAb) subgrouping
13 followed by a three allele DNA-sequence based typing (SBT₃) method then in use. The SBT₃
14 profiles for two of the clinical isolates and isolates from two of the cooling towers were
15 indistinguishable, suggesting that the cooling towers were the likely environmental source.
16 The strains were subsequently re-examined using the current seven allele sequence based
17 typing (SBT) method,¹⁴ with the same outcome.
18
19
20
21
22
23
24
25
26
27
28

29 **DNA extraction and whole genome sequencing**

30 Seven *L. pneumophila* isolates (three clinical and four environmental) were obtained from
31 the reference laboratory where they had been stored at -80°C with minimal passage since
32 the outbreak. DNA was extracted from each *L. pneumophila* isolate (50ng) and prepared for
33 sequencing using the Nextera DNA Sample Prep Kit (Epicentre). Samples were pooled
34 together and then run on a rapid whole-genome sequencing platform (Illumina MiSeq) at
35 Addenbrooke's Hospital, Cambridge, generating 150bp paired-end reads.
36
37
38
39
40
41

42 **Bioinformatic analysis**

43 Bioinformatic analysis was performed at the Wellcome Trust Sanger Institute and blinded to
44 the epidemiological data. The sequencing data from the seven samples were mapped to a
45 reference genome, *L. pneumophila* type strain Philadelphia-1,¹⁵ and compared with eight
46 other publicly available *L. pneumophila* genomes (Table 1). Sequence reads were mapped
47 onto the reference genome using SMALT. Regions containing phage or insertion sequence
48 elements were excluded. Single nucleotide polymorphisms (SNPs) were identified using a
49 standard approach,¹⁶ by removing SNPs with low quality scores and by filtering for SNPs that
50 were present in at least 75% of the mapped reads. A maximum likelihood phylogeny was
51
52
53
54
55
56
57
58
59
60

estimated using RAxML. The general time-reversible model with gamma correction was used for among-site variation.

RESULTS

Phenotypic and typing results

The microbiological characteristics of the *L. pneumophila* isolates included in this study are summarised in Table 1.

Table 1: Clinical, environmental and reference *L. pneumophila* strains

Sample Number	Accession Number	Biological Origin	Type of sample	Serogroup	Monoclonal antibody subgroup	Sequence type*
Reference genome						
Philadelphia	AE017354.1	United States 1974	Clinical	1	Philadelphia	ST36
Published genomes						
ATCC 43290	CP003192.1	United States	Clinical	12	NA	ST187
Alcoy	CP001828.1	Spain	Clinical	1	ND	ST578
Corby	CP000675.2	United Kingdom	Clinical	1	Knoxville	ST51
Lens	CR628337.1	France	Clinical	1	Benidorm	ST15
130b	FR687201.1	United States	Clinical	1	Benidorm	ST42
Paris	CR628336.1	France	Clinical	1	Philadelphia	ST1
Lorraine	FQ958210.1	France	Clinical	1	ND	ST47
LP_HL06041035	FQ958211.1	France	Environmental	1	ND	ST734
Outbreak investigation isolates						
LP_033	ERS166051	Patient 1	Clinical	1	Philadelphia	ST37
LP_035	ERS166045	Patient 2	Clinical	1	Philadelphia	ST37
LP_617	ERS166047	Patient 3	Clinical	1	Allentown / France	ST47
LP_056	ERS166052	Site A cooling tower 1 (CT1)	Environmental	1	Philadelphia	ST37
LP_427	ERS166050	Site A cooling tower 2 (CT2)	Environmental	1	Philadelphia	ST37
LP_467	ERS166049	Domestic spa pool	Environmental	1	Philadelphia	ST37
LP_423	ERS166048	Site B cooling tower 1 (CT1)	Environmental	1	Oxford / OLDA	ST1

*Sequence type was derived from the genome sequence data and was concordant with the results of the seven allele sequenced based typing method.

NA = Not applicable

ND = not determined

Genomic analysis

Whole genome phylogenetic analysis showed that two clinical isolates (LP033 and LP035) and three environmental isolates (LP056, LP427 and LP467) were closely related genetically, and accordingly clustered together on the tree (Figure 1A). These five isolates were therefore considered to be the outbreak isolates, though it was not possible to obtain directional information from this analysis due to the low number of SNPs differentiating isolates; in total, there were less than 15 SNP differences within the outbreak strain cluster (Figure 1B). Furthermore, the genetic variability between isolates from two cooling tower isolates on Site A, and the observation that these intermingled with the clinical isolates on the tree, suggested that some diversity existed in the source population before the onset of the outbreak. Sequence types were derived from the genome sequence data and confirmed that all five isolates were ST37.

The two remaining isolates (LP423 and LP617) were situated ~75,000 to 77,500 SNPs respectively from the outbreak cluster, and thus were not considered to be part of the outbreak. Sequence types were derived from the genomic data and the clinical isolate (LP617) was ST47 whereas the environmental isolate (LP423) was ST1.

The five outbreak isolates were compared to the nine published strains and found to be most closely related to the Philadelphia-1 strain (which is ST36, a single locus variant of ST37) and to the ATCC 43290 strain (which is ST187) (Figure 1A). Both of these isolates were ~10,000 to 13,000 SNPs distant from the outbreak cluster. The LP617 isolate was 56 SNPs different from Lorraine strain (also ST47), and the LP423 isolate was 906 SNPs different from the Paris strain (also ST1).

Comparison of epidemiological investigation and genomic analysis

Two clinical isolates (LP033 and LP035) had been obtained from patients included in the outbreak. Both strains were located within the outbreak cluster in the phylogenetic tree. The third clinical isolate (LP617) was obtained from a patient who had initially been linked to the outbreak. The original epidemiological investigation found, however, that this patient was a lorry driver, who had passed through Hereford at the time of the outbreak, and had likely acquired his infection elsewhere. This isolate was located distant to the outbreak cluster on the phylogenetic tree, and was therefore not considered to be linked to the outbreak. Thus, for the clinical isolates, the genomic data supported the results of the previous epidemiological investigation.

1
2
3 Three environmental isolates were located within the outbreak cluster. Two of these
4 (LP056 and LP427) had been collected from two cooling towers at the same location (Site A)
5 whilst the third environmental isolate (LP467) had been collected from a spa pool in local
6 domestic premises. Given the small number of SNP differences between these three isolates
7 (Figure 1B) it was not possible to determine which of these isolates represented the source
8 of the outbreak using genomic data alone. The original epidemiological investigation had,
9 however, concluded that the cooling towers on Site A were the most likely source.
10

11
12 The fourth environmental isolate (LP423) was obtained from a cooling tower at a
13 different site (Site B), which was considered epidemiologically unlikely to be the source of
14 the outbreak; a view supported by the typing data. This isolate was located away from the
15 outbreak cluster and was most closely related (906 SNPs different) to the Paris strain (Figure
16 1A).
17
18
19
20
21
22
23

24 **Comparison of conventional typing and genomic analysis**

25 We also compared the results of the conventional typing (monoclonal antibody typing and
26 sequence based typing) with WGS. All of the isolates included in this analysis were
27 *L. pneumophila* serogroup 1, apart from the ATCC 43290 strain, which was serogroup 12. All
28 of the outbreak strains belonged to the mAb subgroup 'Philadelphia', and were ST37. The
29 clinical non-outbreak isolate belonged to the mAb subgroup 'Allentown/France' and was
30 ST47, whereas the environmental non-outbreak isolate belonged to the mAb subgroup
31 'Oxford/OLDA' and was ST1. Thus, in this outbreak, the performance of WGS sequence was
32 equivalent to conventional SBT in differentiating the outbreak from the non-outbreak
33 strains. WGS was unable to distinguish the epidemiologically most likely source of the
34 outbreak (Site A cooling towers) from the domestic spa pool.
35
36
37
38
39
40
41
42

43 **DISCUSSION**

44 Here, we have demonstrated the feasibility of using WGS to perform an investigation of a
45 *Legionella* outbreak. Using genomic analysis we were readily able to distinguish outbreak
46 from non-outbreak *Legionella* isolates, and to identify probable environmental sources, thus
47 supporting the findings of the previous epidemiological investigation. The main advantage of
48 WGS over other typing techniques such as monoclonal antibody typing,⁸ amplified fragment
49 length polymorphism,¹⁷ pulsed-field gel electrophoresis,³ and sequence-based typing⁹ is that
50 it interrogates the whole genome thus giving maximum resolution, even within individual
51 sequence types. Current barriers to routine implementation of WGS include the inability to
52
53
54
55
56
57
58
59
60

1
2
3 sequence directly from clinical specimens, the lack of availability of comprehensive open
4 access genomic databases to compare isolates to, the lack of automated data interpretation
5 software to deliver clinically relevant information, and the need for cost-benefit analyses of
6 WGS versus the current typing methods.
7
8

9 We acknowledge several limitations to our study. The study was performed
10 retrospectively and was hampered by the small number of stored *L. pneumophila* isolates
11 available for WGS. This is also a challenge for contemporaneous outbreak investigations for
12 two reasons. Firstly, the diagnosis of LD is usually made by detection of *L. pneumophila*
13 urinary antigen, and is often not confirmed by culture of the organism from clinical
14 specimens, which takes two to three days. Secondly, environmental samples can take even
15 longer to culture than clinical specimens, and are usually not processed in the same
16 laboratory. Thus the number of clinical and environmental samples available for typing from
17 *Legionella* outbreaks is limited.
18
19
20
21
22
23

24 Our analysis was also constrained by the limited available information on the genetic
25 variation and population structure of *L. pneumophila* at the whole genome level.
26 Environmental and clinical isolates are not evenly distributed in the environment based on
27 sequence-based typing observations, suggesting that clinical isolates are a distinct sub-
28 population of environmental strains. Humans are continuously exposed to environmental
29 *Legionellae* and it is not clear why certain sequence types predominate in human disease.
30 One hypothesis is that disease only occurs in those who have increased susceptibility to
31 infection, for example the elderly, and the immunosuppressed.²¹ Whenever a *Legionella*
32 outbreak occurs it usually reflects the breakdown of *Legionella* control measures, with
33 human infections occurring as a consequence.
34
35
36
37
38
39

40 The genetic diversity of *Legionella* strains within an environmental source, as seen in
41 this analysis, could potentially undermine our ability to link environmental and clinical
42 isolates in an outbreak situation. Thus a detailed epidemiological investigation accompanied
43 by thorough environmental sampling, sequencing and comparison with patient isolates will
44 continue to be required to confirm the likely source of an outbreak.
45
46
47

48 Despite these caveats our work here demonstrates that this WGS approach can
49 provide highly discriminatory information within a clinically relevant time frame, but
50 requires a parallel epidemiological investigation to rule in or rule out potential
51 environmental sources. This heralds the opportunity of conducting combined
52 epidemiological and genomic outbreak investigations in real time, as has been performed
53 for other pathogens.¹⁸
54
55
56
57
58
59
60

Acknowledgements

We would like to acknowledge the authors of the original outbreak investigation and the staff of the Respiratory and Systemic Infection Laboratory, Health Protection Agency.

