Reviewer’s report

Title: A landscape of genomic alterations at the root of a near-untreatable tuberculosis epidemic

Version: 0 Date: 10 Sep 2019

Reviewer: Guislaine Refrégier

Reviewer's report:

The manuscript by Klopper et al presents a genomic study on 150 Mycobacterium tuberculosis South African isolates of so-called "atypical" Beijing that are compared to other Beijing isolates to total 211 samples. Atypical Beijing has been previously linked to high resistance levels in South Africa.

The aim of the study was to provide hypotheses on the genetic background favoring resistance emergence, and therefore be able to provide guidelines for more efficient surveillance of TB. The method included the identification of mutations co-occurring with resistance mutations.

The study includes new data and interesting analyses. The discussion section connects isolate diversity to past drug regimens in South Africa, and suggests vigilance points for global management of tuberculosis treatment.

The scheme of the study is of interest to describe some parts of resistance acquisition dynamics among tuberculous isolates in South Africa. Altogether, this study provides interesting data on South African tuberculosis, and articulates promising hypotheses to be tested in other settings.

I however have some questions on the methods, and on the robustness and interpretation of the results.

Major comments:

1. Description of strains. The focus sample is described as a convenient sample, with no information on the patient treatment and outcome. However, the strains are described as "atypical", and containing only 7 "presumed drug-susceptible strains". Please explain what experiments were conducted to identify these characteristics, what strains were left out of the study, and what biases may have been introduced at the corresponding steps (collection, culture, targeted genotyping and/or sequencing, etc.).

2. Description of the methods. The DNA sequence analysis is said to have been conducted as in Black et al (2015) but then, SNP detection is described quite extensively. Please make clear what part is a summary from Black et al procedure and highlight thereafter potential differences, or state clearly that the description is just a reminder of what was previously described.
Regarding filtering of repetitive regions and resistance genes, it is contradictory to read in the phylogeny methods that "all known drug resistance variants [...] were removed" from the SNP matrix and to identify resistance variants in the table describing the mutations characterizing each clade (e.g. ethA A381P in Table 2, katG S315T in table 3.1, etc.). Please check that you performed the phylogenetic analysis independently of the resistance mutations, and clearly describe what you did. Further check possible reasons for the wide differences in terminal branch lengths which casts further doubts on the robustness of the results.

Regarding phylogeny, please make explicit what tree is made with how many samples (Fig1 and S3), with what filtering regarding resistance mutations.

3. Description of the phylogenetic results. No bootstrap values are shown on either of the trees. It is in contrast stated that some had low support in the ML analysis. Please show bootstrap supports. I would advise to name "clade" only sublineages that are well supported. For unsupported branches, the best would be to provide a tree with merged branches. This would help to have a clear view of reliable phylogenetic relationships.

4. Abstract precision. Methods should not claim that the comparative genomic analysis was done to determine changes that are unique to resistant strains: too few strains happen to be susceptible to pretend to determine changes that are present only in resistant strains. It seems also an overstatement to say that the ethA mutation "was instrumental in this resistance acquisition" as no control sample is available. It seems lastly contradictory to me to claim that "subsequent inadequate treatment led to amplification of resistance" (abstract) and that "treatment history and outcomes are unknown for all patients sampled" (methods).

5. Results precision. Among the variants defining AA1SA, consider deleting the mutations conferring resistance as the sampling may not be exhaustive and the existence of susceptible strains in this sublineage may not be fully ruled out.

Minor comments:

6. It is difficult to follow authors' description without S3 tree, so I would put it as a main figure. Clades would be more visible if presented with braces at the right end of the figure instead of as color-coded column.

7. Asia ancestral strains AA3, supposedly Asia Ancestral 3, are intermingled with typical isolates in the S3 figure phylogeny. Please explain how this can be. Moreover AA2 are announced but none is to be seen in the tree according to the color code. Please clarify AA2 presence/absence.

8. If kept in future versions of the manuscript, please highlight that clade D is not monophyletic or discuss its possible monophyly considering bootstrap values.
9. L188-190: transfer to discussion.

10. L193: "definitive" SNPs of AA1 should be listed.

11. L195-196: before stating that Cohen et al did not find a large number of AA1S1, it would be useful for the reader to state that in this study they were abundant not only in EC and WC but also in KZN.

12. SNP distance is not a clear parameter: are you describing pairwise distances or distance to the root? Please clarify.

13. I could not find mprB in Table 2 although described as an essential gene harboring a deleterious mutation. Please explain.

14. What is the "known AA1 progenitor"? What do you mean by "this progenitor was genetically closer to the known AA1 progenitor, or the current definition of AA1SA"?

15. I could not find any description of abbreviations in table S2, and only some of them can be found at other places. Please clarify all, such as AF.

16. Conclusion focus. It would be nice to focus on what is new. The fact that undetected resistance may lead to further resistance does not seem new to me. Also please reduce statements related to the discussion section.

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

No

**Does the work include the necessary controls?**
If not, please specify which controls are required in your comments to the authors.

Unable to assess

**Are the conclusions drawn adequately supported by the data shown?**
If not, please explain in your comments to the authors.

No
Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

Not relevant to this manuscript

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests.

Statement on potential review bias
Please complete a statement on potential review bias, considering the following questions:

1. Did you co-author any publication with an author of this manuscript in the last 5 years?

2. Are you currently or recently affiliated at the same institution as an author of this manuscript?

If you can answer no to all of the above, write 'I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria”. If your reply is yes to any, please give details below.
I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal.