BMJ open
accessible medical research

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

Sandra Reuter,[1] Timothy G Harrison,[2] Claudio U Köser,[3,4] Matthew J Ellington,[4] Geoffrey P Smith,[5] Julian Parkhill,[1] Sharon J Peacock,[1,3,4,6] Stephen D Bentley,[1,3] M Estée Török[3,4,6]

To cite: Reuter S, Harrison TG, Köser CU, *et al*. A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak. *BMJ Open* 2013;**3**:e002175. doi:10.1136/bmjopen-2012-002175

▸ Prepublication history and additional material for this paper are available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2012-002175).

Received 14 October 2012
Revised 27 November 2012
Accepted 11 December 2012

For numbered affiliations see end of article

Correspondence to
Dr M Estée Török; estee.torok@addenbrookes.nhs.uk

ABSTRACT

Objectives: 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 *L 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 *L 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.

BACKGROUND

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 antigen[5] 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 subgrouping[8]) and genotypic (sequence-based typing[9]) methods, which typically take 1–2 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* 0104:H4[12] 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 benchtop sequencing platform (Illumina MiSeq) to retrospectively investigate a *Legionella* outbreak.

DESIGN

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, UK.[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 local hospital and a formal outbreak investigation was carried out. Twenty-four further cases of LD were identified over the next three weeks. All cases had a positive *L pneumophila* urinary antigen test, and three patients' samples were culture-positive for *L pneumophila* serogroup 1. Epidemiological and environmental investigations were undertaken to determine possible sources. A total of 142 environmental samples were collected from potential sources, which included 50 cooling towers on 11 premises. *L pneumophila* serogroup 1 was isolated from samples collected at three cooling towers at two different locations (sites A and B) and a domestic spa pool Clinical and environmental isolates were referred to the Respiratory and Systemic Infection Laboratory, Health Protection Agency, London, for *L pneumophila* monoclonal antibody (mAb) subgrouping followed by a three-allele DNA-sequence-based typing (SBT_3) method then in use. The SBT_3 profiles for two of the clinical isolates and isolates from two of the cooling towers were indistinguishable, suggesting that the cooling towers were the likely environmental source. The strains were subsequently re-examined using the current seven-allele SBT method,[15] with the same outcome.

DNA extraction and whole genome sequencing

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

Bioinformatic analysis

Bioinformatic analysis was performed at the Wellcome Trust Sanger Institute and blinded to the epidemiological data. The sequencing data from the seven samples were mapped to a reference genome, *L pneumophila*-type strain Philadelphia-1,[16] and compared with eight other publicly available *L pneumophila* genomes (table 1). Sequence reads were mapped onto the reference genome using the SMALT software programme. Regions containing phage or insertion sequence elements were excluded from the analysis. Single nucleotide polymorphisms (SNPs) were identified using a standard approach,[17] by removing SNPs with low-quality scores and by filtering for SNPs that were present in at least 75% of the mapped reads. The minimum number of high-quality reads mapping to call a base was set to four, which is equivalent to a minimum coverage of four. Actual coverage ranged between 20× and 100× per isolate. A maximum likelihood phylogeny was estimated using the RA×ML software programme. The general time-reversible model with γ 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 *et al*,[18] the overall topology of the phylogenetic tree remained unchanged and would not have affected the 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.

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 owing to the low number of SNPs differentiating isolates; in total, there were less than 15 SNP

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						
LP Philadelphia	AE017354.1	USA 1974	Clinical	1	Philadelphia	ST36
Published genomes						
LP ATCC 43290	CP003192.1	USA	Clinical	12	NA	ST187
LP Alcoy	CP001828.1	Spain	Clinical	1	ND	ST578
LP Corby	CP000675.2	UK	Clinical	1	Knoxville	ST51
LP Lens	CR628337.1	France	Clinical	1	Benidorm	ST15
LP 130b	FR687201.1	USA	Clinical	1	Benidorm	ST42
LP Paris	CR628336.1	France	Clinical	1	Philadelphia	ST1
LP Lorraine	FQ958210.1	France	Clinical	1	ND	ST47
LPHL06041035	FQ958211.1	France	Environmental	1	ND	ST734
Outbreak investigation isolates						
LP033	ERS166051	Patient 1	Clinical	1	Philadelphia	ST37
LP035	ERS166045	Patient 2	Clinical	1	Philadelphia	ST37
LP617	ERS166047	Patient 3	Clinical	1	Allentown/France	ST47
LP056	ERS166052	Site A cooling tower 1	Environmental	1	Philadelphia	ST37
LP427	ERS166050	Site A cooling tower 2	Environmental	1	Philadelphia	ST37
LP467	ERS166049	Domestic spa pool	Environmental	1	Philadelphia	ST37
LP423	ERS166048	Site B cooling tower 1	Environmental	1	Oxford/OLDA	ST1

*Sequence type was derived from the genome sequence data and was concordant with the results of the seven-allele sequence-based typing method.
NA, Not applicable; ND, not determined.

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 43 290 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.

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

The fourth environmental isolate (LP423) was obtained from a cooling tower at a different site (site B), which was considered epidemiologically unlikely to be the source of the outbreak; a view supported by the typing data. This isolate was located away from the outbreak cluster and

