PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Seeking the source of Pseudomonas aeruginosa infections in a recently opened hospital: an observational study using whole-genome sequencing
AUTHORS	Loman, Nicholas; Quick, Josh; Cumley, Nicola; Wearn, Chris; Niebel, Marc; Constantinidou, Chrystala; Thomas, Chris; Pallen, Mark; Moiemen, Naiem; Bamford, Amy; Oppenheim, Beryl

VERSION 1 - REVIEW

REVIEWER	Dr. Jennifer Gardy, Senior Scientist BC Centre for Disease Control, Vancouver, Canada
REVIEW RETURNED	08-Aug-2014

GENERAL COMMENTS	In this interesting, comprehensive, and well-written report from Quick et al, the authors use a genomic epidemiology approach to examine the contribution of hospital plumbing/water to the spread of Pseudomonas aeruginosa on a burns ward. The study is amongst the first few papers to use whole genome sequencing (WGS) to investigate Pseudomonas epidemiology, and is, to my knowledge, the largest such study to date and the first to examine multiple water sources (e.g. shower, drain, tap, sink) as potential sources of patient infections. The introduction and rationale for the study are clearly presented, and the methods are nicely detailed with the sequences being publicly available, making the study as described quite reproducible. My only suggestions for improvement are below, and relate primarily to making the flow of the analysis in the Results sections clearer to the reader within the text itself. I have numbered them according to the Results section to which they pertain. 1. The Results section begins with a description of recruitment into the screening phase of the study, which identified 30 patients. There should be some follow-up text of how many of these 30 tested positive for Pseudomonas (I had to count out the five from Figure 2). 2A. The paper then goes directly into Results section 2, discussing 83 environmental isolates that were sequenced, and shows them in Figure 1, which also includes the human isolates sequenced, however these are not mentioned in the text. I tried to deduce that if you sequenced 141 genomes as noted in the abstract and 83 were environmental as noted in Results I, 58 had to be human, coming from 5 patients (inferred from Figure 2), but this didn't reconcile with the number of isolates I counted as coming from patients in Figure 2. This section should be modified to make it crystal-clear exactly how many genomes were sequenced and from which sources. Figure 1 is nice to show first, as it sets the stage for all the genomes in a

	global context, followed by Figure 2, as long as I can easily follow all the isolates from Figure 1 over onto Figure 2 (there seem to be fewer circles on Figure 2 than genomes suggested by Figure 1 - does one circle comprise multiple genomes? If so, indicate this with a number inside each circle corresponding to # of sequenced genomes). 2B. In Figure 2, you might want to colour the circles using a different colour scheme - the colours used to denote clades in this figure are the same used to denote isolate source in the other figures, which can create confusion. I would suggest colouring the branches in the Figure 1 phylogeny with new colours, and then using that same branch colouring scheme to colour the circles in Figure 2. The ;tldr version of comments 2A and B is basically that it's quite hard to follow how much was sequenced and from whom, so tweak the second section of the Results to make it much clearer, without the reader having to refer to figures. 3A. In Results section 3, the source of the five patients' infections is discussed in light of the genomic data collected from their wounds and their environments. One patient is referred to as Patient E while the others are denoted by numbers (I'm assuming E=5). 3B. For the clade E patients and data in Figure 3, it would be helpful to label the isolates in Figure 3 according to their patient source. I had to deduce from Figure 2 that the Day 14 isolate was patient 1 and then rest came from patient 4. 3C. A BEAST phylogeny incorporating sampling dates would be most informative here, particular when it comes to sorting out the timing of TMRCAs. As it stands at the moment, one could conceivably think that the source of the Bed 11 similar isolates was patient 1's Day 14 groin sample, given that it was the earliest sequence sampled. BEAST would also allow you to report mutation rates and examine whether they vary across the hospital settings. A Bayesian Skyline analysis of Clade E might also be informative as to outbreak dynamics. 3D. In the online appendices showing the phylogenies for the other three clades, it would also help to convert dates as shown now into days as shown in Figure 3 - it is easier to understand the sequence of events this way. 3E. In the online appendices for the clade genomics, the labels in the Patient column aren't quite clear to me - e.g. some Specimen Types that say Water have varying labels in the Patient column, from P, to SP, to "Water sampling". I am assuming P = patient and SP = screening patient, but how are you getting water out of patients? It doesn't seem to correlate with shower pre/post flush or anything else. I am confuzzled. 3F. In Figure 2, patient 2's box contains isolates from both clades C and E, but this is not touched upon in the text. 3G. Just as a random thought should the opportunity ever arise, it would be quite interesting to root deeper into the plumbing, as with the TMV shotgun section, and see if you could find the TMRCB (The Most Recent Common Biofilm), from which Clade E descended, which presumably is deep in some pipe that feeds all the showers