Study approval

Individual patient consent was not obtained as the study was conducted using stored, anonymized bacterial isolates which had collected at the time of the original outbreak investigation in 2003. Ethical approval was not required as this was a retrospective laboratory-based study using stored anonymized bacterial isolates obtained from a diagnostic archive at the Respiratory and Systemic Infection Laboratory, Health Protection Agency. The study was approved by the Cambridge Health Protection Agency Research and Development Committee and the Cambridge University Hospitals NHS Foundation Trust Research and Development Department.

Funding

This work was supported by grants from the United Kingdom Clinical Research Collaboration (UKCRC) Translational Infection Research Initiative (TIRI); the Medical Research Council (G1000803), with contributions from the Biotechnology and Biological Sciences Research Council, the National Institute for Health Research (NIHR) on behalf of the United Kingdom Department of Health, and the Chief Scientist of the Scottish Government Health Directorate; the Health Protection Agency Strategic Development Research Fund (grant 107514); the NIHR Cambridge Biomedical Research Centre; and the Wellcome Trust (grant number 098051).

Competing interests

The following authors have potential conflicts of interest to declare: GPS (employee and shareholder of Illumina Inc.); JP (travel, accommodation and meeting expenses from Pacific Biosciences and Illumina Ltd); and SJP (consultancy fees from Pfizer).

Data sharing policy

The *L. pneumophila* sequences included in this study have been deposited in the European Nucleotide Archive, under study number ERP001732.

Contributorship statement:

MET, SJP and TGH conceived and designed the study.

CUK and MJE conducted the laboratory experiments.

SR, SDB, JP, SJP and GS analysed and interpreted the data.

SR, TGH, MET wrote the first draft of the manuscript and all authors revised it critically for intellectual content.

All authors reviewed and approved the final manuscript.

REFERENCES

1. Carratala J, Garcia-Vidal C. An update on *Legionella*. *Curr Opin Infect Dis* 2010;23(2):152-7.
2. Tram C, Simonet M, Nicolas MH, et al. Molecular typing of nosocomial isolates of *Legionella pneumophila* serogroup 3. *J Clin Microbiol* 1990;28(2):242-5.
3. Schoonmaker D, Heimberger T, Birkhead G. Comparison of ribotyping and restriction enzyme analysis using pulsed-field gel electrophoresis for distinguishing *Legionella pneumophila* isolates obtained during a nosocomial outbreak. *J Clin Microbiol* 1992;30(6):1491-8.
4. Darelid J, Hallander H, Lofgren S, et al. Community spread of *Legionella pneumophila* serogroup 1 in temporal relation to a nosocomial outbreak. *Scand J Infect Dis* 2001;33(3):194-9.
5. Kirrage D, Reynolds G, Smith GE, et al. Investigation of an outbreak of Legionnaires' disease: Hereford, UK 2003. *Respir Med* 2007;101(8):1639-44.
6. Helbig JH, Bernander S, Castellani Pastoris M, et al. 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):710-6.
7. Fry NK, Alexiou-Daniel S, Bangsberg JM, et al. A multicenter evaluation of genotypic methods for the epidemiologic typing of *Legionella pneumophila* serogroup 1: results of a pan-European study. *Clin Microbiol Infect* 1999;5(8):462-77.
8. Brindle RJ, Stannett PJ, Tobin JO. *Legionella pneumophila*: monoclonal antibody typing of clinical and environmental isolates. *Epidemiol Infect* 1987;99(2):235-9.
9. Gaia V, Fry NK, Afshar B, et al. Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila*. *J Clin Microbiol* 2005;43(5):2047-52.
10. Köser CU, Holden MT, Ellington MJ, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* 2012;366(24):2267-75.
11. Gardy JL, Johnston JC, Ho Sui SJ, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 2011;364(8):730-9.
12. Rohde H, Qin J, Cui Y, et al. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N Engl J Med* 2011;365(8):718-24.
13. Snitkin ES, Zelazny AM, Thomas PJ, et al. Tracking a Hospital Outbreak of Carbapenem-Resistant *Klebsiella pneumoniae* with Whole-Genome Sequencing. *Sci Transl Med* 2012;4(148):148ra16.
14. Gaia V, Fry NK, Harrison TG, et al. Sequence-based typing of *Legionella pneumophila* serogroup 1 offers the potential for true portability in legionellosis outbreak investigation. *J Clin Microbiol* 2003;41(7):2932-9.
15. Chien M, Morozova I, Shi S, et al. The genomic sequence of the accidental pathogen *Legionella pneumophila*. *Science* 2004;305(5692):1966-8.

16. Harris SR, Feil EJ, Holden MT, et al. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 2010;327(5964):469-74.
17. Fry NK, Bangsberg JM, Bergmans A, et al. Designation of the European Working Group on *Legionella* Infection (EWGLI) amplified fragment length polymorphism types of *Legionella pneumophila* serogroup 1 and results of intercentre proficiency testing Using a standard protocol. *Eur J Clin Microbiol Infect Dis* 2002;21(10):722-8.
18. Köser CU, Ellington MJ, Cartwright EJ, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog* 2012;8(8):e1002824.
19. Harrison TG, Afshar B, Doshi N, et al. 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):781-91.
20. Cazalet C, Jarraud S, Ghavi-Helm Y, et al. Multigenome analysis identifies a worldwide distributed epidemic *Legionella pneumophila* clone that emerged within a highly diverse species. *Genome Res* 008;18(3):431-41.
21. Ampel NM, Wing EJ. *Legionella* infection in transplant patients. *Semin Respir Infect* 1990;5(1):30-7.

1
2
3 **Figure 1. Phylogenetic tree of *Legionella pneumophila* strains**
4

5 A. Phylogeny of the species *L. pneumophila*. Clinical, environmental and references isolates
6 are shown in red, blue, and black, respectively. Inset B. Close-up phylogeny of the isolates
7 involved in the outbreak. The branch leading to the reference strain Philadelphia has been
8 truncated for clarity.
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

