nature portfolio

Peer Review File

Multivariable regression models improve accuracy and sensitive grading of antibiotic resistance mutations in Mycobacterium tuberculosis

Corresponding Author: Ms Sanjana Kulkarni

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Here, Kulkarni et. al. introduce a regression-based model for cataloguing resistance mutations in Mycobacterium tuberculosis which outperforms the existing WHO catalogue by identifying 99% of known variants as well as 29% more novel variants. This approach improves sensitivity while maintaining similar specificity, and detects additional mutations linked to drug resistance and hyper-susceptibility. This work is well-constructed, and they address potential limitations and issues in the data and analysis well, for example performing PCA to account for population structure. This work will be of significance in the field where novel mutations associated with AMR can be identified and graded with higher accuracy and can inspire further work to experimentally validate new associations.

I only have minor comments below:

Minors comment:

I understand the main purpose of this study is to compare to the WHO SOLO method and grade associations found using a phenotype/genotype regression but tools that are routinely used in TB genomic analysis to predict resistance in samples (e.g., TBprofiler (though Coll et. al. is mentioned for lineage calling) and Mykrobe) are not evaluated or introduced. Even if the results of these tools are a binary in silico prediction of resistant/susceptible, it would be helpful to mention these tools and whether the mutations identified here are found in these tools, particularly some of the novel variants.

Line 85 – Confusing wording as methods haven't been introduced yet so difficult to interpret why there are ranges in the following statement: "the remaining 560 to 48,236 isolates and 31 to 4,086 variants per drug were used to train regression models for 15 drugs".

Line 85 – Beginning here, lots of abbreviations are used before they are defined (e.g., BQ, BQ-S, pDST, R-associated, LoF). I see that these are in table 1 so maybe add a sentence stating that is where abbreviations are detailed before using them. Also, there is no background on the 'ALL' dataset before it is introduced as it is in the methods.

Line 111 – perhaps put the gradings (e.g., Assoc w R, Uncertain, etc) in italics or quote marks to show these refer to the gradings.

Line 544 – typo "Fig.s"

(Remarks on code availability)
The code is well documented on the supplied GitHub page to allow reproducibility of the analysis.

Reviewer #2

(Remarks to the Author)

Summary/ noteworthy results
The authors utilise a multivariate logistic regression approach using a large dataset comprised of M. tuberculosis samples to build a catalogue of drug-resistance associated mutations. Major strengths of this analysis include the large sample size, large-scale statistical analysis and methods for establishing the significance of model features. The catalogue is benchmarked against the existing World Health Organization catalogue that was developed using the SOLO method and demonstrates improvement for profiling drug-resistance for all drugs, but only by a large amount for some of the drugs. The authors address an existing limitation of the current WHO catalogue in that its development has relied upon univariate association and does not address confounding from population structure using the multivariate approach, but the overall outcomes do not differ drastically from the original catalogue, with a few potential hits in compensatory mutations and other regions. The code associated with the manuscript is detailed and appears reproducible, but cond environment could not be installed on the server. Please see major and minor comments below.