	on the ward. Perhaps it would even contain representatives of the other clades. Overall, this a fascinating and comprehensive paper providing a very interesting and detailed look at the hospital ecology of Pseudomonas and its impact upon patient care. Happy to recommend pending minor revisions for clarity suggested above. NB: There is a rogue track changes comment in the first page of the STROBE checklist in the supplement.

REVIEWER	Alan McNally Nottingham Trent University
REVIEW RETURNED	12-Aug-2014

GENERAL COMMENTS	The authors present an extremely well written and well conducted investigation of the dissemination of Pseudomonas aeruginosa in a Hospital burns unit. The study has been performed to an extremely high level and I consider it ready to publish. I have just a few small points the authors may wish to consider to slightly improve the manuscript for a general audience: Line 23 Page 9: Why were 10 patients expected to have acquired infection by this time point? Line 47 Page 12: The selection of 3 SNPs/1000nt as an indication of possible recombination could be regarded by some readers as arbitrary, particularly in light of Bayesian methods for determining recombination. This would be strenghtened with a small supporting rationale or reference Line 44, page 14: This may seem very pedantic but am not sure this can be considered as detecting transmission. Maybe the authors could consider calling this tracking of strains dissemination? Line 24, page 15: Is Patient E accurate or should this be a numbered patient? Line 55, page 28: I think this should read (panel B)

REVIEWER	David Baltrus University of Arizona Tucson, USA
REVIEW RETURNED	24-Aug-2014

GENERAL COMMENTS	This manuscript from Quick et al. demonstrates, in a very straightforward way, the utility of using whole genome sequencing for epidemiology. The authors do a good job placing this study in context, while also highlighting the power of current sequencing technologies to trace sources of outbreaks. I think this paper will be well received, both from the story point of view and clinical relevance. Well done throughout. I don‟t really have any major critiques aside from a suggestion to be a little bit more cautious (although the authors are usually quite cautious throughout) with nuances in wording. You can‟t really say that any of the clades are truly absent from any of the sites because

	it‟s impossible to sample all cells. I only say this because there are a couple of places, notably in the Fig. 3 legend, where this kind of resolution matters. In this legend, “This also indicated” is better phrased as “This suggests”. Along these lines, and I‟m fully aware of the smallness of the point, but isn‟t it possible given your sampling scheme that patient 1 in room 11 actually seeded the tap samples? Very low probability, but possible. MINOR CRITIQUES I couldn‟t help thinking that the manuscript would be made a bit stronger by a comparison of data obtained by whole genome sequencing with patterns obtained by MLST, just to reinforce the power of newer technologies. Doesn‟t even need to be a figure just a couple of lines in the discussion saying that with MLST you would be able to discriminate across clades (I think) but not within clades. Same with PFGE and VNTR unless you got really lucky and the random bits of genome sampled by these methods actually had SNPs. Page 9: line 57: how was water taken from the shower/tap. Just spraying into container? Page 10: line 43: would be good to include average coverage for each strain somewhere in methods/results. You say later you picked variants with 80% representation, but is this 8/10 reads or 80/100? I‟m guessing the later, but would still be good to include this info for other groups to replicate. Page 10: line 56/57: the nomenclature is a bit confusing historically, but I think you mean SMRT cell instead of zero mode waveguides here Page 11: line 31: would be good to just say that the species specific 16s PCR is a presence/absence, band=P. aeruginosa presence test Page 11: line 58: recombination, Darwinian selection...or could be misalignment of reads no? Page 14: line 36. “in burns unit” better as “in the burn unit‟s” Page 15: line 24. Why is this Patient E when all other patients are numerical? Page 15: line 26. two periods (..) at end of sentence Page 16: line 8 (and throughout manuscript) in various places the word clade is capitalized and in other‟s it‟s not. You also refer to “clone E” --page 19 line 54--instead of clade E at another point. I would suggest modifying each case so the manuscript is consistent throughout Page 16: line 39-41. how did you know these were plasmids if you only had MiSeq paired ends? I doubt they would assemble completely, or did they? Page 16: line 54 There aren‟t any published reports of natural transformation in P. aeruginosa as of 2010. Any as far as I know, nothing has been published on competence in P.aeruginosa since then (I could be wrong...) http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3008168/