A

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

BMJ Open

B

LP_Philadelphia
LP_ATCC43290
LP_056
LP_427
LP_033
LP_035
LP_467

LP Philadelphia
LP056
Site A cooling tower 1

LP035
Patient 2

LP467
Domestic Spa Pool

LP427
Site A cooling tower 2

LP033
Patient 1

1 SNP

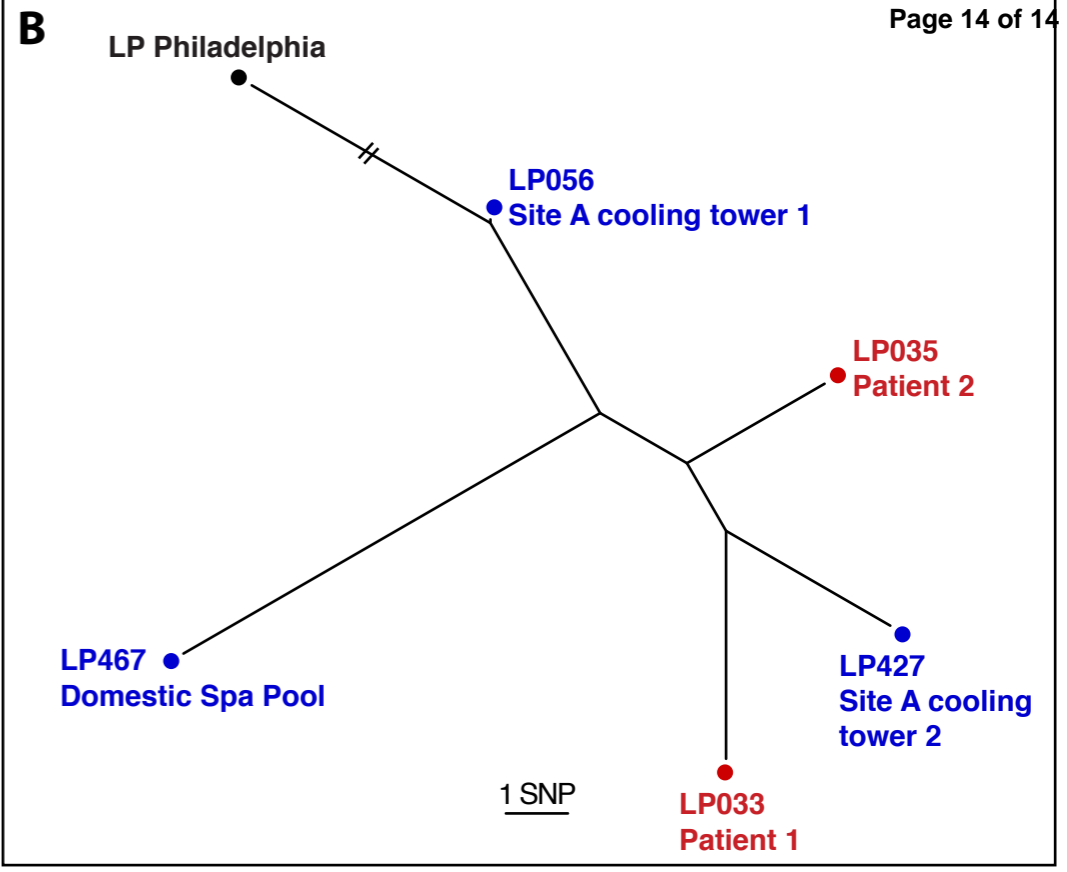
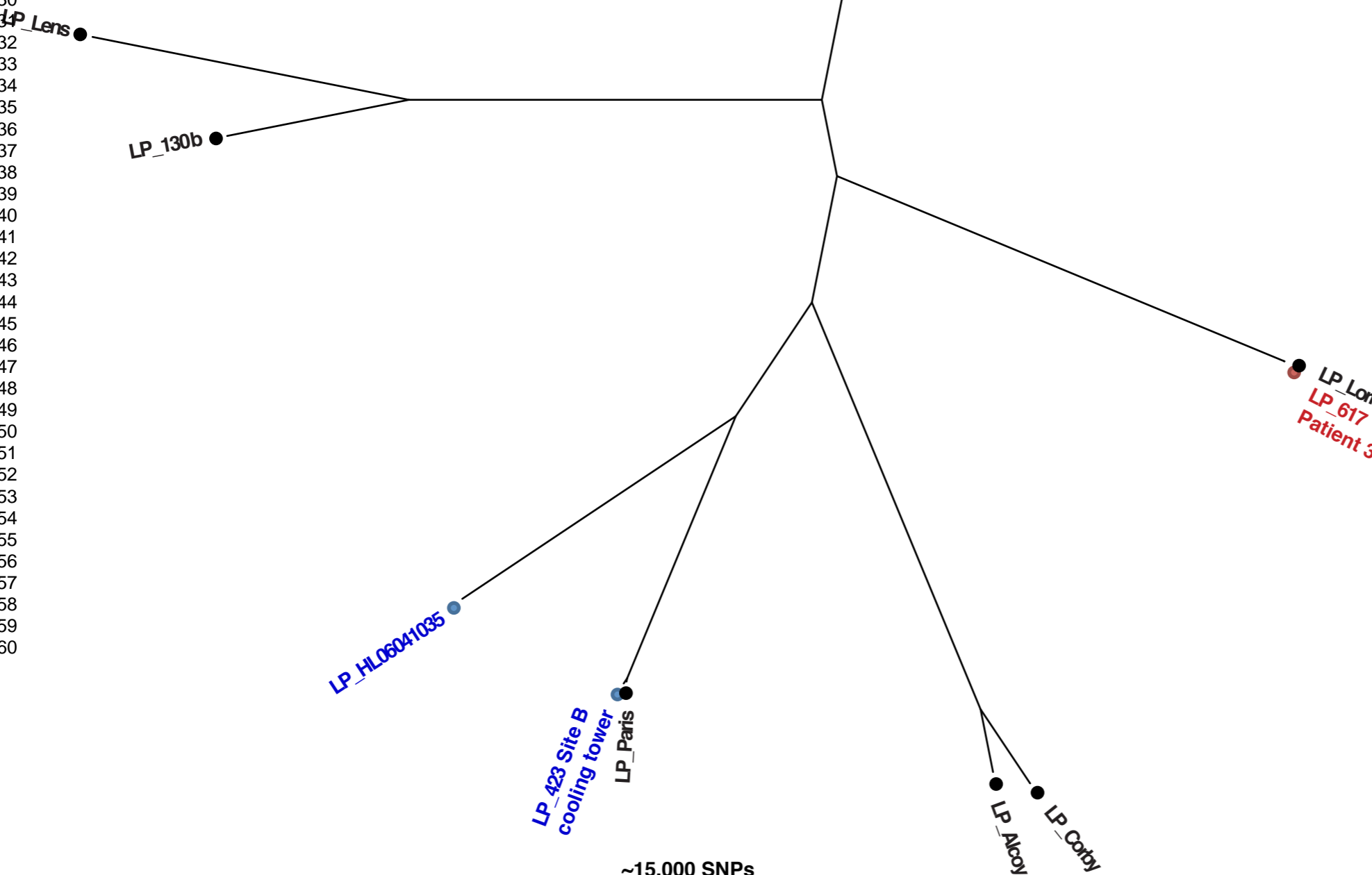
LP Lorraine
LP_617
Patient 3

LP_HL06041035

LP_423 Site B
cooling tower
LP_Paris

LP_Alcoy
LP_Corby

~15,000 SNPs





A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak

Journal:	<i>BMJ Open</i>
Manuscript ID:	bmjopen-2012-002175.R1
Article Type:	Research
Date Submitted by the Author:	27-Nov-2012
Complete List of Authors:	Reuter, Sandra; Wellcome Trust Sanger Institute, Harrison, Tim; Health Protection Agency, Köser, Claudio; University of Cambridge, Ellington, Matthew; Health Protection Agency, Smith, Geoffrey; Illumina Inc, Parkhill, Julian; Wellcome Trust Sanger Institute, Peacock, Sharon; University of Cambridge, Bentley, Stephen; Wellcome Trust Sanger Institute, Torok, Estee; University of Cambridge, Department of Medicine
Primary Subject Heading:	Infectious diseases
Secondary Subject Heading:	Genetics and genomics
Keywords:	INFECTIOUS DISEASES, Diagnostic microbiology < INFECTIOUS DISEASES, Molecular diagnostics < INFECTIOUS DISEASES, Public health < INFECTIOUS DISEASES

SCHOLARONE™
Manuscripts

only

Title

A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak

Authors

Sandra Reuter¹, Timothy G. Harrison², Claudio U. Köser^{3,4}, Matthew J. Ellington⁴, Geoffrey P. Smith⁵, Julian Parkhill¹, Sharon J. Peacock^{1,3,4,6}, Stephen D. Bentley^{1,3}, M. Estée Török^{3,4,6}

Affiliations

1. The Wellcome Trust Sanger Institute, Hinxton, United Kingdom
2. Respiratory and Systemic Infection Laboratory, Health Protection Agency Centre for Infections, London, United Kingdom
3. Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge, United Kingdom
4. Cambridge Public Health and Microbiology Laboratory, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom
5. Illumina Cambridge Ltd, Saffron Walden, United Kingdom
6. Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom

Corresponding author

Dr M. Estée Török

University of Cambridge, Department of Medicine,

Box 157, Addenbrooke's Hospital, Hills Road,

Cambridge CB2 0QQ, United Kingdom

Tel: +44 (0)1223 217520

Fax: +44 (0)1223 336846

Email: estee.torok@addenbrookes.nhs.uk

Keywords

Legionnaires' disease; *Legionella pneumophila*; outbreak; whole-genome sequencing; typing

Word count

Word count 2343; 1 table; 1 figure

ABSTRACT**Introduction**

Epidemiological investigations of Legionnaires' disease outbreaks rely on the rapid identification and typing of clinical and environmental *Legionella* isolates in order to identify and control the source of infection. Rapid bacterial whole-genome sequencing (WGS) is an emerging technology that has the potential to rapidly discriminate outbreak from non-outbreak isolates in a clinically relevant time frame.

Methods

We performed a pilot study to determine the feasibility of using bacterial WGS to differentiate outbreak from non-outbreak isolates collected during an outbreak of Legionnaires' disease. Seven *Legionella* isolates (three clinical and four environmental) were obtained from the reference laboratory and sequenced using the Illumina MiSeq platform at Addenbrooke's Hospital, Cambridge. Bioinformatic analysis was performed blinded to the epidemiological data at the Wellcome Trust Sanger Institute.

Results

We were able to distinguish outbreak from non-outbreak isolates using bacterial WGS, and to confirm the probable environmental source. Our analysis also highlighted constraints, which were the small number of *Legionella pneumophila* isolates available for sequencing, and the limited number of published genomes for comparison.

Conclusions

We have demonstrated the feasibility of using rapid WGS to investigate an outbreak of Legionnaires' disease. Future work includes building larger genomic databases of *Legionella pneumophila* from both clinical and environmental sources, developing automated data interpretation software, and conducting a cost-benefit analysis of WGS versus current typing methods.

ARTICLE SUMMARY**Article focus**

- Epidemiological investigations of Legionnaires' disease outbreaks rely on the rapid identification and typing of clinical and environmental *Legionella pneumophila* isolates in order to identify and control the source of infection
- Rapid bacterial whole genome sequencing (WGS) is an emerging technology that has the ability to identify and discriminate bacterial isolates
- We hypothesised that WGS could be used to discriminate outbreak from non-outbreak *Legionella* isolates in a clinically relevant time frame

Key messages

- We retrospectively applied bacterial WGS to isolates cultured during a previous outbreak investigation, and were able to rapidly distinguish outbreak from non-outbreak isolates, and to identify the probable environmental source
- Our findings were consistent with those of previous epidemiological and microbiological investigations of the same outbreak
- This raises the possibility of conducting combined epidemiological and genomic outbreak investigations in real time

Strengths and limitations of this study

- We have demonstrated the feasibility of using rapid WGS to investigate an outbreak of Legionnaires' disease
- Our study was limited by the small number of *Legionella pneumophila* genomes available for comparison
- Future work includes the development of automated data interpretation software and a cost-benefit analysis of current typing methods compared with WGS

MAIN ARTICLE

Introduction

Legionella pneumophila causes outbreaks of respiratory infection in community settings and results in significant morbidity and mortality.¹ The organism is common in aquatic environments and is spread by aerosol from a contaminated source, often cooling towers and other aerosol-producing devices. Nosocomial outbreaks that are related to contaminated water supplies have also been widely reported.²⁻⁴ The diagnosis of Legionnaires' disease (LD) is based on a compatible clinical syndrome and detection of *L. pneumophila* urinary antigen⁵ or isolation of the organism from respiratory specimens, which requires culture on selective media.⁶ Most cases of human infection are caused by *L. pneumophila* serogroup 1. During *Legionella* outbreaks, clinical and environmental isolates are collected and sent to the reference laboratory for typing.⁷ Epidemiological investigations are dependent on the rapid identification and typing of the associated organisms in order to identify and control the source of infection. Current typing methods include phenotypic (monoclonal antibody subgrouping⁸) and genotypic (sequence-based typing⁹) methods, which typically take one to two days. High-throughput sequencing technology has the potential to rapidly provide information on organism identity and genetic relatedness, and has been shown to provide a high degree of discrimination for a range of other bacteria such as methicillin-resistant *Staphylococcus aureus*,¹⁰ *Mycobacterium tuberculosis*,¹¹ *Escherichia coli* O104:H4¹² and *Klebsiella pneumoniae*.¹³ We hypothesised that WGS could be used to discriminate outbreak from non-outbreak isolates of *L. pneumophila* in a comparable time frame, and with a higher level of discrimination, when compared with current typing methods. Therefore we conducted a pilot study to determine the feasibility of using a rapid bench-top sequencing platform (Illumina MiSeq) to retrospectively investigate a *Legionella* outbreak.