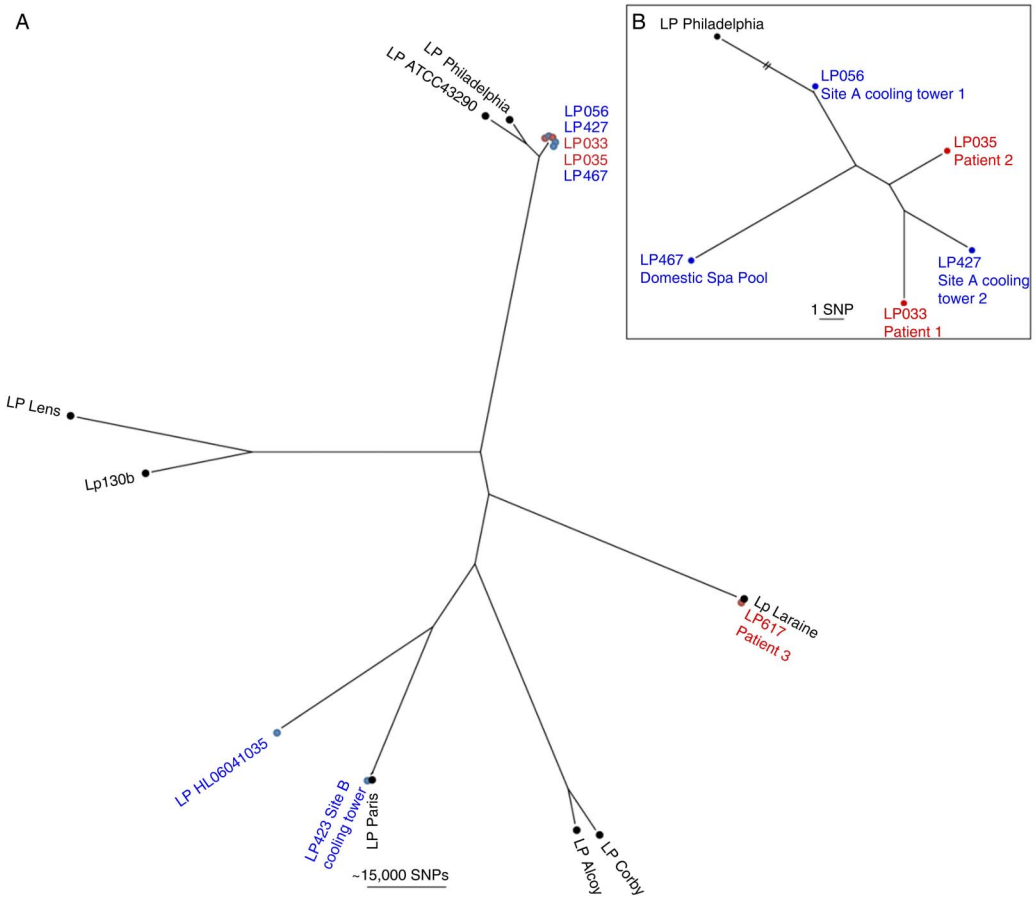


Figure 1 Phylogenetic tree of *Legionella pneumophila* strains. (A) Phylogeny of the species *L pneumophila*. Clinical, environmental and reference isolates are shown in red, blue and black, respectively. Inset (B) close-up phylogeny of the isolates involved in the outbreak. The branch leading to the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

was most closely related (906 SNPs different) to the Paris strain (figure 1A).

Comparison of conventional typing and genomic analysis

We also compared the results of the conventional typing (monoclonal antibody typing and sequence-based typing) with WGS. All of the isolates included in this analysis were *L pneumophila* serogroup 1, apart from the ATCC 43 290 strain, which was serogroup 12. All of the outbreak strains belonged to the mAb subgroup 'Philadelphia', and were ST37. The clinical non-outbreak isolate belonged to the mAb subgroup 'Allentown/France' and was ST47, whereas the environmental non-outbreak isolate belonged to the mAb subgroup 'Oxford/OLDA' and was ST1. Thus, in this outbreak, the performance of WGS sequence was equivalent to conventional SBT in differentiating the outbreak from the non-outbreak strains. WGS was unable to distinguish the epidemiologically most likely source of the outbreak (site A cooling towers) from the domestic spa pool.

DISCUSSION

Here, we have demonstrated the feasibility of using WGS to perform an investigation of a *Legionella* outbreak. Using genomic analysis, we were readily able to distinguish outbreak from non-outbreak *Legionella* isolates, and to identify probable environmental sources, thus supporting the findings of the previous epidemiological investigation. The main advantage of WGS over other typing techniques such as monoclonal antibody typing,[8] amplified fragment length polymorphism,[19] pulsed-field gel electrophoresis,[3] and sequence-based typing[9] is that it interrogates the whole genome, thus giving maximum resolution, even within individual sequence types. Current barriers to routine implementation of WGS include the inability to sequence directly from clinical specimens, the lack of availability of comprehensive open-access genomic databases to compare isolates to, the lack of automated data interpretation software to deliver clinically relevant information and the need for cost-benefit analyses of WGS versus the current typing methods.

We acknowledge several limitations to our study. The study was performed retrospectively and was hampered by the small number of stored *L pneumophila* isolates available for WGS. In the original investigation, we examined multiple isolates from each environmental sample to confirm their phenotype (species, serogroup and monoclonal antibody subgroup). Each sample (and source)

contained a single phenotype—hence only a single colony for each sample was characterised genotypically and archived for later use. For the clinical samples, five colonies were taken from each positive patient sample and characterised phenotypically. Again, only a single phenotype was identified in each patient and hence only a single colony from each was characterised genotypically. This issue remains a challenge for contemporaneous outbreak investigations for two reasons. First, the diagnosis of LD is usually made by the detection of *L pneumophila* urinary antigen, and is often not confirmed by culture of the organism from clinical specimens, which takes 2–3 days. Second, environmental samples can take even longer to culture than clinical specimens, and are usually not processed in the same laboratory. Thus, the number of clinical and environmental samples available for typing from *Legionella* outbreaks is likely to be limited.

Our analysis was also constrained by the limited available information on the genetic variation and population structure of *L pneumophila* at the whole genome level. Environmental and clinical isolates are not evenly distributed in the environment, based on sequence-based typing observations, suggesting that clinical isolates are a distinct subpopulation of environmental strains. Humans are continuously exposed to environmental *Legionellae* and it is not clear why certain sequence types predominate in human disease.[20] One hypothesis is that disease only occurs in those who have increased susceptibility to infection, for example, the elderly, and the immunosuppressed.[21] Whenever a *Legionella* outbreak occurs, it usually reflects the breakdown of *Legionella* control measures, with human infections occurring as a consequence.