Major comments:
1. The authors state that they exclude Tier 2 genes from their analysis because of lack of previous association. The authors should comment on whether they investigated whether the multivariate logistic regression approach could have been used to identify mutations in Tier 2 genes that may have not been detected previously due to univariate association (line 378).
2. Additional clarity on the features used within each model should be provided (line 397). Were non-silent mutations pooled at any point, why were they 'unpooled'? Additionally, whilst it is agreed that silent mutations are unlikely to confer a larger effect, there are some cases where they have been identified as potential drug-resistance mutations. What did the authors do in the case of rrs for example? It would be useful to reference a supplementary file with the mutations used in each model.
3. The authors have used a systematic approach to look at the number of times a mutation occurs in each lineage and comment on their rarity. How did authors deal with such results, the authors should comment on whether they be downgraded due to this (although lineage-specific drug-resistance mutations are known to occur). The authors also mention in the discussion that principal components were used for correction, previous genome-wide association studies have used 5-10 principal components. How was a threshold of 50 decided in this case and how did this compare to models with no correction? This is an interesting approach to potentially limit confounding in machine learning analysis.
4. The concordance of variants identified using the binary and MIC regression approaches is 34%: "233 variants have a significant regression OR in the WHO dataset, 214 and of these, 79 (34%) are also significantly associated with MIC (73 with resistance, 6 with 215 susceptibility), with all significant associations reproducing the direction of effect measured in 216 the binary WHO models.". If the interpretation of this is correct, would the authors expect this number to be higher and could they provide a further explanation of what might be limiting this. Have the authors provided a supplementary file they could reference that details these specific variants in addition to Figure 4?
5. In the abstract, the authors suggest an average improvement of 3.2 pp for sensitivity. However, the improvement of performance has a large range (0.081-10.2 pp). This should be more accurately represented in the abstract as the improved performance by larger amounts is for specific drugs as opposed to the entire panel. They also suggest INH and RIF do not meet the target product profile required by WHO in terms of specificity (line 280) . Therefore, the conclusion in the abstract regarding assays should be adjusted accordingly, perhaps highlighting further validation is required. Also, the final sentence in the abstract suggests the logistic regression models themselves are used as the predictors and not necessarily the mutation catalogue developed so this could be reworded to make this more clear as performance of the model themselves has not been evaluated.
6. The authors discuss the potential for ahpC mutations as having a compensatory mechanism but they have also detected mutations outside of the RDDR in rpoB that contain another drug-resistance mutation. These could also be potential compensatory mutations and could be compared to existing lists within the literature. On a similar note, WHO previously excluded compensatory mutations from the mutation catalogue. Does this support that the inclusion of compensatory mutations may improve classification?

Minor comments:
1. Mycobacterium tuberculosis should be italicised (line 78).
2. The WHO mutation catalogue used should be included as a supplementary file for ease of reference and comparison to readers (line 361).
3. Format of genotypic and phenotypic data should be described i.e are phenotypes binary?
4. There seems to be a typo on line 386, there should be a closed ')' instead of ']'. The same typo occurs in the table legend for supplementary data 1.
5. Authors should state how thresholds used to determine significance were decided (line 472), a table with significance thresholds would also be useful or could be included in figure 1.
6. There is a typo on line 89 'bedaquiline(BDQ)'
7. There is a typo on line 175 '(0.01, 0.05]'.
8. rpoB should be italicised on line 226
9. Line 237-240 requires a reference
10. Does line 310 need an embedded reference?

(Remarks on code availability)
The code has a README file detailing how to execute and installation. It appears detailed and reproducible. Could not install conda environment due to strict priorities which could be made flexible.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed my comments well and I now believe the manuscript should be accepted.

(Remarks on code availability)
The code appears well documented and I was able to install and run without issue.

Reviewer #2

(Remarks to the Author)
All my original comments have been addressed by the authors. I have no additional comments for the authors. Thank you.

(Remarks on code availability)
The code is reproducible and provided in a well-guided repository.

Manuscript NCOMMS-24-37063A

Response to Reviewers

Reviewer #1 (Remarks to the Author):

Here, Kulkarni et. al. introduce a regression-based model for cataloguing resistance mutations in Mycobacterium tuberculosis which outperforms the existing WHO catalogue by identifying 99% of known variants as well as 29% more novel variants. This approach improves sensitivity while maintaining similar specificity, and detects additional mutations linked to drug resistance and hyper-susceptibility. This work is well-constructed, and they address potential limitations and issues in the data and analysis well, for example performing PCA to account for population structure. This work will be of significance in the field where novel mutations associated with AMR can be identified and graded with higher accuracy and can inspire further work to experimentally validate new associations.