Objectives

The aim of this pilot study was to determine the feasibility of using bacterial WGS for the investigation of a previous *Legionella* outbreak.

Epidemiological and microbiological investigation

In 2003, an outbreak of LD occurred in Hereford, United Kingdom.¹⁴ The outbreak started with two community cases that presented with clinical features of infection within a few days of each other, one of whom died. Active case finding identified two further cases in the

1
2
3 local hospital and a formal outbreak investigation was carried out. Twenty-four further cases
4 of LD were identified over the next three weeks. All cases had a positive *L. pneumophila*
5 urinary antigen test, and three patients' samples were culture-positive for *L. pneumophila*
6 serogroup 1. Epidemiological and environmental investigations were undertaken to
7 determine possible sources. One hundred and forty-two environmental samples were
8 collected from potential sources, which included 50 cooling towers on 11 premises.
9 *L. pneumophila* serogroup 1 was isolated from samples collected at three cooling towers at
10 two different locations (sites A and B) and a domestic spa pool. Clinical and environmental
11 isolates were referred to the Respiratory and Systemic Infection Laboratory, Health
12 Protection Agency, London, for *L. pneumophila* monoclonal antibody (mAb) subgrouping
13 followed by a three allele DNA-sequence based typing (SBT₃) method then in use. The SBT₃
14 profiles for two of the clinical isolates and isolates from two of the cooling towers were
15 indistinguishable, suggesting that the cooling towers were the likely environmental source.
16 The strains were subsequently re-examined using the current seven allele sequence based
17 typing (SBT) method,¹⁵ with the same outcome.
18
19
20
21
22
23
24
25
26
27
28

29 **DNA extraction and whole genome sequencing**

30 Seven *L. pneumophila* isolates (three clinical and four environmental) were obtained from
31 the reference laboratory where they had been stored at -80°C with minimal passage since
32 the outbreak. DNA was extracted from each *L. pneumophila* isolate (50ng) and prepared for
33 sequencing using the Nextera DNA Sample Prep Kit (Epicentre). Samples were pooled
34 together and then run on a rapid whole-genome sequencing platform (Illumina MiSeq) at
35 Addenbrooke's Hospital, Cambridge, generating 150bp paired-end reads.
36
37
38
39
40
41

42 **Bioinformatic analysis**

43 Bioinformatic analysis was performed at the Wellcome Trust Sanger Institute and blinded to
44 the epidemiological data. The sequencing data from the seven samples were mapped to a
45 reference genome, *L. pneumophila* type strain Philadelphia-1,¹⁶ and compared with eight
46 other publicly available *L. pneumophila* genomes (Table 1). Sequence reads were mapped
47 onto the reference genome using SMALT. Regions containing phage or insertion sequence
48 elements were excluded from the analysis. Single nucleotide polymorphisms (SNPs) were
49 identified using a standard approach,¹⁷ by removing SNPs with low quality scores and by
50 filtering for SNPs that were present in at least 75% of the mapped reads. The minimum
51 number of high quality reads mapping to call a base was set to four, which is equivalent to a
52
53
54
55
56
57
58
59
60

minimum coverage of four. Actual coverage ranged between 20x and 100x per isolate. A maximum likelihood phylogeny was estimated using RAxML. The general time-reversible model with gamma correction was used for among-site variation. Tandem repeats were not considered in the original analysis, although we did re-run the analysis excluding the 23 repetitive genes mentioned in the paper by Coil and colleagues;¹⁸ the overall topology of the phylogenetic tree remained unchanged and would not have affected interpretation of our data.

RESULTS

Phenotypic and typing results

The microbiological characteristics of the *L. pneumophila* isolates included in this study are summarised in Table 1.

Table 1: Clinical, environmental and reference *L. pneumophila* strains

Sample Number	Accession Number	Biological Origin	Type of sample	Serogroup	Monoclonal antibody subgroup	Sequence type*
Reference genome						
Philadelphia	AE017354.1	United States 1974	Clinical	1	Philadelphia	ST36
Published genomes						
ATCC 43290	CP003192.1	United States	Clinical	12	NA	ST187
Alcoy	CP001828.1	Spain	Clinical	1	ND	ST578
Corby	CP000675.2	United Kingdom	Clinical	1	Knoxville	ST51
Lens	CR628337.1	France	Clinical	1	Benidorm	ST15
130b	FR687201.1	United States	Clinical	1	Benidorm	ST42
Paris	CR628336.1	France	Clinical	1	Philadelphia	ST1
Lorraine	FQ958210.1	France	Clinical	1	ND	ST47
LP_HL06041035	FQ958211.1	France	Environmental	1	ND	ST734
Outbreak investigation isolates						
LP_033	ERS166051	Patient 1	Clinical	1	Philadelphia	ST37
LP_035	ERS166045	Patient 2	Clinical	1	Philadelphia	ST37
LP_617	ERS166047	Patient 3	Clinical	1	Allentown / France	ST47
LP_056	ERS166052	Site A cooling tower 1 (CT1)	Environmental	1	Philadelphia	ST37
LP_427	ERS166050	Site A cooling tower 2 (CT2)	Environmental	1	Philadelphia	ST37
LP_467	ERS166049	Domestic spa pool	Environmental	1	Philadelphia	ST37
LP_423	ERS166048	Site B cooling tower 1	Environmental	1	Oxford / OLDA	ST1

		(CT1)				
--	--	-------	--	--	--	--

*Sequence type was derived from the genome sequence data and was concordant with the results of the seven allele sequenced based typing method.

NA = Not applicable

ND = not determined

Genomic analysis

Whole genome phylogenetic analysis showed that two clinical isolates (LP033 and LP035) and three environmental isolates (LP056, LP427 and LP467) were closely related genetically, and accordingly clustered together on the tree (Figure 1A). These five isolates were therefore considered to be the outbreak isolates, though it was not possible to obtain directional information from this analysis due to the low number of SNPs differentiating isolates; in total, there were less than 15 SNP differences within the outbreak strain cluster (Figure 1B). Furthermore, the genetic variability between isolates from two cooling tower isolates on Site A, and the observation that these intermingled with the clinical isolates on the tree, suggested that some diversity existed in the source population before the onset of the outbreak. Sequence types were derived from the genome sequence data and confirmed that all five isolates were ST37.

The two remaining isolates (LP423 and LP617) were situated ~75,000 to 77,500 SNPs respectively from the outbreak cluster, and thus were not considered to be part of the outbreak. Sequence types were derived from the genomic data and the clinical isolate (LP617) was ST47 whereas the environmental isolate (LP423) was ST1.

The five outbreak isolates were compared to the nine published strains and found to be most closely related to the Philadelphia-1 strain (which is ST36, a single locus variant of ST37) and to the ATCC 43290 strain (which is ST187) (Figure 1A). Both of these isolates were ~10,000 to 13,000 SNPs distant from the outbreak cluster. The LP617 isolate was 56 SNPs different from Lorraine strain (also ST47), and the LP423 isolate was 906 SNPs different from the Paris strain (also ST1).

Comparison of epidemiological investigation and genomic analysis

Two clinical isolates (LP033 and LP035) had been obtained from patients included in the outbreak. Both strains were located within the outbreak cluster in the phylogenetic tree. The third clinical isolate (LP617) was obtained from a patient who had initially been linked to the outbreak. The original epidemiological investigation found, however, that this patient was a lorry driver, who had passed through Hereford at the time of the outbreak, and had likely acquired his infection elsewhere. This isolate was located distant to the outbreak cluster on the phylogenetic tree, and was therefore not considered to be linked to the

1
2
3 outbreak. Thus, for the clinical isolates, the genomic data supported the results of the
4 previous epidemiological investigation.
5

6 Three environmental isolates were located within the outbreak cluster. Two of these
7 (LP056 and LP427) had been collected from two cooling towers at the same location (Site A)
8 whilst the third environmental isolate (LP467) had been collected from a spa pool in local
9 domestic premises. Given the small number of SNP differences between these three isolates
10 (Figure 1B) it was not possible to determine which of these isolates represented the source
11 of the outbreak using genomic data alone. The original epidemiological investigation had,
12 however, concluded that the cooling towers on Site A were the most likely source.
13
14
15
16

17 The fourth environmental isolate (LP423) was obtained from a cooling tower at a
18 different site (Site B), which was considered epidemiologically unlikely to be the source of
19 the outbreak; a view supported by the typing data. This isolate was located away from the
20 outbreak cluster and was most closely related (906 SNPs different) to the Paris strain (Figure
21 1A).
22
23
24
25
26

27 **Comparison of conventional typing and genomic analysis**

28 We also compared the results of the conventional typing (monoclonal antibody typing and
29 sequence based typing) with WGS. All of the isolates included in this analysis were
30 *L. pneumophila* serogroup 1, apart from the ATCC 43290 strain, which was serogroup 12. All
31 of the outbreak strains belonged to the mAb subgroup 'Philadelphia', and were ST37. The
32 clinical non-outbreak isolate belonged to the mAb subgroup 'Allentown/France' and was
33 ST47, whereas the environmental non-outbreak isolate belonged to the mAb subgroup
34 'Oxford/OLDA' and was ST1. Thus, in this outbreak, the performance of WGS sequence was
35 equivalent to conventional SBT in differentiating the outbreak from the non-outbreak
36 strains. WGS was unable to distinguish the epidemiologically most likely source of the
37 outbreak (Site A cooling towers) from the domestic spa pool.
38
39
40
41
42
43
44
45
46