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

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

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

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, London.

Contributors 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.

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 UK 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).

Ethics approval Cambridge University Hospitals NHS Foundation Trust Research and Development Department.

Provenance and peer review Not commissioned; externally peer reviewed.

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

REFERENCES

1. Carratala J, Garcia-Vidal C. An update on Legionella. *Curr Opin Infect Dis* 2010;23:152–7.
2. Tram C, Simonet M, Nicolas MH, *et al*. Molecular typing of nosocomial isolates of Legionella pneumophila serogroup 3. *J Clin Microbiol* 1990;28:242–5.
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;30:1491–8.
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;33:194–9.
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;43:685–90.
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. *Eur J Clin Microbiol Infect Dis* 2002;21:710–16.
7. Fry NK, Alexiou-Daniel S, Bangsborg 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;5:462–77.
8. Brindle RJ, Stannett PJ, Tobin JO. Legionella pneumophila: monoclonal antibody typing of clinical and environmental isolates. *Epidemiol Infect* 1987;99:235–9.
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;43:2047–52.
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;366:2267–75.
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;364:730–9.
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;365:718–24.
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;4:148ra16.
14. Kirrage D, Reynolds G, Smith GE, *et al*. Investigation of an outbreak of Legionnaires' disease: Hereford, UK 2003. *Respir Med* 2007;101:1639–44.
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;41:2932–9.
16. Chien M, Morozova I, Shi S, *et al*. The genomic sequence of the accidental pathogen Legionella pneumophila. *Science* 2004;305:1966–8.
17. Harris SR, Feil EJ, Holden MT, *et al*. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 2010;327:469–74.
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, Bangsborg 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;21:722–8.
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;18:431–41.
21. Ampel NM, Wing EJ. Legionella infection in transplant patients. *Semin Respir Infect* 1990;5:30–7.

BMJ open

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

Journal:	*BMJ Open*
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

Title

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

Authors

Sandra Reuter[1], Timothy G. Harrison[2], Claudio U. Köser[3,4], Matthew J. Ellington[4], Geoffrey P. Smith[5], Julian Parkhill[1], 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 antigen[5] 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 subgrouping[8]) and genotypic (sequence-based typing[9]) 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* 0104:H4[12] 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

local hospital and a formal outbreak investigation was conducted. Twenty-four further cases of LD were identified over the next three weeks. All cases had a positive *L. pneumophila* urinary antigen test, and three patients' samples were culture-positive for *L. pneumophila* serogroup 1. Epidemiological and environmental investigations were undertaken to determine possible sources. One hundred and forty-two environmental samples were collected from potential sources, which included 50 cooling towers on 11 premises. *L. pneumophila* serogroup 1 was isolated from samples collected at three cooling towers at two different locations (sites A and B) and a domestic spa pool. Clinical and environmental isolates were referred to the Respiratory and Systemic Infection Laboratory, Health Protection Agency, London, for *L. pneumophila* monoclonal antibody (mAb) subgrouping followed by a three allele DNA-sequence based typing (SBT_3) method then in use. The SBT_3 profiles for two of the clinical isolates and isolates from two of the cooling towers were indistinguishable, suggesting that the cooling towers were the likely environmental source. The strains were subsequently re-examined using the current seven allele sequence based typing (SBT) method,[14] with the same outcome.

DNA extraction and whole genome sequencing

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

Bioinformatic analysis

Bioinformatic analysis was performed at the Wellcome Trust Sanger Institute and blinded to the epidemiological data. The sequencing data from the seven samples were mapped to a reference genome, *L. pneumophila* type strain Philadelphia-1,[15] and compared with eight other publicly available *L. pneumophila* genomes (Table 1). Sequence reads were mapped onto the reference genome using SMALT. Regions containing phage or insertion sequence elements were excluded. Single nucleotide polymorphisms (SNPs) were identified using a standard approach,[16] by removing SNPs with low quality scores and by filtering for SNPs that were present in at least 75% of the mapped reads. 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 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.

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

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

Comparison of conventional typing and genomic analysis

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

DISCUSSION

Here, we have demonstrated the feasibility of using WGS to perform an investigation of a *Legionella* outbreak. Using genomic analysis we were readily able to distinguish outbreak from non-outbreak *Legionella* isolates, and to identify probable environmental sources, thus supporting the findings of the previous epidemiological investigation. The main advantage of WGS over other typing techniques such as monoclonal antibody typing,[8] amplified fragment length polymorphism,[17] pulsed-field gel electrophoresis,[3] and sequence-based typing[9] is that it interrogates the whole genome thus giving maximum resolution, even within individual sequence types. Current barriers to routine implementation of WGS include the inability to