	Figure legend 1: Maybe include a reference to the Stewart paper in legend (reference 27) as well as any other references needed in order to say where all other strains in tree came from. Figure legend 2: Reading the legend and seeing the figure, I"m still a bit unclear as to what the Y-axis means. Maybe reference appendix 5 in legend and rewrite to be more descriptive.

REVIEWER	Thomas Connor Cardiff University, UK
REVIEW RETURNED	27-Aug-2014

GENERAL COMMENTS	I enjoyed the paper very much, and think that it is both scientifically sound, and thought provoking, and think that it warrants publication. I do have a few minor questions/corrections that I would like to put to the authors. Some of these (1-3) relate to added detail that makes it easier for a reader to assess the strength of the conclusions drawn, while the others (4-6) are questions that are a little outside the scope of the current paper, but could be of interest to other bacterial genomics researchers. Questions 1. The authors state that the samples were mapped, and then recombinant regions were removed, but don't provide specific information on what the SNP scale on their trees refers to; I am interested to know what % mapping/core genome size you were left with for your isolates on a population wide level (i.e. the data used to generate figure 1A) when you came to draw your tree. Also, and related to this, I am curious how well the use of a variant density filter and the VarScan thresholds worked to remove phage and other mobile elements from the data for the population wide results. Also, how did you simulate the 600,000 reads for previously published strains? 2. In terms of the fine-scale trees, I am would like to know a little more about your mapping results. Did your mapping approach lead to 'N's being called (i.e. the mapping software is unable to make a call) or did you only include SNPs where you had a definitive mapped base for every sample? If so, how many SNP sites did you discard? Also what software did you use to detect indels? did you use a standalone piece of software (eg GATK), or was this just done based on the BWA results followed by variant calling? 3. The two plasmids pBURNS 1 and 2 are covered only briefly, did the authors look at these in more detail? specifically, how large are they? are they similar to other plasmids that have been reported elsewhere? and did the authors examine the phylogeny of pBURNS 1, and was it consistent with the chromosomal phylogeny, or was there more diversity present than might be expected (which might tell you something about acquisitions of the plasmids in the environment)? 4.

	The collection includes a number of isolates from environmental sources, I wondered about their genomic content too; did you find any other plasmids (other than pBURNS1 and 2) in their assembled genomes, and did you look for (or find) evidence of historical recombination between isolates that had been isolated from the water supply? 5. Did the Pseudomonas colonise the same site as the Acinetobacter baumannii? and did you sequence the Acinetobacter strain? (and, if so, did you see any evidence of genetic exchange?) 6. Did the you try mapping the DNA in the biofilm to phage or other microbial species such as those from the archaea? Other minor points Page 12 line 21/22 - does this mean that you only used positions that were present in all samples? is a little unclear. Page 13, line 7; I think pubMLST like the citation at the end of this link ; http://www.biomedcentral.com/1471-2105/11/595/abstract You may also / probably should include a short description of what MLST is (just a sentence), and probably a citation to the MLST paper (available at - http://www.pnas.org/cgi/pmidlookup?view=long&pmid=9501229) Page 14 line 2 -this is a repeat of what is on page 12/13 - "phylogenetic trees were generated using FastTree and FigTree" but I think this should read "phylogenetic trees were generated using FastTree and visualised using FigTree", although as it is redundant, you could probably just remove it.