47 **DISCUSSION**

48 Here, we have demonstrated the feasibility of using WGS to perform an investigation of a
49 *Legionella* outbreak. Using genomic analysis we were readily able to distinguish outbreak
50 from non-outbreak *Legionella* isolates, and to identify probable environmental sources, thus
51 supporting the findings of the previous epidemiological investigation. The main advantage of
52 WGS over other typing techniques such as monoclonal antibody typing,⁸ amplified fragment
53 length polymorphism,¹⁹ pulsed-field gel electrophoresis,³ and sequence-based typing⁹ is that
54
55
56
57
58
59
60

1
2
3 it interrogates the whole genome thus giving maximum resolution, even within individual
4 sequence types. Current barriers to routine implementation of WGS include the inability to
5 sequence directly from clinical specimens, the lack of availability of comprehensive open
6 access genomic databases to compare isolates to, the lack of automated data interpretation
7 software to deliver clinically relevant information, and the need for cost-benefit analyses of
8 WGS versus the current typing methods.
9

10
11
12 We acknowledge several limitations to our study. The study was performed
13 retrospectively and was hampered by the small number of stored *L. pneumophila* isolates
14 available for WGS. In the original investigation we examined multiple isolates from each
15 environmental sample to confirm their phenotype (species, serogroup and monoclonal
16 antibody subgroup). Each sample (and source) contained a single phenotype – hence only a
17 single colony for each sample was characterised genotypically and archived for later use. For
18 the clinical samples five colonies were taken from each positive patient sample and
19 characterised phenotypically. Again only a single phenotype was identified in each patient
20 and hence only a single colony from each was characterised genotypically. This issue
21 remains a challenge for contemporaneous outbreak investigations for two reasons. Firstly,
22 the diagnosis of LD is usually made by detection of *L. pneumophila* urinary antigen, and is
23 often not confirmed by culture of the organism from clinical specimens, which takes two to
24 three days. Secondly, environmental samples can take even longer to culture than clinical
25 specimens, and are usually not processed in the same laboratory. Thus the number of
26 clinical and environmental samples available for typing from *Legionella* outbreaks is likely to
27 be limited.
28

29
30
31 Our analysis was also constrained by the limited available information on the genetic
32 variation and population structure of *L. pneumophila* at the whole genome level.
33 Environmental and clinical isolates are not evenly distributed in the environment based on
34 sequence-based typing observations, suggesting that clinical isolates are a distinct sub-
35 population of environmental strains. Humans are continuously exposed to environmental
36 *Legionellae* and it is not clear why certain sequence types predominate in human disease²⁰.
37 One hypothesis is that disease only occurs in those who have increased susceptibility to
38 infection, for example the elderly, and the immunosuppressed.²¹ Whenever a *Legionella*
39 outbreak occurs it usually reflects the breakdown of *Legionella* control measures, with
40 human infections occurring as a consequence.
41

42
43
44 The genetic diversity of *Legionella* strains within an environmental source, as seen in
45 this analysis, could potentially undermine our ability to link environmental and clinical
46
47
48
49
50
51
52
53
54

1
2
3 isolates in an outbreak situation. Thus a detailed epidemiological investigation accompanied
4 by thorough environmental sampling, sequencing and comparison with patient isolates will
5 continue to be required to confirm the likely source of an outbreak.
6
7

8 Despite these caveats our work here demonstrates that this WGS approach can
9 provide highly discriminatory information within a clinically relevant time frame, but
10 requires a parallel epidemiological investigation to rule in or rule out potential
11 environmental sources. This heralds the opportunity of conducting combined
12 epidemiological and genomic outbreak investigations in real time, as has been performed
13 for other pathogens.¹⁸
14
15
16
17

18 **Acknowledgements**

19 We would like to acknowledge the authors of the original outbreak investigation and the
20 staff of the Respiratory and Systemic Infection Laboratory, Health Protection Agency.
21
22
23

24 **Study approval**

25 Individual patient consent was not obtained as the study was conducted using stored,
26 anonymized bacterial isolates which had collected at the time of the original outbreak
27 investigation in 2003. Ethical approval was not required as this was a retrospective
28 laboratory-based study using stored anonymized bacterial isolates obtained from a
29 diagnostic archive at the Respiratory and Systemic Infection Laboratory, Health Protection
30 Agency. The study was approved by the Cambridge Health Protection Agency Research and
31 Development Committee and the Cambridge University Hospitals NHS Foundation Trust
32 Research and Development Department.
33
34
35
36
37
38
39
40

41 **Funding**

42 This work was supported by grants from the United Kingdom Clinical Research Collaboration
43 (UKCRC) Translational Infection Research Initiative (TIRI); the Medical Research Council
44 (G1000803), with contributions from the Biotechnology and Biological Sciences Research
45 Council, the National Institute for Health Research (NIHR) on behalf of the United Kingdom
46 Department of Health, and the Chief Scientist of the Scottish Government Health
47 Directorate; the Health Protection Agency Strategic Development Research Fund (grant
48 107514); the NIHR Cambridge Biomedical Research Centre; and the Wellcome Trust (grant
49 number 098051).
50
51
52
53
54
55
56
57
58
59
60

Competing interests

The following authors have potential conflicts of interest to declare: GPS (employee and shareholder of Illumina Inc.; JP (travel, accommodation and meeting expenses from Pacific Biosciences and Illumina Ltd); and SJP (consultancy fees from Pfizer).

Data sharing policy

The *L. pneumophila* sequences included in this study have been deposited in the European Nucleotide Archive, under study number ERP001732.

Contributorship

MET, SJP and TGH conceived and designed the study.

CUK and MJE conducted the laboratory experiments.

SR, SDB, JP, SJP and GS analysed and interpreted the data.

SR, TGH, MET wrote the first draft of the manuscript and all authors revised it critically for intellectual content.

All authors reviewed and approved the final manuscript.

REFERENCES

1. Carratala J, Garcia-Vidal C. An update on Legionella. *Current opinion in infectious diseases* 2010;23(2):152-7.
2. Tram C, Simonet M, Nicolas MH, et al. Molecular typing of nosocomial isolates of Legionella pneumophila serogroup 3. *J Clin Microbiol* 1990;28(2):242-5.
3. Schoonmaker D, Heimberger T, Birkhead G. Comparison of ribotyping and restriction enzyme analysis using pulsed-field gel electrophoresis for distinguishing Legionella pneumophila isolates obtained during a nosocomial outbreak. *J Clin Microbiol* 1992;30(6):1491-8.
4. Darelid J, Hallander H, Lofgren S, et al. Community spread of Legionella pneumophila serogroup 1 in temporal relation to a nosocomial outbreak. *Scand J Infect Dis* 2001;33(3):194-9.
5. Birtles RJ, Harrison TG, Samuel D, et al. Evaluation of urinary antigen ELISA for diagnosing Legionella pneumophila serogroup 1 infection. *J Clin Pathol* 1990;43(8):685-90.
6. Helbig JH, Bernander S, Castellani Pastoris M, et al. Pan-European study on culture-proven Legionnaires' disease: distribution of Legionella pneumophila serogroups and monoclonal subgroups. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology* 2002;21(10):710-6.
7. Fry NK, Alexiou-Daniel S, Bangsberg JM, et al. A multicenter evaluation of genotypic methods for the epidemiologic typing of Legionella pneumophila serogroup 1: results of a pan-European study. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 1999;5(8):462-77.

- 1
 - 2
 - 3
 - 4
 - 5
 - 6
 - 7
 - 8
 - 9
 - 10
 - 11
 - 12
 - 13
 - 14
 - 15
 - 16
 - 17
 - 18
 - 19
 - 20
 - 21
 - 22
 - 23
 - 24
 - 25
 - 26
 - 27
 - 28
 - 29
 - 30
 - 31
 - 32
 - 33
 - 34
 - 35
 - 36
 - 37
 - 38
 - 39
 - 40
 - 41
 - 42
 - 43
 - 44
 - 45
 - 46
 - 47
 - 48
 - 49
 - 50
 - 51
 - 52
 - 53
 - 54
 - 55
 - 56
 - 57
 - 58
 - 59
 - 60
8. Brindle RJ, Stannett PJ, Tobin JO. Legionella pneumophila: monoclonal antibody typing of clinical and environmental isolates. *Epidemiol Infect* 1987;99(2):235-9.
9. Gaia V, Fry NK, Afshar B, et al. Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of Legionella pneumophila. *J Clin Microbiol* 2005;43(5):2047-52.
10. Koser CU, Holden MT, Ellington MJ, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* 2012;366(24):2267-75.
11. Gardy JL, Johnston JC, Ho Sui SJ, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 2011;364(8):730-9.
12. Rohde H, Qin J, Cui Y, et al. Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *N Engl J Med* 2011;365(8):718-24.
13. Snitkin ES, Zelazny AM, Thomas PJ, et al. Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella pneumoniae with Whole-Genome Sequencing. *Sci Transl Med* 2012;4(148):148ra16.
14. Kirrage D, Reynolds G, Smith GE, et al. Investigation of an outbreak of Legionnaires' disease: Hereford, UK 2003. *Respiratory medicine* 2007;101(8):1639-44.
15. Gaia V, Fry NK, Harrison TG, et al. Sequence-based typing of Legionella pneumophila serogroup 1 offers the potential for true portability in legionellosis outbreak investigation. *J Clin Microbiol* 2003;41(7):2932-9.
16. Chien M, Morozova I, Shi S, et al. The genomic sequence of the accidental pathogen Legionella pneumophila. *Science* 2004;305(5692):1966-8.
17. Harris SR, Feil EJ, Holden MT, et al. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 2010;327(5964):469-74.
18. Coil DA, Vandersmissen L, Ginevra C, et al. Intragenic tandem repeat variation between Legionella pneumophila strains. *BMC Microbiol* 2008;8:218.
19. Fry NK, Bangsberg JM, Bergmans A, et al. Designation of the European Working Group on Legionella Infection (EWGLI) amplified fragment length polymorphism types of Legionella pneumophila serogroup 1 and results of intercentre proficiency testing Using a standard protocol. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology* 2002;21(10):722-8.
20. Cazalet C, Jarraud S, Ghavi-Helm Y, et al. Multigenome analysis identifies a worldwide distributed epidemic Legionella pneumophila clone that emerged within a highly diverse species. *Genome Res* 2008;18(3):431-41.
21. Ampel NM, Wing EJ. Legionella infection in transplant patients. *Semin Respir Infect* 1990;5(1):30-7.