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

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

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

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

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

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

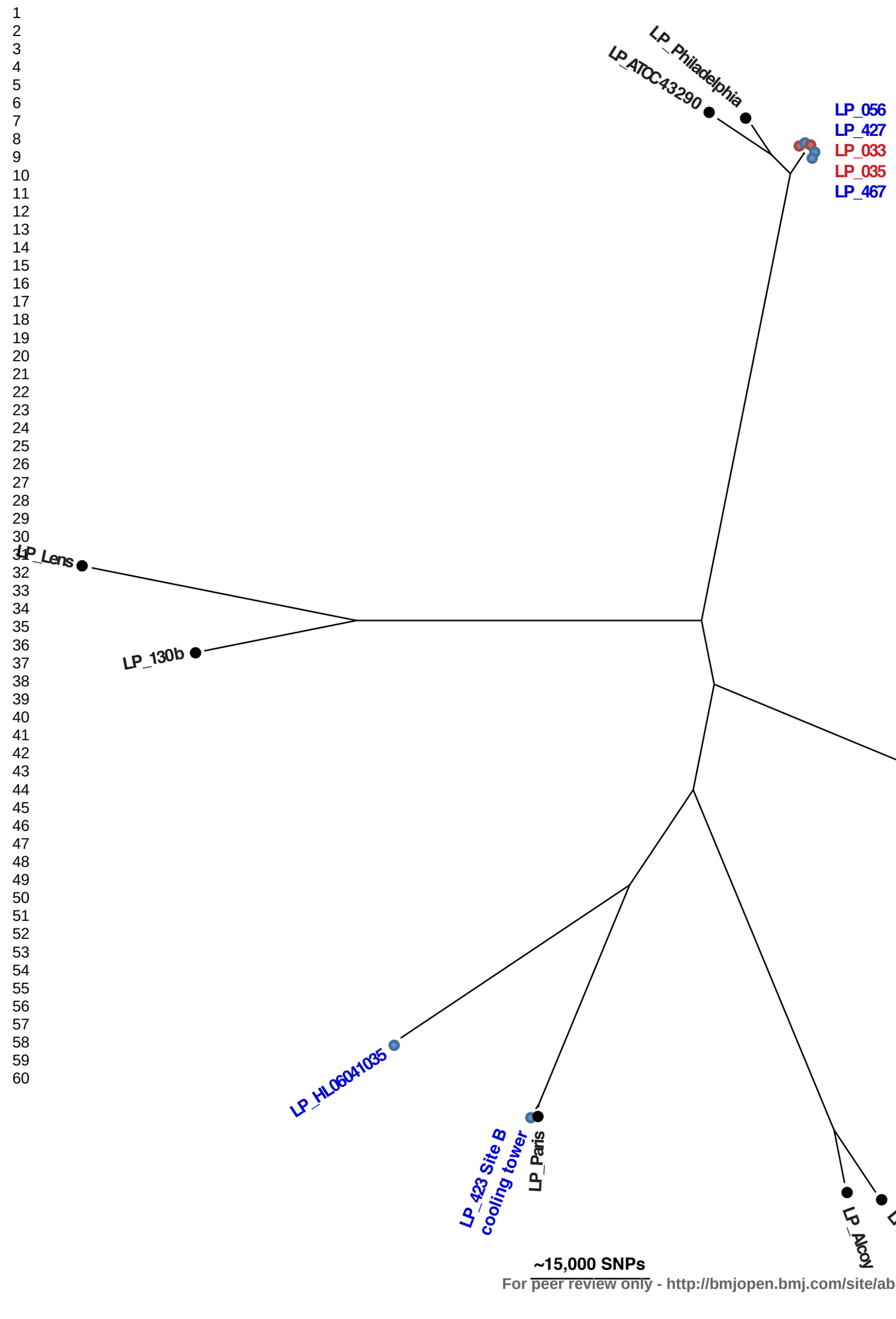
1. Carratala J, Garcia-Vidal C. An update on *Legionella*. *Curr Opin Infect Dis* 2010;23(2):152-7.
2. Tram C, Simonet M, Nicolas MH, et al. Molecular typing of nosocomial isolates of *Legionella pneumophila* serogroup 3. *J Clin Microbiol* 1990;28(2):242-5.
3. Schoonmaker D, Heimberger T, Birkhead G. 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, Bangsborg 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, Bangsborg 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.

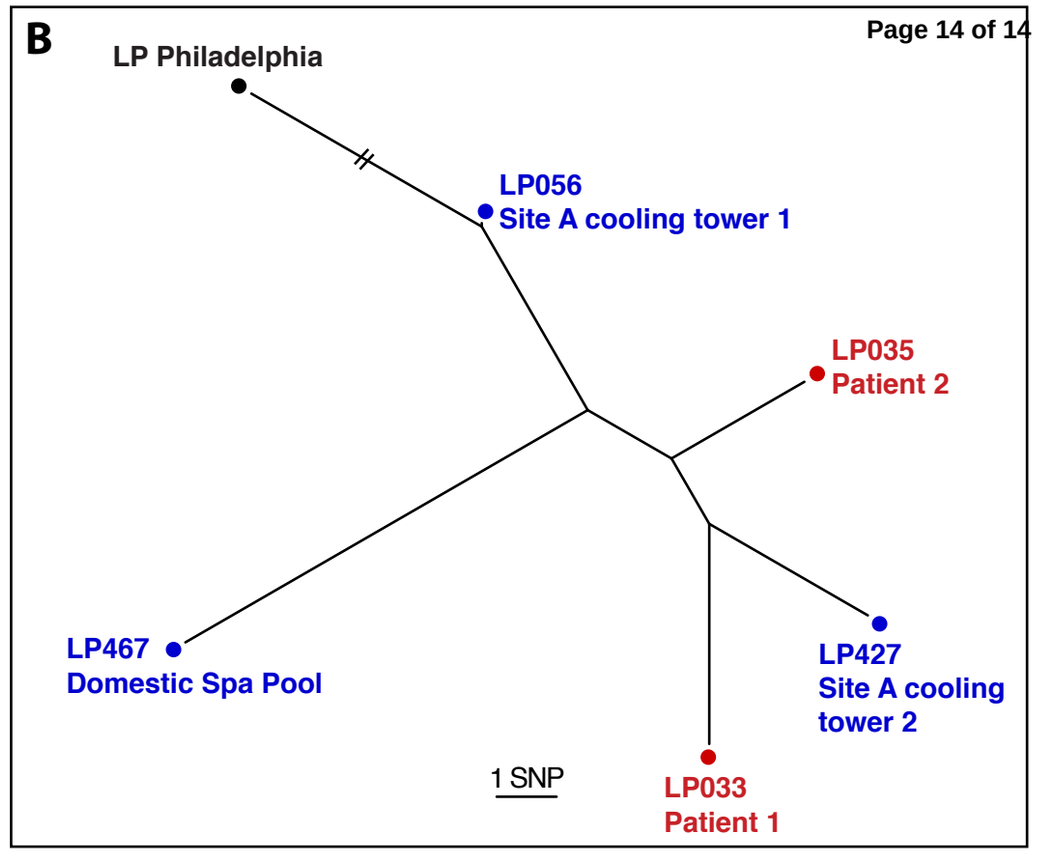
Figure 1. Phylogenetic tree of *Legionella pneumophila* strains

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.