I only have minor comments below:

Minors comment:

I understand the main purpose of this study is to compare to the WHO SOLO method and grade associations found using a phenotype/genotype regression but tools that are routinely used in TB genomic analysis to predict resistance in samples (e.g., TBprofiler (though Coll et. al. is mentioned for lineage calling) and Mykrobe) are not evaluated or introduced. Even if the results of these tools are a binary in silico prediction of resistant/susceptible, it would be helpful to mention these tools and whether the mutations identified here are found in these tools, particularly some of the novel variants.

We have added references to TBprofiler and Mykrobe in the introduction. Both tools make predictions using a defined catalogue of resistance mutations, typically the World Health Organization's published catalogue. A comparison to TBprofiler and Mykrobe would be identical to the comparison to the published catalogue, other than the bioinformatics/variant calling, which is not the focus of this work.

Line 85 – Confusing wording as methods haven't been introduced yet so difficult to interpret why there are ranges in the following statement: "the remaining 560 to 48,236 isolates and 31 to 4,086 variants per drug were used to train regression models for 15 drugs".

We have removed the isolate and variant counts so that the wording is less confusing. This is now line 92 in the revised manuscript.

Line 85 – Beginning here, lots of abbreviations are used before they are defined (e.g., BQ, BQ-

S, pDST, R-associated, LoF). I see that these are in table 1 so maybe add a sentence stating that is where abbreviations are detailed before using them. Also, there is no background on the 'ALL' dataset before it is introduced as it is in the methods.

We have added a line referencing the abbreviations in Table 1 and added additional in-text full length words before the abbreviations. We also added an explanation of the "ALL dataset" in the Results section.

Line 111 – perhaps put the gradings (e.g., Assoc w R, Uncertain, etc) in italics or quote marks to show these refer to the gradings.

We have added quotations throughout the text for the gradings.

Line 544 – typo "Fig.s"

Fixed. This is now line 591 in the revised manuscript.

Reviewer #1 (Remarks on code availability):

The code is well documented on the supplied GitHub page to allow reproducibility of the analysis.

Reviewer #2 (Remarks to the Author):

Summary/ noteworthy results
The authors utilise a multivariate logistic regression approach using a large dataset comprised of M. tuberculosis samples to build a catalogue of drug-resistance associated mutations. Major strengths of this analysis include the large sample size, large-scale statistical analysis and methods for establishing the significance of model features. The catalogue is benchmarked against the existing World Health Organization catalogue that was developed using the SOLO method and demonstrates improvement for profiling drug-resistance for all drugs, but only by a large amount for some of the drugs. The authors address an existing limitation of the current WHO catalogue in that its development has relied upon univariate association and does not address confounding from population structure using the multivariate approach, but the overall outcomes do not differ drastically from the original catalogue, with a few potential hits in compensatory mutations and other regions. The code associated with the manuscript is detailed and appears reproducible, but cond environment could not be installed on the server. Please see major and minor comments below.

Major comments:
1. The authors state that they exclude Tier 2 genes from their analysis because of lack of previous association. The authors should comment on whether they investigated whether the

multivariate logistic regression approach could have been used to identify mutations in Tier 2 genes that may have not been detected previously due to univariate association (line 378).

Yes, regression would allow for testing for association between any mutation and the resistance phenotype provided the variant occurs at sufficient frequency and its correlation with existing resistance mutations isn't too high. Our open-source code currently allows for the addition of tier 2 mutations to tier 1 mutations in the association, but we have not tested if it associates tier 2 mutations as there is no accepted benchmark for these given that the 2nd edition of the WHO mutation catalogue graded all variants in Tier 2 genes as "Uncertain." In response to the reviewer's comment, we have now added this to lines 420-424:

"We focused the regression on associating variants in Tier 1 genes **(Supplementary Table 1)** because evidence from the 2nd edition of the mutation catalogue did not support associations between drug resistance and variants in Tier 2 genes[8]. However, regression models can be fit on variants in both Tier 1 and 2 genes by adding an additional flag to the model scripts."