VERSION 1 – AUTHOR RESPONSE

Reviewer Name Dr. Jennifer Gardy, Senior Scientist
Institution and Country BC Centre for Disease Control, Vancouver, Canada
Please state any competing interests or state „None declared": None declared

In this interesting, comprehensive, and well-written report from Quick et al, the authors use a genomic epidemiology approach to examine the contribution of hospital plumbing/water to the spread of Pseudomonas aeruginosa on a burns ward. The study is amongst the first few papers to use whole genome sequencing (WGS) to investigate Pseudomonas epidemiology, and is, to my knowledge, the largest such study to date and the first to examine multiple water sources (e.g. shower, drain, tap, sink) as potential sources of patient infections.

The introduction and rationale for the study are clearly presented, and the methods are nicely detailed with the sequences being publicly available, making the study as described quite reproducible. My only suggestions for improvement are below, and relate primarily to making the flow of the analysis in the Results sections clearer to the reader within the text itself. I have numbered them according to the Results section to which they pertain.

We are very grateful to Dr. Gardy for her enthusiastic reaction to our paper, and to the extremely useful comments and suggestions which are primarily aimed at clarity of presentation.

1. The Results section begins with a description of recruitment into the screening phase of the study, which identified 30 patients. There should be some follow-up text of how many of these 30 tested positive for Pseudomonas (I had to count out the five from Figure 2).

Five patients were positive for P. aeruginosa (3 in burns wound only, 1 in burn and urine and 1 in sputum). We have added two explanatory sentences to the Results section to clarify this.

2A. The paper then goes directly into Results section 2, discussing 83 environmental isolates that were sequenced, and shows them in Figure 1, which also includes the human isolates sequenced, however these are not mentioned in the text. I tried to deduce that if you sequenced 141 genomes as noted in the abstract and 83 were environmental as noted in Results I, 58 had to be human, coming from 5 patients (inferred from Figure 2), but this didn't reconcile with the number of isolates I counted as coming from patients in Figure 2. This section should be modified to make it crystal-clear exactly how many genomes were sequenced and from which sources. Figure 1 is nice to show first, as it sets the stage for all the genomes in a global context, followed by Figure 2, as long as I can easily follow all the isolates from Figure 1 over onto Figure 2 (there seem to be fewer circles on Figure 2 than genomes suggested by Figure 1 - does one circle comprise multiple genomes? If so, indicate this with a number inside each circle corresponding to # of sequenced genomes).

Of the 141 genomes sequenced, 86 isolates were environmental and 55 were patient isolates. We have added a sentence to the text to describe the number of positive samples from patients which was an important omission. The reason that the number of patient isolates could not be deduced from Figure 2 was because if two or more isolates of the same type were collected on the same day then they were collapsed into a single icon. This point has also been clarified in the figure legend.

2B. In Figure 2, you might want to colour the circles using a different colour scheme - the colours used to denote clades in this figure are the same used to denote isolate source in the other figures, which can create confusion. I would suggest colouring the branches in the Figure 1 phylogeny with new colours, and then using that same branch colouring scheme to colour the circles in Figure 2. The ;tldr version of comments 2A and B is basically that it's quite hard to follow how much was sequenced and from whom, so tweak the second section of the Results to make it much clearer, without the reader having to refer to figures.

Thanks for this very useful suggestion. We agree that consistent use of colours would aid interpretation. Consequently we have changed Figure 2 so that the colours represent sample type so they are now consistent with those in Figure 1 and 3, reducing the chance of confusion. In order to do this the clade is now represented by its letter, which simplifies the look of the figure and offers an overall improvement.

3A. In Results section 3, the source of the five patients' infections is discussed in light of the genomic data collected from their wounds and their environments. One patient is referred to as Patient E while the others are denoted by numbers (I'm assuming E=5).

Patient E is Patient 5. We apologise for this error and have corrected it accordingly in the manuscript (this error was also noted by other reviewers).

3B. For the clade E patients and data in Figure 3, it would be helpful to label the isolates in Figure 3 according to their patient source. I had to deduce from Figure 2 that the Day 14 isolate was patient 1 and then rest came from patient 4.

A extra column has been added to the figure to display this information.