1
2
3 **Figure 1. Phylogenetic tree of *Legionella pneumophila* strains**
4

5 A. Phylogeny of the species *L. pneumophila*. Clinical, environmental and references isolates
6 are shown in red, blue, and black, respectively. Inset B. Close-up phylogeny of the isolates
7 involved in the outbreak. The branch leading to the reference strain Philadelphia has been
8 truncated for clarity.
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

Title

A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak

Authors

Sandra Reuter¹, Timothy G. Harrison², Claudio U. Köser^{3,4}, Matthew J. Ellington⁴, Geoffrey P. Smith⁵, Julian Parkhill¹, Sharon J. Peacock^{1,3,4,6}, Stephen D. Bentley^{1,3}, M. Estée Török^{3,4,6}

Affiliations

1. The Wellcome Trust Sanger Institute, Hinxton, United Kingdom
2. Respiratory and Systemic Infection Laboratory, Health Protection Agency Centre for Infections, London, United Kingdom
3. Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge, United Kingdom
4. Cambridge Public Health and Microbiology Laboratory, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom
5. Illumina Cambridge Ltd, Saffron Walden, United Kingdom
6. Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom

Corresponding author

Dr M. Estée Török

University of Cambridge, Department of Medicine,
Box 157, Addenbrooke's Hospital, Hills Road,
Cambridge CB2 0QQ, United Kingdom

Tel: +44 (0)1223 217520

Fax: +44 (0)1223 336846

Email: estee.torok@addenbrookes.nhs.uk

Keywords

Legionnaires' disease; *Legionella pneumophila*; outbreak; whole-genome sequencing; typing

Word count

Word count ~~23432374~~; 1 table; 1 figure

ABSTRACT**Introduction**

Epidemiological investigations of Legionnaires' disease outbreaks rely on the rapid identification and typing of clinical and environmental *Legionella* isolates in order to identify and control the source of infection. Rapid bacterial whole-genome sequencing (WGS) is an emerging technology that has the potential to rapidly discriminate outbreak from non-outbreak isolates in a clinically relevant time frame.

Methods

We performed a pilot study to determine the feasibility of using bacterial WGS to differentiate outbreak from non-outbreak isolates collected during an outbreak of Legionnaires' disease. Seven *Legionella* isolates (three clinical and four environmental) were obtained from the reference laboratory and sequenced using the Illumina MiSeq platform at Addenbrooke's Hospital, Cambridge. Bioinformatic analysis was performed blinded to the epidemiological data at the Wellcome Trust Sanger Institute.

Results

We were able to distinguish outbreak from non-outbreak isolates using bacterial WGS, and to confirm the probable environmental source. Our analysis also highlighted constraints, which were the small number of *Legionella pneumophila* isolates available for sequencing, and the limited number of published genomes for comparison.

Conclusions

We have demonstrated the feasibility of using rapid WGS to investigate an outbreak of Legionnaires' disease. Future work includes building larger genomic databases of *Legionella pneumophila* from both clinical and environmental sources, developing automated data interpretation software, and conducting a cost-benefit analysis of WGS versus current typing methods.

ARTICLE SUMMARY**Article focus**

- Epidemiological investigations of Legionnaires' disease outbreaks rely on the rapid identification and typing of clinical and environmental *Legionella pneumophila* isolates in order to identify and control the source of infection
- Rapid bacterial whole genome sequencing (WGS) is an emerging technology that has the ability to identify and discriminate bacterial isolates
- We hypothesised that WGS could be used to discriminate outbreak from non-outbreak *Legionella* isolates in a clinically relevant time frame

Key messages

- We retrospectively applied bacterial WGS to isolates cultured during a previous outbreak investigation, and were able to rapidly distinguish outbreak from non-outbreak isolates, and to identify the probable environmental source
- Our findings were consistent with those of previous epidemiological and microbiological investigations of the same outbreak
- This raises the possibility of conducting combined epidemiological and genomic outbreak investigations in real time

Strengths and limitations of this study

- We have demonstrated the feasibility of using rapid WGS to investigate an outbreak of Legionnaires' disease
- Our study was limited by the small number of *Legionella pneumophila* genomes available for comparison
- Future work includes the development of automated data interpretation software and a cost-benefit analysis of current typing methods compared with WGS

MAIN ARTICLE

Introduction

Legionella pneumophila causes outbreaks of respiratory infection in community settings and results in significant morbidity and mortality.¹ The organism is common in aquatic environments and is spread by aerosol from a contaminated source, often cooling towers and other aerosol-producing devices. Nosocomial outbreaks that are related to contaminated water supplies have also been widely reported.²⁻⁴ The diagnosis of Legionnaires' disease (LD) is based on a compatible clinical syndrome and detection of *L. pneumophila* urinary antigen⁵ or isolation of the organism from respiratory specimens, which requires culture on selective media.⁶ Most cases of human infection are caused by *L. pneumophila* serogroup 1. During *Legionella* outbreaks, clinical and environmental isolates are collected and sent to the reference laboratory for typing.⁷ Epidemiological investigations are dependent on the rapid identification and typing of the associated organisms in order to identify and control the source of infection. Current typing methods include phenotypic (monoclonal antibody subgrouping⁸) and genotypic (sequence-based typing⁹) methods, which typically take one to two days. High-throughput sequencing technology has the potential to rapidly provide information on organism identity and genetic relatedness, and has been shown to provide a high degree of discrimination for a range of other bacteria such as methicillin-resistant *Staphylococcus aureus*,¹⁰ *Mycobacterium tuberculosis*,¹¹ *Escherichia coli* O104:H4¹² and *Klebsiella pneumoniae*.¹³ We hypothesised that WGS could be used to discriminate outbreak from non-outbreak isolates of *L. pneumophila* in a comparable time frame, and with a higher level of discrimination, when compared with current typing methods. ~~Therefore we~~~~We therefore~~ conducted a pilot study to determine the feasibility of using a rapid bench-top sequencing platform (Illumina MiSeq) to retrospectively investigate a *Legionella* outbreak.

Objectives

The aim of this pilot study was to determine the feasibility of using bacterial WGS for the investigation of a previous *Legionella* outbreak.

Epidemiological and microbiological investigation

In 2003, an outbreak of LD occurred in Hereford, United Kingdom.¹⁴⁵ The outbreak started with two community cases that presented with clinical features of infection within a few days of each other, one of whom died. Active case finding identified two further cases in the

Field Code Changed

Field Code Changed

Field Code Changed

1
2
3
4
5
6 local hospital and a formal outbreak investigation was ~~carried out~~^{conducted}. Twenty-four
7 further cases of LD were identified over the next three weeks. All cases had a positive
8 *L. pneumophila* urinary antigen test, and three patients' samples were culture-positive for
9 *L. pneumophila* serogroup 1. Epidemiological and environmental investigations were
10 undertaken to determine possible sources. One hundred and forty-two environmental
11 samples were collected from potential sources, which included 50 cooling towers on 11
12 premises. *L. pneumophila* serogroup 1 was isolated from samples collected at three cooling
13 towers at two different locations (sites A and B) and a domestic spa pool. Clinical and
14 environmental isolates were referred to the Respiratory and Systemic Infection Laboratory,
15 Health Protection Agency, London, for *L. pneumophila* monoclonal antibody (mAb)
16 subgrouping followed by a three allele DNA-sequence based typing (SBT₃) method then in
17 use. The SBT₃ profiles for two of the clinical isolates and isolates from two of the cooling
18 towers were indistinguishable, suggesting that the cooling towers were the likely
19 environmental source. The strains were subsequently re-examined using the current seven
20 allele sequence based typing (SBT) method,¹⁵⁴⁴ with the same outcome.

Field Code Changed

Field Code Changed

29 DNA extraction and whole genome sequencing

30 Seven *L. pneumophila* isolates (three clinical and four environmental) were obtained from
31 the reference laboratory where they had been stored at -80°C with minimal passage since
32 the outbreak. DNA was extracted from each *L. pneumophila* isolate (50ng) and prepared for
33 sequencing using the Nextera DNA Sample Prep Kit (Epicentre). Samples were pooled
34 together and then run on a rapid whole-genome sequencing platform (Illumina MiSeq) at
35 Addenbrooke's Hospital, Cambridge, generating 150bp paired-end reads.

40 Bioinformatic analysis

41 Bioinformatic analysis was performed at the Wellcome Trust Sanger Institute and blinded to
42 the epidemiological data. The sequencing data from the seven samples were mapped to a
43 reference genome, *L. pneumophila* type strain Philadelphia-1,¹⁶⁴⁵ and compared with eight
44 other publicly available *L. pneumophila* genomes (Table 1). Sequence reads were mapped
45 onto the reference genome using SMALT. Regions containing phage or insertion sequence
46 elements were excluded ~~from the analysis~~. Single nucleotide polymorphisms (SNPs) were
47 identified using a standard approach,¹⁷⁴⁶ by removing SNPs with low quality scores and by
48 filtering for SNPs that were present in at least 75% of the mapped reads. The minimum
49 number of high quality reads mapping to call a base was set to four, which is equivalent to a

Formatted: Plain Text, Left

Field Code Changed

Field Code Changed

~~minimum coverage of four. Actual coverage ranged between 20x and 100x per isolate. A maximum likelihood phylogeny was estimated using RAxML. The general time-reversible model with gamma correction was used for among-site variation. Tandem repeats were not considered in the original analysis, although we did re-run the analysis excluding the 23 repetitive genes mentioned in the paper by Coil and colleagues;¹⁸ the overall topology of the phylogenetic tree remained unchanged and would not have affected interpretation of our data. A maximum likelihood phylogeny was estimated using RAxML. The general time-reversible model with gamma correction was used for among-site variation.~~

RESULTS

Phenotypic and typing results

The microbiological characteristics of the *L. pneumophila* isolates included in this study are summarised in Table 1.