We also have the following text in the discussion section in lines 333-336:

"Although the primary goal of this work was to benchmark the regression catalogue against the published catalogue, which only identified associations with variants in genes known or highly suspected to be relevant[8], regression can be used to associate variants in additional loci to discover novel low-level contributors to resistance."

2. Additional clarity on the features used within each model should be provided (line 397). Were non-silent mutations pooled at any point, why were they 'unpooled'? Additionally, whilst it is agreed that silent mutations are unlikely to confer a larger effect, there are some cases where they have been identified as potential drug-resistance mutations. What did the authors do in the case of rrs for example? It would be useful to reference a supplementary file with the mutations used in each model.

To clarify, the models tested each individual non-silent variant (unpooled) because we were interested in measuring the association between each variant and the resistance phenotype. We only pooled putative loss-of-function (LoF) mutations (start lost, stop gained, large deletion, or frameshift mutations according to the WHO catalogue) as is currently done in the WHO v2 catalogue. This pooling allows the measurement of a combined/average effect for several rare mutations with expected large effects on protein function and consequently, drug resistance.

The rarity of these variants limits the statistical power for measuring their individual effect on resistance and justifies their pooling into a single combined LoF variable for each gene in both regression pipeline and the published catalogue method. Given their expected large effect on protein function, pooling them helped preserve statistical power akin to a burden test in human genetics. Variants in the *rrs* and *rrl* genes were not pooled as they do not code for protein and LoF cannot be defined in the same manner as for protein coding genes.

We did not pool silent variants. The latter were tested in secondary models with a higher FDR threshold (0.01) than the primary models that tested only non-silent variants with or without pooling LoF variants.

The files of the results in individual models are available in the Github repository at https://github.com/farhat-lab/who-analysis/tree/main/results/BINARY. Each drug has an Excel file, and each file has 6 sheets corresponding to the 6 binary models fit for each drug (3 pooling types x 2 phenotypic datasets). The MIC models are at https://github.com/farhat-lab/who-analysis/tree/main/results/MIC, and each file has 3 sheets corresponding to the 3 types of variant inclusions – no pooling, pooled LoF, and with silent variants.

3. The authors have used a systematic approach to look at the number of times a mutation occurs in each lineage and comment on their rarity. How did authors deal with such results, the authors should comment on whether they be downgraded due to this (although lineage-specific drug-resistance mutations are known to occur). The authors also mention in the discussion that principal components were used for correction, previous genome-wide association studies have used 5-10 principal components. How was a threshold of 50 decided in this case and how did this compare to models with no correction? This is an interesting approach to potentially limit confounding in machine learning analysis.

We thank the reviewer for this comment. We examine the lineage distribution of mutations to validate the grading, but we do not alter the grading based on whether mutations were lineage-restricted. This is because benchmarking for the use of lineage distributions or homoplasy for grading mutations is not well developed. In other recent and yet unpublished work by the WHO TB drug resistance catalogue scientific committee, the use of lineage distribution preliminarily does not change the SOLO grading of mutations significantly as most mutations graded as resistant are either homoplastic (i.e. occur in different mutations) or are very rare rendering the homoplasy assessment less reliable. We anticipate these results to be published separately as part of the WHO drug resistance catalogue version 3 and are currently outside of the scope of the regression effort. We have added the following text to lines 271-275:

“Changing the grading of a mutation based on its lineage distribution has not been benchmarked because most resistance-associated mutations are homoplastic, occurring independently in multiple lineages. The use of lineage distribution to inform gradings is currently being investigated for the 3rd edition of the WHO resistance mutation catalogue.”