3C. A BEAST phylogeny incorporating sampling dates would be most informative here, particular when it comes to sorting out the timing of TMRCAs. As it stands at the moment, one could conceivably think that the source of the Bed 11 similar isolates was patient 1's Day 14 groin sample, given that it was the earliest sequence sampled. BEAST would also allow you to report mutation rates and examine whether they vary across the hospital settings. A Bayesian Skyline analysis of Clade E might also be informative as to outbreak dynamics.

This is an interesting suggestion. In order to investigate whether this would be a useful approach with our dataset, we plotted sampling date against root-to-tip distance plot using the Path-O-Gen software (http://tree.bio.ed.ac.uk/software/pathogen/). Using the tree rooted according to an outgroup (PAO1) the software produced the following plot:

Allowing the Patho-O-Gen software to choose the best-fitting root gives a poor correlation coefficient and R-squared values:

Given these results, we did not proceed to a BEAST analysis.

Our hypothesis is that our sampling period may be too short to get useful signal in this dataset. Sampling is also necessarily uneven, with many samples being taken around the time of positive patients. We are in the process of extending this study to the wider hospital and, assuming this clone is detected again, we will re-visit this problem when we have a larger dataset collected over a longer timescale (2-3 years).

3D. In the online appendices showing the phylogenies for the other three clades, it would also help to convert dates as shown now into days as shown in Figure 3 - it is easier to understand the sequence of events this way.

Thank you, we have made this change.

3E. In the online appendices for the clade genomics, the labels in the Patient column aren't quite clear to me - e.g. some Specimen Types that say Water have varying labels in the Patient column, from P, to SP, to "Water sampling". I am assuming P = patient and SP = screening patient, but how are you getting water out of patients? It doesn't seem to correlate with shower pre/post flush or anything else. I am confuzzled.

Yes, P = patient, SP = screening patient. If water samples were collected from a patient‟s room during their stay they were described as „water sample‟ but also ascribed to a patient number. In the cases where water sampling was performed as per protocol but there was no study patient in the room then it was just described at „water sampling‟. We have added some explanatory text to the figure legend.

3F. In Figure 2, patient 2's box contains isolates from both clades C and E, but this is not touched upon in the text.

Clade C isolates were detected in wound of patient 2 however not from the water in the patient‟s room. Our interpretation is that the patient was admitted to hospital already colonised with an alternate strain to the clade E „water clone‟ isolated in his room and across the ward. Environmental samples of both clades were collected, this is because the patient‟s environment can be contaminated by water (usually the wet environment) or contaminated by the patient themselves.

3G. Just as a random thought should the opportunity ever arise, it would be quite interesting to root deeper into the plumbing, as with the TMV shotgun section, and see if you could find the TMRCB (The Most Recent Common Biofilm), from which Clade E descended, which presumably is deep in some pipe that feeds all the showers on the ward. Perhaps it would even contain representatives of the other clades.

This is a great suggestion. We did start to look at this after the study ended by attempting to culture P. aeruginosa from plumbing parts taken from the common water supply to the burns ward rooms but we were not able to grow any. This of course does not conclusively rule out the water supply as the original source of P. aeruginosa as Pseudomonas may have been present transiently. However, we note that other groups have found P. aeruginosa in new plumbing fittings supplied by manufacturers.

Overall, this a fascinating and comprehensive paper providing a very interesting and detailed look at the hospital ecology of Pseudomonas and its impact upon patient care. Happy to recommend pending minor revisions for clarity suggested above.

Thank you!

NB: There is a rogue track changes comment in the first page of the STROBE checklist in the supplement.

This has been resolved.

Reviewer Name Alan McNally
Institution and Country Nottingham Trent University
Please state any competing interests or state „None declared": None declared

The authors present an extremely well written and well conducted investigation of the dissemination of Pseudomonas aeruginosa in a Hospital burns unit. The study has been performed to an extremely high level and I consider it ready to publish. I have just a few small points the authors may wish to consider to slightly improve the manuscript for a general audience:

We are pleased that Dr McNally had such a positive reaction to our manuscript and thank him for his comments.