A


LP_Philadelphia
LP_ATCC43290
LP_056
LP_427
LP_033
LP_035
LP_467
LP_Lens
LP_130b
LP_Lorraine
LP_617
Patient 3
LP_HL06041035
LP_423 Site B
cooling tower
LP_Paris
LP_Alcoy
LP_Cotby
~15,000 SNPs

B


LP Philadelphia
LP056
Site A cooling tower 1
LP035
Patient 2
LP467
Domestic Spa Pool
LP427
Site A cooling
tower 2
1 SNP
LP033
Patient 1

BMJ open

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

Journal:	*BMJ Open*
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

Title

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

Authors

Sandra Reuter[1], Timothy G. Harrison[2], Claudio U. Köser[3,4], Matthew J. Ellington[4], Geoffrey P. Smith[5], Julian Parkhill[1], 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 antigen[5] 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 subgrouping[8]) and genotypic (sequence-based typing[9]) 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* 0104:H4[12] 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

local hospital and a formal outbreak investigation was carried out. Twenty-four further cases of LD were identified over the next three weeks. All cases had a positive *L. pneumophila* urinary antigen test, and three patients' samples were culture-positive for *L. pneumophila* serogroup 1. Epidemiological and environmental investigations were undertaken to determine possible sources. One hundred and forty-two environmental samples were collected from potential sources, which included 50 cooling towers on 11 premises. *L. pneumophila* serogroup 1 was isolated from samples collected at three cooling towers at two different locations (sites A and B) and a domestic spa pool. Clinical and environmental isolates were referred to the Respiratory and Systemic Infection Laboratory, Health Protection Agency, London, for *L. pneumophila* monoclonal antibody (mAb) subgrouping followed by a three allele DNA-sequence based typing (SBT_3) method then in use. The SBT_3 profiles for two of the clinical isolates and isolates from two of the cooling towers were indistinguishable, suggesting that the cooling towers were the likely environmental source. The strains were subsequently re-examined using the current seven allele sequence based typing (SBT) method,[15] with the same outcome.

DNA extraction and whole genome sequencing

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

Bioinformatic analysis

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

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

outbreak. Thus, for the clinical isolates, the genomic data supported the results of the previous epidemiological investigation.

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

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

Comparison of conventional typing and genomic analysis

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

DISCUSSION

Here, we have demonstrated the feasibility of using WGS to perform an investigation of a *Legionella* outbreak. Using genomic analysis we were readily able to distinguish outbreak from non-outbreak *Legionella* isolates, and to identify probable environmental sources, thus supporting the findings of the previous epidemiological investigation. The main advantage of WGS over other typing techniques such as monoclonal antibody typing,[8] amplified fragment length polymorphism,[19] pulsed-field gel electrophoresis,[3] and sequence-based typing[9] is that

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

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

Our analysis was also constrained by the limited available information on the genetic variation and population structure of *L. pneumophila* at the whole genome level. Environmental and clinical isolates are not evenly distributed in the environment based on sequence-based typing observations, suggesting that clinical isolates are a distinct sub-population of environmental strains. Humans are continuously exposed to environmental *Legionellae* and it is not clear why certain sequence types predominate in human disease[20]. One hypothesis is that disease only occurs in those who have increased susceptibility to infection, for example the elderly, and the immunosuppressed.[21] Whenever a *Legionella* outbreak occurs it usually reflects the breakdown of *Legionella* control measures, with human infections occurring as a consequence.

The genetic diversity of *Legionella* strains within an environmental source, as seen in this analysis, could potentially undermine our ability to link environmental and clinical

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

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

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

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, Bangsborg 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.

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, Bangsborg 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.

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.

Title

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

Authors

Sandra Reuter[1], Timothy G. Harrison[2], Claudio U. Köser[3,4], Matthew J. Ellington[4], Geoffrey P. Smith[5], Julian Parkhill[1], 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~~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 antigen[5] 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 subgrouping[8]) and genotypic (sequence-based typing[9]) 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* 0104:H4[12] 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.

Field Code Changed

Field Code Changed

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~~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

Field Code Changed

local hospital and a formal outbreak investigation was carried out.~~conducted.~~ Twenty-four further cases of LD were identified over the next three weeks. All cases had a positive *L. pneumophila* urinary antigen test, and three patients' samples were culture-positive for *L. pneumophila* serogroup 1. Epidemiological and environmental investigations were undertaken to determine possible sources. One hundred and forty-two environmental samples were collected from potential sources, which included 50 cooling towers on 11 premises. *L. pneumophila* serogroup 1 was isolated from samples collected at three cooling towers at two different locations (sites A and B) and a domestic spa pool. Clinical and environmental isolates were referred to the Respiratory and Systemic Infection Laboratory, Health Protection Agency, London, for *L. pneumophila* monoclonal antibody (mAb) subgrouping followed by a three allele DNA-sequence based typing (SBT_3) method then in use. The SBT_3 profiles for two of the clinical isolates and isolates from two of the cooling towers were indistinguishable, suggesting that the cooling towers were the likely environmental source. The strains were subsequently re-examined using the current seven allele sequence based typing (SBT) method,[15~~14~~] with the same outcome.

Field Code Changed

Field Code Changed

DNA extraction and whole genome sequencing

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

Bioinformatic analysis

Bioinformatic analysis was performed at the Wellcome Trust Sanger Institute and blinded to the epidemiological data. The sequencing data from the seven samples were mapped to a reference genome, *L. pneumophila* type strain Philadelphia-1,[16~~15~~] and compared with eight other publicly available *L. pneumophila* genomes (Table 1). Sequence reads were mapped onto the reference genome using SMALT. Regions containing phage or insertion sequence elements were excluded from the analysis.~~.~~ Single nucleotide polymorphisms (SNPs) were identified using a standard approach,[17~~16~~] by removing SNPs with low quality scores and by filtering for SNPs that were present in at least 75% of the mapped reads. The minimum 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

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.