We thank the reviewer for highlighting the novelty of prediction approach using principal components. With regards to the number of principal components (PCs) used in the regression model, we chose 50 PCs to be conservative and correct for as much genetic variation outside of the known or suspected drug resistance genes as possible. We have now added a plot of the proportion of variance explained (pve) per PC to **Supplementary Figure 2a** that demonstrates a decreasing pve with an early elbow point around PC ~16 but with a tail of higher PVE for PCs 20-50. At PC50, 95% of the genetic variance was cumulatively explained by PC1-PC50.

4. The concordance of variants identified using the binary and MIC regression approaches is 34%: “233 variants have a significant regression OR in the WHO dataset, and of these, 79 (34%) are also significantly associated with MIC (73 with resistance, 6 with susceptibility), with all significant associations reproducing the direction of effect measured in the binary WHO models.”. If the interpretation of this is correct, would the authors expect this number to be higher and could they provide a further explanation of what might be limiting this. Have the authors provided a supplementary file they could reference that details these specific variants in addition to Figure 4?

Because of the bug fix to the principal components analysis, the new text reads as follows, in lines 235-238 in the revised manuscript:

“232 of these variants have a significant regression OR in the WHO dataset, and of these, 83 (36%) are also significantly associated with MIC (76 with resistance, 7 susceptibility), with all significant associations reproducing the direction of effect measured in the binary WHO models.”

The number is low because the MIC dataset was smaller and power was more limited for associating many of the mutations were associated in the binary regression on the WHO dataset; 52% (78) of the 149 mutations associated in the binary WHO dataset regression but not associated in the MIC models occurred in fewer than 5 isolates in the MIC dataset. We identified an additional 9 association in the MIC dataset at a less conservative FDR threshold than we report in the paper (FDR<0.1 vs 0.05) and their direction of association was also congruent with that measured in the binary regression adding further validity to our results.

We have added the following similar text describing these counts to lines 238-243 in the revised manuscript:

“Of the 232 – 83 = 149 remaining variants associated in the binary WHO dataset regression but not associated in the MIC models, 78 (52%) occurred in fewer than 5 isolates in the MIC datasets, which could explain the inability to detect these associations. An additional 9 variants are significantly associated with MIC at a less conservative FDR threshold of 0.1, and their directions of association are also congruent with those measured in the binary regression.”

The variants with MIC results available and with significant odds ratios in the WHO dataset have now been added to sheet 2 of **Supplementary Data 5**.

5. In the abstract, the authors suggest an average improvement of 3.2 pp for sensitivity. However, the improvement of performance has a large range (0.081-10.2 pp). This should be more accurately represented in the abstract as the improved performance by larger amounts is for specific drugs as opposed to the entire panel. They also suggest INH and RIF do not meet the target product profile required by WHO in terms of specificity (line 280). Therefore, the conclusion in the abstract regarding assays should be adjusted accordingly, perhaps highlighting further validation is required. Also, the final sentence in the abstract suggests the logistic regression models themselves are used as the predictors and not necessarily the

mutation catalogue developed so this could be reworded to make this more clear as performance of the model themselves has not been evaluated.

We have removed the reference to logistic regression models for prediction as it is unclear. We have clarified that the average sensitivity gain is due to a subset of the drugs. The updated abstract with additions highlighted is copied below from lines 29-35:

"Regression detects 450/457 (98%) resistance-associated variants identified using the existing method (*a.k.a,* SOLO method) and grades 221 (29%) more total variants than SOLO. The regression-based catalogue achieves higher sensitivity on average (+3.2 percentage points, pp) than SOLO with smaller average decreases in specificity (-1.0 pp) and positive predictive value (-1.7 pp). Sensitivity gains are highest for ethambutol, clofazimine, streptomycin, and ethionamide as regression grades considerably more resistance-associated variants than SOLO for these drugs."

6. The authors discuss the potential for ahpC mutations as having a compensatory mechanism but they have also detected mutations outside of the RDDR in rpoB that contain another drug-resistance mutation. These could also be potential compensatory mutations and could be compared to existing lists within the literature. On a similar note, WHO previously excluded compensatory mutations from the mutation catalogue. Does this support that the inclusion of compensatory mutations may improve classification?

