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.