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

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

Comparison of conventional typing and genomic analysis

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

DISCUSSION

Here, we have demonstrated the feasibility of using WGS to perform an investigation of a *Legionella* outbreak. Using genomic analysis we were readily able to distinguish outbreak from non-outbreak *Legionella* isolates, and to identify probable environmental sources, thus

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

Field Code Changed

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

Our analysis was also constrained by the limited available information on the genetic variation and population structure of *L. pneumophila* at the whole genome level. Environmental and clinical isolates are not evenly distributed in the environment based on sequence-based typing observations, suggesting that clinical isolates are a distinct sub-population of environmental strains. Humans are continuously exposed to environmental *Legionellae* and it is not clear why certain sequence types predominate in human disease[20]. One hypothesis is that disease only occurs in those who have increased susceptibility to infection, for example the elderly, and the immunosuppressed.[21~~21~~] Whenever a *Legionella*

Field Code Changed

outbreak occurs it usually reflects the breakdown of *Legionella* control measures, with human infections occurring as a consequence.

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

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

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.

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, Bangsborg 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, Brodkin 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, Bangsborg 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.

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.

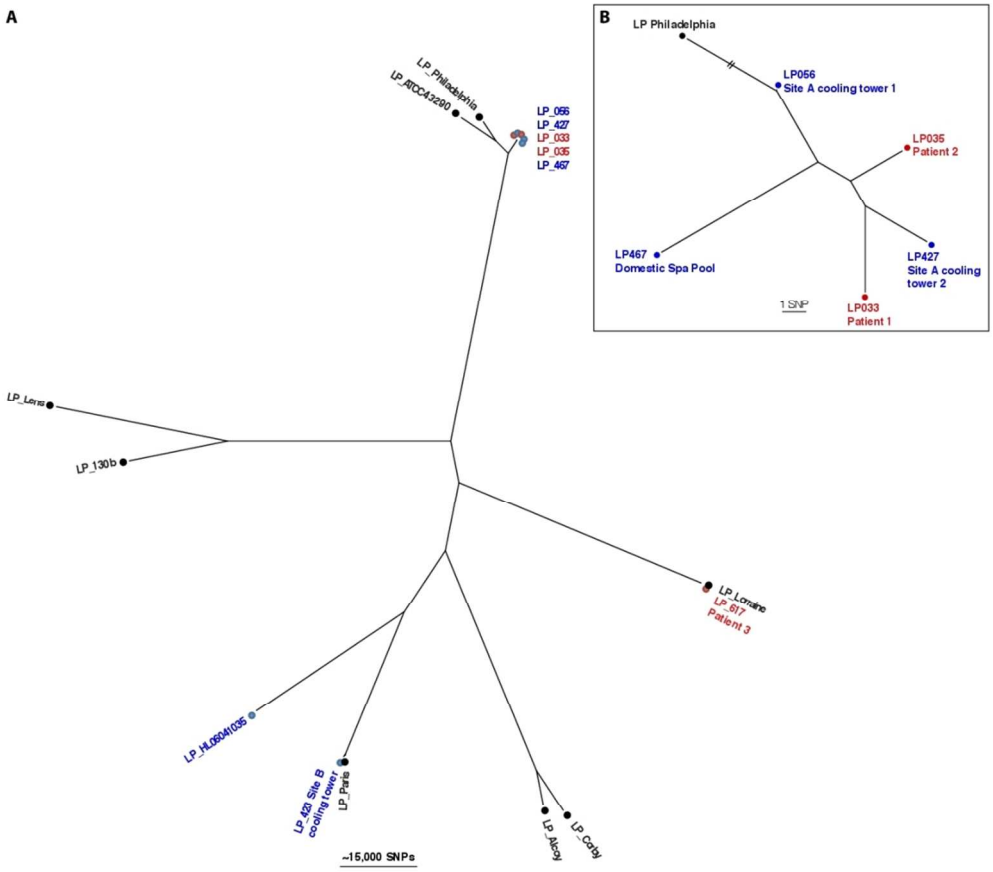


Figure 1. Phylogenetic tree of Legionella pneumophila strains
A. Phylogeny of the species L. pneumophila. Clinical, environmental and references isolates are shown in red, blue, and black, respectively. Inset B. Close-up phylogeny of the isolates involved in the outbreak. The branch leading to the reference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

99x90mm (300 x 300 DPI)

PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak
AUTHORS	Torok, Estee; Reuter, Sandra; Harrison, Tim; Köser, Claudio; Ellington, Matthew; Smith, Geoffrey; Parkhill, Julian; Peacock, Sharon; Bentley, Stephen

VERSION 1 - REVIEW

REVIEWER	Morag Graham Ph.D. Research Scientist and Chief, Genomics Public Health Agency of Canada Canada This peer reviewer has no competing interests
REVIEW RETURNED	09-Nov-2012

THE STUDY	Statistical analysis is not necessry for phylogenomic analysis of bacterial WGS with such a small study size. Essentially N/A. The authors satisfactorily address in the discussion the limited sampling size.
GENERAL COMMENTS	General comments: Owing to widespread occurrence of Legionella pneumophila (Lpn) in both natural and artificial aquatic systems, it is important to implement early prevention measures. To do so, it is necessary to quickly identify potential environmental sources of infection by comparing clinical and environmental isolates. This manuscript explores whole-genome sequencing (WGS) as an approach for analyzing a L. pneumophila outbreak. As a pilot study, a 2003 outbreak was retrospectively explored - initially comprised of 28 cases, but only 3 cases were Lpn culture-positive. The team sequenced these 3 clinical and 4 isolates (of 142) environmental samples (from over 50 cooling towers on 11 premises). They then analyzed the WGS output to identify which isolates were closest at the nucleotide level. The approach for generating the inferred phylogeny was to reference map the Illumina MiSeq read data for each sequenced isolate against the most closely related Lpn bacterial genome (Philadelphia strain) as reference using SMALT; then extract high-quality single nucleotide polymorphisms (SNPs) and build a maximum likelihood phylogeny from the SNP data set. Phylogenetic analysis showed that two clinical isolates (LP033 and LP035) and three environmental isolates (LP056, LP427 and LP467) were closely related (Figure 1), with only a small number of SNPs between them. One clinical (LP617) and the two remaining environmental isolates (LP423 and LP617) were genetically more distant; thus, it was concluded they were not part of the outbreak. The data was congruent with the previous epidemiological analysis. Overall this is a technically sound paper and well written, albeit brief.