Line 23 Page 9: Why were 10 patients expected to have acquired infection by this time point?

This was based on the results of a previously performed local audit (unpublished) which identified the colonisation rate of burns patients to be approximately one third. We have clarified this point in the text.

Line 47 Page 12: The selection of 3 SNPs/1000nt as an indication of possible recombination could be regarded by some readers as arbitrary, particularly in light of Bayesian methods for determining recombination. This would be strenghtened with a small supporting rationale or reference

This rate of filtering has been used in previous studies (e.g. Holt 2008 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2652037/), although the use of more sophisticated statistical approaches such as BRATNextGen or Gubbins is preferred for highly recombinogenic organisms like S. pneumoniae or N. meningitidis. The validity of this heuristic filter was determined by plotting SNP density across a sliding window. The 3 SNPs in 1000 bases was deemed suitable as a

threshold that would identify clear SNP clusters, which may represent recombination.

Line 44, page 14: This may seem very pedantic but am not sure this can be considered as detecting transmission. Maybe the authors could consider calling this tracking of strains dissemination?

We appreciate the point made here, we have changed this to “Inferring potential transmission events by whole-genome sequencing” to make it clearer what the limits of WGS analysis are.

Line 24, page 15: Is Patient E accurate or should this be a numbered patient?

This should be Patient 5 and this has been corrected in the manuscript.

Line 55, page 28: I think this should read (panel B)

Thank you, this has been corrected in the manuscript to read panel B.

Reviewer Name David Baltrus
Institution and Country University of Arizona
Tucson, USA
Please state any competing interests or state „None declared": none declared

This manuscript from Quick et al. demonstrates, in a very straightforward way, the utility of using whole genome sequencing for epidemiology. The authors do a good job placing this study in context, while also highlighting the power of current sequencing technologies to trace sources of outbreaks. I think this paper will be well received, both from the story point of view and clinical relevance. Well done throughout.

I don"t really have any major critiques aside from a suggestion to be a little bit more cautious (although the authors are usually quite cautious throughout) with nuances in wording. You can"t really say that any of the clades are truly absent from any of the sites because it"s impossible to sample all cells. I only say this because there are a couple of places, notably in the Fig. 3 legend, where this kind of resolution matters. In this legend, “This also indicated” is better phrased as “This suggests”. Along these lines, and I"m fully aware of the smallness of the point, but isn"t it possible given your sampling scheme that patient 1 in room 11 actually seeded the tap samples? Very low probability, but possible.

We thank Dr Baltrus for his positive comments on our manuscript and for his useful suggestions. We have made the required change to the Figure 3 legend. We agree it is theoretically possible that patient 1 in room 11 seeded the tap samples, but we provide multiple lines of evidence in the discussion to argue that this is unlikely to be the case. The most important is that this clone is distributed across many rooms, and therefore multiple transmission events would be required from patient to multiple water outlets, which seems highly unlikely.

MINOR CRITIQUES

I couldn"t help thinking that the manuscript would be made a bit stronger by a comparison of data obtained by whole genome sequencing with patterns obtained by MLST, just to reinforce the power of newer technologies. Doesn"t even need to be a figure just a couple of lines in the discussion saying that with MLST you would be able to discriminate across clades (I think) but not within clades. Same with PFGE and VNTR unless you got really lucky and the random bits of genome sampled by these methods actually had SNPs.

We did not want to focus too much on MLST because these points have been made before, but we have added a sentence to the discussion.

Page 9: line 57: how was water taken from the shower/tap. Just spraying into container?

Essentially, yes. At least 200ml of shower water was collected into a vessel containing sodium thiosulphate as a neutralizer. We have added this to the text.

Page 10: line 43: would be good to include average coverage for each strain somewhere in methods/results. You say later you picked variants with 80% representation, but is this 8/10 reads or 80/100? I"m guessing the later, but would still be good to include this info for other groups to replicate.

Our strains had a mean coverage of 24.4x, with the lowest covered strain being 14x and highest 64.7x. We have added this detail to the results. Short read data is available in the SRA if per-sample coverage is required.

Page 10: line 56/57: the nomenclature is a bit confusing historically, but I think you mean SMRT cell instead of zero mode waveguides here

Thank you, we have corrected this.