There is precedent for the use of ahpC compensatory mutations for diagnostic purposes even when they do not directly encode resistance. AhpC mutations were used in several commercial diagnostics including the Xpert MTB/XDR assay for detecting INH resistance. Some compensatory mutations are more frequent than resistance causing mutations and when highly associated with resistance can serve as excellent predictors for resistance even when not causative. The WHO SOLO method used for grading mutations in the WHO catalogue does not explicitly exclude compensatory mutations, but by design it is limited in detecting compensatory mutations due to the requirement for mutations occur alone (without other known resistance mutations) in a resistant isolate. We agree with the reviewer that it is possible that some of the mutations outside of the RRDR in rpoB are compensatory and not causative. As we list in the limitations section, regression cannot confirm causation but it is very useful for association with the resistance phenotype. We computed binary classification metrics with the removal of the 6 ahpC compensatory mutations that regression associates with isoniazid resistance, finding that it reduces sensitivity by 0.5 percentage points (pp) and increases specificity by 0.1 pp. This data has been added to Supplementary Data 8. We have added text to the discussion section on the potential usefulness of compensatory mutations for improving resistance classification, copied below from lines 354-358.

“Our results support the use of the six associated *ahpC* mutations **(Table 3)** to classify resistance because their inclusion changes INH resistance prediction sensitivity by +0.5 percentage points (pp) and specificity by -0.1 pp **(Supplementary Data 8)**. Further research is

needed to assess if multivariate models can offer more accuracy when combining resistance causing and compensatory mutations in diagnostics use."

Minor comments:
1. Mycobacterium tuberculosis should be italicised (line 78).

Fixed. This is now line 86 in the revised manuscript.

2. The WHO mutation catalogue used should be included as a supplementary file for ease of reference and comparison to readers (line 361).

It has been added as Supplementary Data 11.

3. Format of genotypic and phenotypic data should be described i.e are phenotypes binary?

Yes, the phenotypes are binary, except for the MICs. We have added text to make this clearer.

4. There seems to be a typo on line 386, there should be a closed ')' instead of ']'. The same typo occurs in the table legend for supplementary data 1.

This is actually not a typo. The left bound is exclusive, and the right bound is inclusive. Isolates with a within-sample allele frequency of 0.25 are considered to not have the allele, and isolates with a within-sample allele frequency of 0.75 are considered missing with respect to that allele. We have added this clarification to the legend for Supplementary Data 1. This is now line 430 in the revised manuscript. We also clarified this in line 95 of the revised manuscript.

5. Authors should state how thresholds used to determine significance were decided (line 472), a table with significance thresholds would also be useful or could be included in figure 1.

We have added a clarification to the legend for Figure 1. A description of the significance thresholds is in Methods Section F in lines 516-520.

6. There is a typo on line 89 'bedaquiline(BDQ)'

Fixed. This is now line 105 in the revised manuscript.

7. There is a typo on line 175 '(0.01, 0.05]'.

This is actually not a typo. The left bound is exclusive, and the right bound is inclusive. In the legend for Figure 1, we state that we use inclusive bounds for determining significance. Non-silent variants with FDR p-value ≤ 0.05 and silent variants with FDR p-value ≤ 0.01 are significant. We have reworded lines 195-197 to the following for clarity:

“Seven silent variants have OR < 1 with an FDR p-value greater than 0.01 but less than or equal to 0.05, but they are graded “Uncertain” because we apply a stricter cut-off of 0.01 for silent variants.”

8. rpoB should be italicised on line 226

Fixed. This is now in line 252 in the revised manuscript.

9. Line 237-240 requires a reference

Added. This text has been reworded and moved to the discussion at lines 345-348.

10. Does line 310 need an embedded reference?

We have added a reference to Results Section D. This is now line 341 in the revised manuscript.