	Although the approach is not really ground-breaking, the conclusions are valid and I thought they introduced the topic well. Given this journal is aimed at an open medical community; the brevity of the manuscript is probably fine. I was pleased to see the authors rightfully discuss the limitation of their sample size as a major study caveat. And the authors are correct in concluding “the genetic diversity of Legionella strains within an environmental source, as seen in this analysis, could potentially undermine our ability to link environmental and clinical isolates in an outbreak situation. Thus a detailed epidemiological investigation accompanied by thorough environmental sampling, sequencing and comparison with patient isolates will continue to be required to confirm the likely source of an outbreak.” Future real-time applications of WGS for outbreak investigations will most certainly require expanded sampling and sequencing, including repeat sequencing of templates from single isolated colonies for the same sampling source. Fortunately, the throughput and cost of sequencing technologies today are no longer limiting. A few minor questions: 1. Was there any citation for the original 2003 Lpn outbreak epidemiological investigation? If so, then it should be included. 2. The bioinformatics methods section is very brief. Although cut-off values used for identifying SNPs were mentioned (SNP needs to be present in at least 75% of mapped reads) and paired-end reads were generated, there was no mention of a minimum coverage value for identifying SNPs. Minimum read coverage is a relevant value to include as it conveys information about the SNP call confidence. 3. Although mentioned that regions containing phage or insertion sequence were removed from the analysis, it was not mentioned whether repeat regions on the reference genome also were removed from the analysis, whether manual curation was conducted or whether there were any issues in repetitive regions. The genome of Philadelphia strain has 26 intragenic tandem repeat sequences, many of which have been found to be "polymorphic" in repeat copy number (PMIDs:19077205; 21821761). As written, it was hard to determine whether this was captured in the whole-genome analysis? Moreover, it would be interesting to look at the difference in tandem repeat distribution as a function of clinical or environmental origin. 4. The isolates identified as outbreak isolates were found to have at most 15 SNPs between them. These SNPs were identified by reference mapping to Philadelphia, which was found to be the closest publicly available genome. It would be interesting to see if more SNPs could be identified by reference mapping the reads to an assembly of one of the outbreak strains (a within outbreak analysis). It may be that many more SNPs may not be identified given they already excluded phage/insertion sequences. However, it might provide more information regarding Lpn intra-outbreak diversity. Of course, as mentioned, this would also be benefited from a larger number of sequenced isolates. 5. The results indicate that two environmental isolates (LP423 and LP617) were ~75,000 to 77,500 SNPs away from the outbreak cluster. Could it be that more than one Lpn population existed within the environmental templates grown in broth? i.e., did each DNA template originate from an independent individual bacterial colony from a culture plate? Given Legionella are so ubiquitous, perhaps sequencing of templates recovered from several individual colonies

	in parallel would rule out mixed Lpn populations and increase confidence that all environmental sources have been exhaustively analyzed. As sequencing technologies are more affordable and increasingly require less template to prepare libraries, this conservative and prudent approach is becoming feasible. 6. Ref 20 needs to have the year corrected.

REVIEWER	Sophie Jarraud, PharmD, PhD, National Reference Centre for Legionella, France. No conflict of interests.
REVIEW RETURNED	11-Nov-2012

THE STUDY	statistical methods: data not necessary
GENERAL COMMENTS	Manuscript very interesting describing the WGS approach for the investigation of a Legionella outbreak. The authors described well the potentiel power of this method but also the limits especially the limited available information on the genetic variation of L. pneumophila at the whole genome level.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

General comments:
Owing to widespread occurrence of Legionella pneumophila (Lpn) in both natural and artificial aquatic systems, it is important to implement early prevention measures. To do so, it is necessary to quickly identify potential environmental sources of infection by comparing clinical and environmental isolates. This manuscript explores whole-genome sequencing (WGS) as an approach for analyzing a L. pneumophila outbreak. As a pilot study, a 2003 outbreak was retrospectively explored - initially comprised of 28 cases, but only 3 cases were Lpn culture-positive. The team sequenced these 3 clinical and 4 isolates (of 142) environmental samples (from over 50 cooling towers on 11 premises). They then analyzed the WGS output to identify which isolates were closest at the nucleotide level. The approach for generating the inferred phylogeny was to reference map the Illumina MiSeq read data for each sequenced isolate against the most closely related Lpn bacterial genome (Philadelphia strain) as reference using SMALT; then extract high-quality single nucleotide polymorphisms (SNPs) and build a maximum likelihood phylogeny from the SNP data set. Phylogenetic analysis showed that two clinical isolates (LP033 and LP035) and three environmental isolates (LP056, LP427 and LP467) were closely related (Figure 1), with only a small number of SNPs between them. One clinical (LP617) and the two remaining environmental isolates (LP423 and LP617) were genetically more distant; thus, it was concluded they were not part of the outbreak. The data was congruent with the previous epidemiological analysis.

Overall this is a technically sound paper and well written, albeit brief. Although the approach is not really ground-breaking, the conclusions are valid and I thought they introduced the topic well. Given this journal is aimed at an open medical community; the brevity of the manuscript is probably fine.

We thank the reviewer for this comment – the length of the manuscript is dictated by the Journal's instructions to authors.