Page 11: line 31: would be good to just say that the species specific 16s PCR is a presence/absence, band=P. aeruginosa presence test

We have clarified this point.

Page 11: line 58: recombination, Darwinian selection...or could be misalignment of reads no?

That is a reasonable option and we have added this option to the text.

Page 14: line 36. “in burns unit” better as “in the burn unit"s”

We have made this change.

Page 15: line 24. Why is this Patient E when all other patients are numerical?

We have corrected this.

Page 15: line 26. two periods (..) at end of sentence

We have corrected this.

Page 16: line 8 (and throughout manuscript) in various places the word clade is capitalized and in other"s it"s not. You also refer to “clone E” --page 19 line 54--instead of clade E at another point. I would suggest modifying each case so the manuscript is consistent throughout

We thank the reviewer for this helpful point, we have standardised on “Clade X” throughout, with Clade capitalised.

Page 16: line 39-41. how did you know these were plasmids if you only had MiSeq paired ends? I doubt they would assemble completely, or did they?

Plasmids were detected through homology searching for plasmid-related genes such as replication initiators, and looking for contigs with abnormal coverage in Velvet assemblies. In some cases we found contigs which appeared to be complete contigs, in other strains they seemed to be fragmented. We have not performed any additional validation. The plasmids may be looked at in more detail in future studies.

Page 16: line 54 There aren‟t any published reports of natural transformation in P. aeruginosa as of 2010. Any as far as I know, nothing has been published on competence in P.aeruginosa since then (I could be wrong...) http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3008168/

We appreciate this point and as it was mentioned only by way of a general introduction to antibiotic resistance we feel comfortable removing this reference to natural transformability.

Figure legend 1: Maybe include a reference to the Stewart paper in legend (reference 27) as well as any other references needed in order to say where all other strains in tree came from.

We have added this reference, which should also suffice to provide references to the other strains.

Figure legend 2: Reading the legend and seeing the figure, I‟m still a bit unclear as to what the Y-axis means. Maybe reference appendix 5 in legend and rewrite to be more descriptive.

We have added detail here.

Reviewer Name Thomas Connor
Institution and Country Cardiff University, UK
Please state any competing interests or state „None declared‟: None declared

I enjoyed the paper very much, and think that it is both scientifically sound, and thought provoking, and think that it warrants publication. I do have a few minor questions/corrections that I would like to put to the authors. Some of these (1-3) relate to added detail that makes it easier for a reader to assess the strength of the conclusions drawn, while the others (4-6) are questions that are a little outside the scope of the current paper, but could be of interest to other bacterial genomics researchers.

We thank Dr. Connor for his kind comments on our paper and thank him for the suggestions made.

Questions

1. The authors state that the samples were mapped, and then recombinant regions were removed, but don't provide specific information on what the SNP scale on their trees refers to; I am interested to know what % mapping/core genome size you were left with for your isolates on a population wide level (i.e. the data used to generate figure 1A) when you came to draw your tree. Also, and related to this, I am curious how well the use of a variant density filter and the VarScan thresholds worked to remove phage and other mobile elements from the data for the population wide results. Also, how did you simulate the 600,000 reads for previously published strains?

Unfortunately we do not have the consensus VCF (i.e. non-variant positions) files to hand in order to compute the precise core genome size, and the limited time given to respond to reviewer‟s comments precluded us from generating one. However, by reference to the variant VCF files we found there were 234,879 variant positions in the core genome analysis against PAO1 used to generate Figure 1. This number was reduced to 218,314 after filtering for variant density and other filtering parameters. When we removed positions that were not called in all samples we had 96,246 positions remaining,

which was used to build the tree. Assuming variant positions are randomly distributed across the genome, we can infer that the core genome size is approximately 96246/234879*6300000 or 2.58 megabases.

Simulated paired-end 250 base reads were generated with wgsim (https://github.com/lh3/wgsim)

2. In terms of the fine-scale trees, I am would like to know a little more about your mapping results. Did your mapping approach lead to 'N's being called (i.e. the mapping software is unable to make a call) or did you only include SNPs where you had a definitive mapped base for every sample? If so, how many SNP sites did you discard? Also what software did you use to detect indels? did you use a standalone piece of software (eg GATK), or was this just done based on the BWA results followed by variant calling?

