Predicting antibiotic resistance in *Mycobacterium tuberculosis* with genomic machine learning

Chang Ho Yoon MD[1], Anna G. Green PhD[1], Michael L. Chen[1], Luca Freschi PhD[1], Isaac Kohane MD PhD[1], Andrew Beam PhD[1,2], Maha Farhat MD MSc[1,3]

[1] Department of Biomedical Informatics, Harvard Medical School [2]Department of Epidemiology, Harvard T.H. Chan School of Public Health [3]Division of Pulmonary & Critical Care, Massachusetts General Hospital

Background. Tuberculosis (TB) is the foremost cause of death worldwide from a single infectious pathogen, with over 1.5 million people succumbing annually to the disease.(1) While the global incidence of TB is declining, the proportion of multidrug-resistant tuberculosis (MDR-TB) is increasing, posing an urgent threat to public health.(2) Mortality is significantly higher in MDR-TB and XDR-TB (extensively drug-resistant TB): ca. 48% and 72%, respectively.(1) These cases experience a long lag in the initiation of appropriate treatment due to TB's slow growth in culture required for traditional antimicrobial drug susceptibility testing. Established molecular techniques for detecting MDR also lack sensitivity, accurately predicting resistance only to 2 of more than 10 possible drugs.(3) Recently, our group deployed a wide and deep neural network (WDNN) to predict antimicrobial resistance in TB using curated variants, gene-gene interactions and indeterminate variants in known resistance loci.(4) Rapid machine-learning-based diagnostics will likely lead to faster initiation of appropriate treatment for MDR-TB, reducing morbidity and mortality, and improving health economic endpoints.(1, 5)

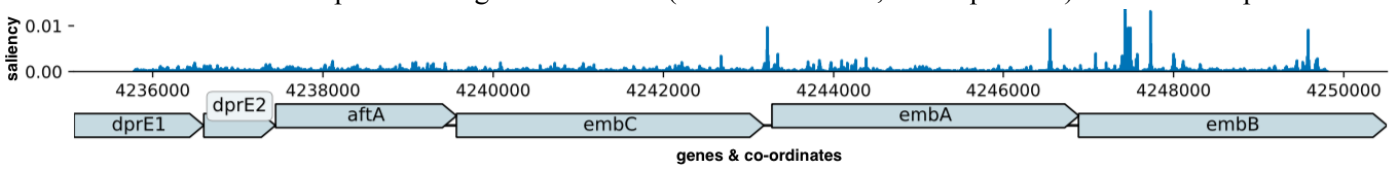
Hypothesis. We hypothesize that a multi-task convolutional neural network is well-suited to the prediction of antimicrobial resistance in TB. Because a CNN incorporates information about genetic proximity, we expect that it will perform at least as well as our group's previous WDNN, while at the same time achieving increased interpretability through saliency maps. We expect this model to have implications for trust (in diagnosis) and hypothesis generation about molecular mechanisms of antimicrobial resistance.

Methods. Using genotypic and phenotypic data from 10,204 MTB isolates, we investigated the predictive performance of a multi-task CNN, designed to directly analyse the genetic sequences and discern motifs contributing to antimicrobial resistance. The isolates' phenotype comprises resistance, susceptibility, or unavailable information with respect to 13 anti-TB drugs. We trained the CNN to predict resistance or susceptibility to these drugs for a given isolate, based on 15 genetic loci (each one having multiple genes and flanking regions) known to confer resistance to at least one of the anti-TB drugs. Gradient-based saliency maps highlighted the genetic regions most pertinent to the model's predictions.

Results. Table: TB drug resistance prediction AUC (area under the receiver operating characteristic) performance of the comparative models using 5-fold cross-validation (mean AUC of 10 iterations ± 95% confidence interval)

	Mean AUC for a given model & drug (±95% confidence interval)	
Drug	**WDNN**	**Multi-task CNN**
Rifampicin	0.980 (±0.0007)	0.974 (±0.0005)
Isoniazid	0.974 (±0.0006)	0.969 (±0.0006)
Pyrazinamide	0.948 (±0.0017)	0.925 (±0.0019)
Ethambutol	0.937 (±0.0016)	0.930 (±0.0013)
Streptomycin	0.937 (±0.0027)	0.925 (±0.0013)
Capreomycin	0.894 (±0.0043)	0.886 (±0.0029)
Kanamycin	0.942 (±0.0034)	0.935 (±0.0017)
Amikacin	0.908 (±0.0039)	0.904 (±0.0036)
Ciprofloxacin	0.992 (±0.0030)	0.991 (±0.0053)
Ofloxacin	0.908 (±0.0035)	0.902 (±0.0039)
Moxifloxacin	0.883 (±0.0097)	0.879 (±0.0025)
Levofloxacin	Not performed	0.941 (±0.0114)
Ethionamide	Not performed	0.829 (±0.0047)

Figure: An example saliency map of TB isolates for ethambutol and the *dprE1-embB* intergenic region. The saliency values reflect the relative importance of genetic variants (some established, others putative) in the CNN's predictions:



Conclusion. The multi-task CNN has comparable predictive performance to our previously-reported WDNN with multiple advantages that will render it more clinically applicable: saliency maps provide insight into which genetic regions are contributing to its predictions, thereby promoting greater interpretability and trust, and enabling hypotheses about putative molecular mechanisms; by directly analysing genetic sequences, the CNN requires significantly less pre-processing of input data and can consider regions neighbouring resistance-associated genes. Our model represents a scalable, comprehensive diagnostic tool that holds potential in translating sequencing technology to the clinic.

References.

1. Organization. WH. Global tuberculosis report 2018.; 2018.
2. Lange C, Chesov D, Heyckendorf J, Leung CC, Udwadia Z, Dheda K. Drug-resistant tuberculosis: An update on disease burden, diagnosis and treatment. Respirology. 2018;23(7):656-73.
3. Farhat MR, Sultana R, Iartchouk O, Bozeman S, Galagan J, Sisk P, et al. Genetic Determinants of Drug Resistance in Mycobacterium tuberculosis and Their Diagnostic Value. Am J Respir Crit Care Med. 2016;194(5):621-30.
4. Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, et al. Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. EBioMedicine. 2019;43:356-69.
5. Chen Y, Yuan Z, Shen X, Wu J, Wu Z, Xu B. Time to Multidrug-Resistant Tuberculosis Treatment Initiation in Association with Treatment Outcomes in Shanghai, China. Antimicrobial Agents and Chemotherapy. 2018;62(4):e02259-17.