Table 1: Clinical, environmental and reference *L. pneumophila* strains

Sample Number	Accession Number	Biological Origin	Type of sample	Serogroup	Monoclonal antibody subgroup	Sequence type*
Reference genome						
Philadelphia	AE017354.1	United States 1974	Clinical	1	Philadelphia	ST36
Published genomes						
ATCC 43290	CP003192.1	United States	Clinical	12	NA	ST187
Alcoy	CP001828.1	Spain	Clinical	1	ND	ST578
Corby	CP000675.2	United Kingdom	Clinical	1	Knoxville	ST51
Lens	CR628337.1	France	Clinical	1	Benidorm	ST15
130b	FR687201.1	United States	Clinical	1	Benidorm	ST42
Paris	CR628336.1	France	Clinical	1	Philadelphia	ST1
Lorraine	FQ958210.1	France	Clinical	1	ND	ST47
LP_HL06041035	FQ958211.1	France	Environmental	1	ND	ST734
Outbreak investigation isolates						
LP_033	ERS166051	Patient 1	Clinical	1	Philadelphia	ST37
LP_035	ERS166045	Patient 2	Clinical	1	Philadelphia	ST37
LP_617	ERS166047	Patient 3	Clinical	1	Allentown / France	ST47
LP_056	ERS166052	Site A cooling tower 1 (CT1)	Environmental	1	Philadelphia	ST37
LP_427	ERS166050	Site A cooling tower 2 (CT2)	Environmental	1	Philadelphia	ST37
LP_467	ERS166049	Domestic spa pool	Environmental	1	Philadelphia	ST37
LP_423	ERS166048	Site B cooling	Environmental	1	Oxford / OLDA	ST1

		tower 1 (CT1)				
--	--	------------------	--	--	--	--

*Sequence type was derived from the genome sequence data and was concordant with the results of the seven allele sequenced based typing method.

NA = Not applicable

ND = not determined

Genomic analysis

Whole genome phylogenetic analysis showed that two clinical isolates (LP033 and LP035) and three environmental isolates (LP056, LP427 and LP467) were closely related genetically, and accordingly clustered together on the tree (Figure 1A). These five isolates were therefore considered to be the outbreak isolates, though it was not possible to obtain directional information from this analysis due to the low number of SNPs differentiating isolates; in total, there were less than 15 SNP differences within the outbreak strain cluster (Figure 1B). Furthermore, the genetic variability between isolates from two cooling tower isolates on Site A, and the observation that these intermingled with the clinical isolates on the tree, suggested that some diversity existed in the source population before the onset of the outbreak. Sequence types were derived from the genome sequence data and confirmed that all five isolates were ST37.

The two remaining isolates (LP423 and LP617) were situated ~75,000 to 77,500 SNPs respectively from the outbreak cluster, and thus were not considered to be part of the outbreak. Sequence types were derived from the genomic data and the clinical isolate (LP617) was ST47 whereas the environmental isolate (LP423) was ST1.

The five outbreak isolates were compared to the nine published strains and found to be most closely related to the Philadelphia-1 strain (which is ST36, a single locus variant of ST37) and to the ATCC 43290 strain (which is ST187) (Figure 1A). Both of these isolates were ~10,000 to 13,000 SNPs distant from the outbreak cluster. The LP617 isolate was 56 SNPs different from Lorraine strain (also ST47), and the LP423 isolate was 906 SNPs different from the Paris strain (also ST1).

Comparison of epidemiological investigation and genomic analysis

Two clinical isolates (LP033 and LP035) had been obtained from patients included in the outbreak. Both strains were located within the outbreak cluster in the phylogenetic tree. The third clinical isolate (LP617) was obtained from a patient who had initially been linked to the outbreak. The original epidemiological investigation found, however, that this patient

1
2
3
4
5
6 was a lorry driver, who had passed through Hereford at the time of the outbreak, and had
7 likely acquired his infection elsewhere. This isolate was located distant to the outbreak
8 cluster on the phylogenetic tree, and was therefore not considered to be linked to the
9 outbreak. Thus, for the clinical isolates, the genomic data supported the results of the
10 previous epidemiological investigation.
11

12
13 Three environmental isolates were located within the outbreak cluster. Two of these
14 (LP056 and LP427) had been collected from two cooling towers at the same location (Site A)
15 whilst the third environmental isolate (LP467) had been collected from a spa pool in local
16 domestic premises. Given the small number of SNP differences between these three isolates
17 (Figure 1B) it was not possible to determine which of these isolates represented the source
18 of the outbreak using genomic data alone. The original epidemiological investigation had,
19 however, concluded that the cooling towers on Site A were the most likely source.
20
21

22
23 The fourth environmental isolate (LP423) was obtained from a cooling tower at a
24 different site (Site B), which was considered epidemiologically unlikely to be the source of
25 the outbreak; a view supported by the typing data. This isolate was located away from the
26 outbreak cluster and was most closely related (906 SNPs different) to the Paris strain (Figure
27 1A).
28
29
30
31

32 **Comparison of conventional typing and genomic analysis**

33 We also compared the results of the conventional typing (monoclonal antibody typing and
34 sequence based typing) with WGS. All of the isolates included in this analysis were
35 *L. pneumophila* serogroup 1, apart from the ATCC 43290 strain, which was serogroup 12. All
36 of the outbreak strains belonged to the mAb subgroup 'Philadelphia', and were ST37. The
37 clinical non-outbreak isolate belonged to the mAb subgroup 'Allentown/France' and was
38 ST47, whereas the environmental non-outbreak isolate belonged to the mAb subgroup
39 'Oxford/OLDA' and was ST1. Thus, in this outbreak, the performance of WGS sequence was
40 equivalent to conventional SBT in differentiating the outbreak from the non-outbreak
41 strains. WGS was unable to distinguish the epidemiologically most likely source of the
42 outbreak (Site A cooling towers) from the domestic spa pool.
43
44
45
46
47
48

49 **DISCUSSION**

50 Here, we have demonstrated the feasibility of using WGS to perform an investigation of a
51 *Legionella* outbreak. Using genomic analysis we were readily able to distinguish outbreak
52 from non-outbreak *Legionella* isolates, and to identify probable environmental sources, thus
53
54
55
56
57
58
59
60

1
2
3
4
5
6 supporting the findings of the previous epidemiological investigation. The main advantage of
7 WGS over other typing techniques such as monoclonal antibody typing,⁸ amplified fragment
8 length polymorphism,¹⁹⁺⁷ pulsed-field gel electrophoresis,³ and sequence-based typing⁹ is
9 that it interrogates the whole genome thus giving maximum resolution, even within
10 individual sequence types. Current barriers to routine implementation of WGS include the
11 inability to sequence directly from clinical specimens, the lack of availability of
12 comprehensive open access genomic databases to compare isolates to, the lack of
13 automated data interpretation software to deliver clinically relevant information, and the
14 need for cost-benefit analyses of WGS versus the current typing methods.
15
16
17
18

19 We acknowledge several limitations to our study. The study was performed
20 retrospectively and was hampered by the small number of stored *L. pneumophila* isolates
21 available for WGS. In the original investigation we examined multiple isolates from each
22 environmental sample to confirm their phenotype (species, serogroup and monoclonal
23 antibody subgroup). Each sample (and source) contained a single phenotype – hence only a
24 single colony for each sample was characterised genotypically and archived for later use. For
25 the clinical samples five colonies were taken from each positive patient sample and
26 characterised phenotypically. Again only a single phenotype was identified in each patient
27 and hence only a single colony from each was characterised genotypically. This issue
28 remains~~This is also~~ a challenge for contemporaneous outbreak investigations for two
29 reasons. Firstly, the diagnosis of LD is usually made by detection of *L. pneumophila* urinary
30 antigen, and is often not confirmed by culture of the organism from clinical specimens,
31 which takes two to three days. Secondly, environmental samples can take even longer to
32 culture than clinical specimens, and are usually not processed in the same laboratory. Thus
33 the number of clinical and environmental samples available for typing from *Legionella*
34 outbreaks is likely to be limited.
35
36
37
38
39
40
41

42 Our analysis was also constrained by the limited available information on the genetic
43 variation and population structure of *L. pneumophila* at the whole genome level.
44 Environmental and clinical isolates are not evenly distributed in the environment based on
45 sequence-based typing observations, suggesting that clinical isolates are a distinct sub-
46 population of environmental strains. Humans are continuously exposed to environmental
47 *Legionellae* and it is not clear why certain sequence types predominate in human disease²⁰.
48 One hypothesis is that disease only occurs in those who have increased susceptibility to
49 infection, for example the elderly, and the immunosuppressed.²¹⁺²⁴ Whenever a *Legionella*
50
51
52
53
54
55
56
57
58
59
60

Field Code Changed

Field Code Changed

1
2
3
4
5
6 outbreak occurs it usually reflects the breakdown of *Legionella* control measures, with
7 human infections occurring as a consequence.
8

9 The genetic diversity of *Legionella* strains within an environmental source, as seen in
10 this analysis, could potentially undermine our ability to link environmental and clinical
11 isolates in an outbreak situation. Thus a detailed epidemiological investigation accompanied
12 by thorough environmental sampling, sequencing and comparison with patient isolates will
13 continue to be required to confirm the likely source of an outbreak.
14
15