Our variant calling and filtering approach discards positions with no-calls. The only exception to this is the biofilm metagenomic phylogenetic placement technique, which includes uncalled positions as missing data due to the low coverage. We ensured that no samples had more than 20% no-calls to prevent poorly covered samples resulting in a drastic reduction in the size of the core genome. We use VarScan to detect both SNPs and small (split read/paired end) indels.

3. The two plasmids pBURNS 1 and 2 are covered only briefly, did the authors look at these in more detail? specifically, how large are they? are they similar to other plasmids that have been reported elsewhere? and did the authors examine the phylogeny of pBURNS 1, and was it consistent with the chromosomal phylogeny, or was there more diversity present than might be expected (which might tell you something about acquisitions of the plasmids in the environment)?

Because the paper is fairly rich in data analysis already we have not included any such analysis but plan to discuss the potential role of these plasmids in a future publication.

4. The collection includes a number of isolates from environmental sources, I wondered about their genomic content too; did you find any other plasmids (other than pBURNS1 and 2) in their assembled genomes, and did you look for (or find) evidence of historical recombination between isolates that had been isolated from the water supply?

pBURNS 1 and 2 are the only plasmids we detected and we would like to explore this interesting question in a future study.

5. Did the Pseudomonas colonise the same site as the Acinetobacter baumannii? and did you sequence the Acinetobacter strain? (and, if so, did you see any evidence of genetic exchange?)

This is a very interesting question. The patient had both P. aeruginosa and A. baumannii in their sputum (they were mechanically ventilated). We did not sequence the Acinetobacter strain in this study, but the movement of mobile elements between pathogenic organisms in a mixed infection is a fascinating question to study in the future.

6. Did the you try mapping the DNA in the biofilm to phage or other microbial species such as those from the archaea?

Analysis of the biofilm sample was hampered by low coverage of the organisms in there. Even the most abundant organism, P. aeruginosa, had such only 5x coverage. We would anticipate collecting more biofilm samples in future and conducting a full-scale metagenomics analysis on them.

Other minor points

Page 12 line 21/22 - does this mean that you only used positions that were present in all samples? is a little unclear.

We always call „core" SNPs, and because the size of the core genome is reduced when a more unrelated reference is used, we decided to map read against a finished representative of this clone.

Page 13, line 7; I think pubMLST like the citation at the end of this link ; http://www.biomedcentral.com/1471-2105/11/595/abstract

Thank you, we have added this reference to the text.

You may also / probably should include a short description of what MLST is (just a sentence), and probably a citation to the MLST paper (available at - http://www.pnas.org/cgi/pmidlookup?view=long&pmid=9501229)

Thank you, we have added a reference to MLST but have stopped short of describing it.

Page 14 line 2 -this is a repeat of what is on page 12/13 - "phylogenetic trees were generated using FastTree and FigTree" but I think this should read "phylogenetic trees were generated using FastTree and visualised using FigTree", although as it is redundant, you could probably just remove it.

Thank you, we have clarified.

We would like to thank all the reviewers for their positive reaction to the manuscript and their extremely helpful comments. We believe the manuscript is now greatly improved and we have added our personal thanks in the acknowledgements section.

VERSION 2 – REVIEW

REVIEWER	Jennifer Gardy BC Centre for Disease Control, Vancouver, Canada
REVIEW RETURNED	16-Sep-2014

GENERAL COMMENTS	This reviewer is more than happy with the authors' responses to her suggestions, and their responses to the thoughtful comments of the other reviewers, and she is looking forward to seeing this final version in print - well done!

REVIEWER	Thomas Connor Cardiff University School of Biosciences No competing interests (as per your definition) - for the avoidance of doubt, I currently collaborate with Pallen / Loman as part of the CLoud Infrastructure for Medical Bioinformatics consortium which is unrelated to the work here.
REVIEW RETURNED	26-Sep-2014

- The reviewer completed the checklist but made no further comments.