I was pleased to see the authors rightfully discuss the limitation of their sample size as a major study caveat. And the authors are correct in concluding "the genetic diversity of Legionella strains within an

environmental source, as seen in this analysis, could potentially undermine our ability to link environmental and clinical isolates in an outbreak situation. Thus a detailed epidemiological investigation accompanied by thorough environmental sampling, sequencing and comparison with patient isolates will continue to be required to confirm the likely source of an outbreak.

We agree entirely with the reviewer

Future real-time applications of WGS for outbreak investigations will most certainly require expanded sampling and sequencing, including repeat sequencing of templates from single isolated colonies for the same sampling source. Fortunately, the throughput and cost of sequencing technologies today are no longer limiting.

We agree with the reviewer on this point

A few minor questions:
1. Was there any citation for the original 2003 Lpn outbreak epidemiological investigation? If so, then it should be included.

Response: The original legionella outbreak investigation is described in reference number 14 (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)

2. The bioinformatics methods section is very brief. Although cut-off values used for identifying SNPs were mentioned (SNP needs to be present in at least 75% of mapped reads) and paired-end reads were generated, there was no mention of a minimum coverage value for identifying SNPs. Minimum read coverage is a relevant value to include as it conveys information about the SNP call confidence.

Response: We agree with the reviewer on this point and have accordingly added the relevant detail to the manuscript. Briefly, the minimum number of high quality reads mapping to call a base is set to 4. This is equivalent to a minimum coverage of 4.

3. Although mentioned that regions containing phage or insertion sequence were removed from the analysis, it was not mentioned whether repeat regions on the reference genome also were removed from the analysis, whether manual curation was conducted or whether there were any issues in repetitive regions. The genome of Philadelphia strain has 26 intragenic tandem repeat sequences, many of which have been found to be "polymorphic" in repeat copy number (PMIDs:19077205; 21821761). As written, it was hard to determine whether this was captured in the whole-genome analysis? Moreover, it would be interesting to look at the difference in tandem repeat distribution as a function of clinical or environmental origin.

Response: Thank you for this comment. Tandem repeats were not considered in the analysis, We did, however, re-run the analysis excluding the 23 repetitive genes mentioned in Coil et al 2008 (PMID 19077205), to show that the overall topology of the phylogenetic tree remains unchanged so would not have affected interpretation of the data.

4. The isolates identified as outbreak isolates were found to have at most 15 SNPs between them. These SNPs were identified by reference mapping to Philadelphia, which was found to be the closest publicly available genome. It would be interesting to see if more SNPs could be identified by reference mapping the reads to an assembly of one of the outbreak strains (a within outbreak analysis). It may be that many more SNPs may not be identified given they already excluded phage/insertion sequences. However, it might provide more information regarding Lpn intra-outbreak diversity. Of course, as mentioned, this would also be benefited from a larger number of sequenced isolates.

Response: We thank the reviewer for this useful comment. We agree that using the draft assembly of one of the outbreak strains as the reference for mapping may have the potential to identify SNPs not detectable when mapping to the more distant Philadelphia strain. However, this would be balanced with the potential for false positive SNP calls due to base mis-calling in the draft assembly. Draft assemblies also have poor base quality at the ends of the many contigs introducing further potential for miscalling of SNPs. Some errors could be ruled out by manual curation but we propose that it would be inappropriate to pursue such an approach in an outbreak situation unless an appropriate reference is available. Overall, given the small numbers of SNP differences between the outbreak isolates, we feel that using a draft assembly of one of the outbreak isolates would give no advantage over using the finished Philadelphia strain sequence.

5. The results indicate that two environmental isolates (LP423 and LP617) were ~75,000 to 77,500 SNPs away from the outbreak cluster. Could it be that more than one Lpn population existed within the environmental templates grown in broth? i.e., did each DNA template originate from an independent individual bacterial colony from a culture plate? Given Legionella are so ubiquitous, perhaps sequencing of templates recovered from several individual colonies in parallel would rule out mixed Lpn populations and increase confidence that all environmental sources have been exhaustively analyzed. As sequencing technologies are more affordable and increasingly require less template to prepare libraries, this conservative and prudent approach is becoming feasible.

Response: The reviewer is correct in assuming that environmental samples frequently contain multiple strains. However in the original investigation we examined multiple isolates from each environmental sample to confirm their phenotype (species, serogroup and monoclonal antibody subgroup). In this investigation each sample (and source) only contained a single phenotype – hence only a single colony for each sample was characterised genotypically and archived for later use. For the clinical samples five colonies were taken from each positive patient sample and characterised phenotypically (species, serogroup and monoclonal antibody subgroup). Again only a single phenotype was identified in each patient and hence only a single colony from each was characterised genotypically.

6. Ref 20 needs to have the year corrected.

Response: We have corrected this reference.

Reviewer 2

Manuscript very interesting describing the WGS approach for the investigation of a Legionella outbreak. The authors described well the potentiel power of this method but also the limits especially the limited available information on the genetic variation of L. pneumophila at the whole genome level.

We thank the reviewer for their comments.

VERSION 2 – REVIEW

REVIEWER	Morag Graham Ph.D. Research Scientist and Chief, Genomics Public Health Agency of Canada Canada This peer reviewer has no competing interests.
REVIEW RETURNED	03-Dec-2012

THE STUDY	Statistical analysis not essential for this small sample size phylogenomic analysis. The authors adequately addressed the limited sample size in the discussion.
GENERAL COMMENTS	Although 20x genome coverage and 4 high quality reads to call each base [with 75% (or 3 reads) then determining a SNP relative to reference genome] are already considered (in the field) to be low genome/read coverage - this is a proof of concept study. Such a relaxed approach will enable SNP detection even for strains with low read coverage. However, it is important to note that inherent errors owing to platform bias may get through with such low coverage and lab experiments to verify SNPs is advisable to rule out false positives. With enhanced genome coverage and reads, overall accuracy and data confidence will improve, with the added advantage of reduced need for extra wet-lab work.