16 Despite these caveats our work here demonstrates that this WGS approach can
17 provide highly discriminatory information within a clinically relevant time frame, but
18 requires a parallel epidemiological investigation to rule in or rule out potential
19 environmental sources. This heralds the opportunity of conducting combined
20 epidemiological and genomic outbreak investigations in real time, as has been performed
21 for other pathogens.¹⁸
22
23
24
25

26 **Acknowledgements**

27 We would like to acknowledge the authors of the original outbreak investigation and the
28 staff of the Respiratory and Systemic Infection Laboratory, Health Protection Agency.
29
30
31

32 **Study approval**

33 Individual patient consent was not obtained as the study was conducted using stored,
34 anonymized bacterial isolates which had collected at the time of the original outbreak
35 investigation in 2003. Ethical approval was not required as this was a retrospective
36 laboratory-based study using stored anonymized bacterial isolates obtained from a
37 diagnostic archive at the Respiratory and Systemic Infection Laboratory, Health Protection
38 Agency. The study was approved by the Cambridge Health Protection Agency Research and
39 Development Committee and the Cambridge University Hospitals NHS Foundation Trust
40 Research and Development Department.
41
42
43
44
45

46 **Funding**

47 This work was supported by grants from the United Kingdom Clinical Research Collaboration
48 (UKCRC) Translational Infection Research Initiative (TIRI); the Medical Research Council
49 (G1000803), with contributions from the Biotechnology and Biological Sciences Research
50 Council, the National Institute for Health Research (NIHR) on behalf of the United Kingdom
51 Department of Health, and the Chief Scientist of the Scottish Government Health
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6 Directorate; the Health Protection Agency Strategic Development Research Fund (grant
7 107514); the NIHR Cambridge Biomedical Research Centre; and the Wellcome Trust (grant
8 number 098051).
9

10 11 12 **Competing interests**

13 The following authors have potential conflicts of interest to declare: GPS (employee and
14 shareholder of Illumina Inc.; JP (travel, accommodation and meeting expenses from Pacific
15 Biosciences and Illumina Ltd); and SJP (consultancy fees from Pfizer).
16
17
18

19 **Data sharing policy**

20 The *L. pneumophila* sequences included in this study have been deposited in the European
21 Nucleotide Archive, under study number ERP001732.
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

REFERENCES

1. Carratala J, Garcia-Vidal C. An update on Legionella. *Current opinion in infectious diseases* 2010;23(2):152-7.
2. Tram C, Simonet M, Nicolas MH, Offredo C, Grimont F, Lefevre M, et al. Molecular typing of nosocomial isolates of Legionella pneumophila serogroup 3. *J Clin Microbiol* 1990;28(2):242-5.
3. Schoonmaker D, Heimberger T, Birkhead G. Comparison of ribotyping and restriction enzyme analysis using pulsed-field gel electrophoresis for distinguishing Legionella pneumophila isolates obtained during a nosocomial outbreak. *J Clin Microbiol* 1992;30(6):1491-8.
4. Darelid J, Hallander H, Lofgren S, Malmvall BE, Olinder-Nielsen AM. Community spread of Legionella pneumophila serogroup 1 in temporal relation to a nosocomial outbreak. *Scand J Infect Dis* 2001;33(3):194-9.
5. Birtles RJ, Harrison TG, Samuel D, Taylor AG. Evaluation of urinary antigen ELISA for diagnosing Legionella pneumophila serogroup 1 infection. *J Clin Pathol* 1990;43(8):685-90.
6. Helbig JH, Bernander S, Castellani Pastoris M, Etienne J, Gaia V, Lauwers S, et al. Pan-European study on culture-proven Legionnaires' disease: distribution of Legionella pneumophila serogroups and monoclonal subgroups. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology* 2002;21(10):710-6.
7. Fry NK, Alexiou-Daniel S, Bangsberg JM, Bernander S, Castellani Pastoris M, Etienne J, et al. A multicenter evaluation of genotypic methods for the epidemiologic typing of Legionella pneumophila serogroup 1: results of a pan-European study. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 1999;5(8):462-77.
8. Brindle RJ, Stannett PJ, Tobin JO. Legionella pneumophila: monoclonal antibody typing of clinical and environmental isolates. *Epidemiol Infect* 1987;99(2):235-9.
9. Gaia V, Fry NK, Afshar B, Luck PC, Meugnier H, Etienne J, et al. Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of Legionella pneumophila. *J Clin Microbiol* 2005;43(5):2047-52.
10. Koser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* 2012;366(24):2267-75.
11. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodtkin E, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 2011;364(8):730-9.
12. Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, et al. Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *N Engl J Med* 2011;365(8):718-24.
13. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, et al. Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella pneumoniae with Whole-Genome Sequencing. *Sci Transl Med* 2012;4(148):148ra16.
14. Kirrage D, Reynolds G, Smith GE, Olowokure B. Investigation of an outbreak of Legionnaires' disease: Hereford, UK 2003. *Respiratory medicine* 2007;101(8):1639-44.
15. Gaia V, Fry NK, Harrison TG, Peduzzi R. Sequence-based typing of Legionella pneumophila serogroup 1 offers the potential for true portability in legionellosis outbreak investigation. *J Clin Microbiol* 2003;41(7):2932-9.
16. Chien M, Morozova I, Shi S, Sheng H, Chen J, Gomez SM, et al. The genomic sequence of the accidental pathogen Legionella pneumophila. *Science* 2004;305(5692):1966-8.

17. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 2010;327(5964):469-74.
18. Coil DA, Vandersmissen L, Ginevra C, Jarraud S, Lammertyn E, Anne J. Intragenic tandem repeat variation between *Legionella pneumophila* strains. *BMC Microbiol* 2008;8:218.
19. Fry NK, Bangsberg JM, Bergmans A, Bernander S, Etienne J, Franzin L, et al. Designation of the European Working Group on Legionella Infection (EWGLI) amplified fragment length polymorphism types of *Legionella pneumophila* serogroup 1 and results of intercentre proficiency testing Using a standard protocol. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology* 2002;21(10):722-8.
20. Cazalet C, Jarraud S, Ghavi-Helm Y, Kunst F, Glaser P, Etienne J, et al. Multigenome analysis identifies a worldwide distributed epidemic *Legionella pneumophila* clone that emerged within a highly diverse species. *Genome Res* 2008;18(3):431-41.
21. Ampel NM, Wing EJ. Legionella infection in transplant patients. *Semin Respir Infect* 1990;5(1):30-7.

1
2
3
4
5
6 **Figure 1. Phylogenetic tree of *Legionella pneumophila* strains**

7
8 A. Phylogeny of the species *L. pneumophila*. Clinical, environmental and references isolates
9 are shown in red, blue, and black, respectively. Inset B. Close-up phylogeny of the isolates
10 involved in the outbreak. The branch leading to the reference strain Philadelphia has been
11 truncated for clarity.
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

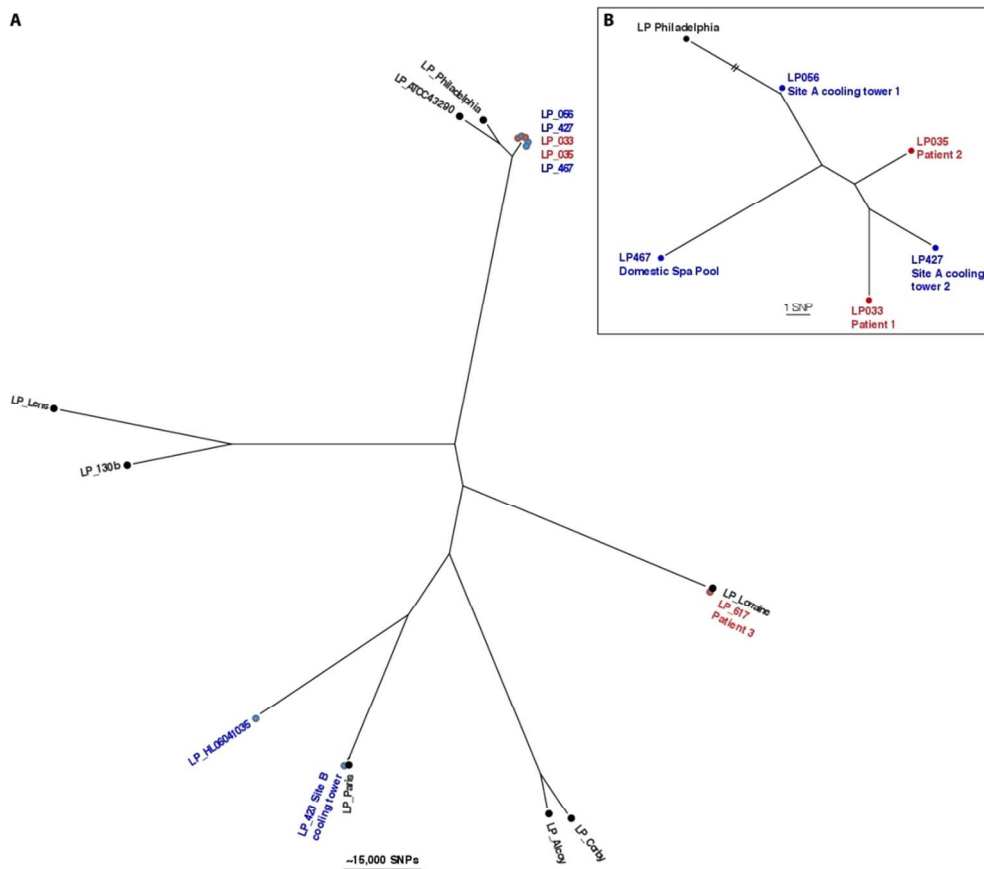


Figure 1. Phylogenetic tree of *Legionella pneumophila* strains
 A. Phylogeny of the species *L. pneumophila*. Clinical, environmental and references isolates are shown in red, blue, and black, respectively. Inset B. Close-up phylogeny of the isolates involved in the outbreak. The branch leading to the reference strain Philadelphia has been truncated for clarity.

99x90mm (300 x 300 DPI)

