

## Commentary

## Realising the Potential of Genomics for *M. tuberculosis*: A Silver Lining to the Pandemic?

Timothy M Walker<sup>1</sup>; Derrick W Crook<sup>2</sup>

The impact of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic on tuberculosis has been profound. Globally, there were 1.4 million fewer patients receiving tuberculosis treatment in 2020, representing not a fall in the global case load but a failure to serve millions of patients with the disease (1–2). Morbidity, mortality, and onward transmission are inevitable outcomes of such failure, setting the world back a decade of progress.

Even before the pandemic, there was a severe diagnostic deficit with only 61% of patients with bacteriologically confirmed tuberculosis tested for rifampicin resistance, let alone other drugs relevant to their disease (3). Culture-based susceptibility testing is too slow, expensive, and demanding of infrastructure and skills to provide a global solution, not least because the global burden of disease is predominantly found in low- and middle-income countries. Interim molecular solutions have come in cartridge form, but even the widely available MTB/RIF Xpert<sup>®</sup> and its descendants cannot provide full drug susceptibility testing results that clinicians would optimally want when designing personalised regimens for their patients.

Attention has therefore rightly turned to the potential of genomic solutions. Whole genome sequencing (WGS) has already been adopted as a routine test in a number of high-income settings such as the UK (4). It offers not only genome-based drug susceptibility testing for many drugs where we understand the molecular mechanisms of resistance, but also vital information on transmission to inform public health interventions. A remaining disadvantage of this technology is that good quality sequence data generally still requires DNA extraction from culture. Targeted Next Generation Sequencing (tNGS) offers a solution here by amplifying a fraction of the genome directly from the clinical sample and sequencing that to great depth. Clinically useful coverage of the “resistome” is thus generated, but at the cost of losing the potential for identifying genomic relatedness, inferring transmission, and informing public health action focussed on outbreaks.

All molecular assays, whether targeted or not, require interpretation. Whilst there are some canonical mutations conferring resistance to key drugs, there are many more mutations with less supporting data (5). The challenge for targeted molecular assays based purely around polymerase chain reaction (PCR) technology has been to identify a small number of high-value mutations restricted to the short target DNA sequence available to the assay. As many mutations outside the target sequence are therefore missed, results need to be interpreted with caution and with regard to the circumstances (6). The challenge for sequencing solutions continues to be selecting the infrequent, often more poorly supported, but collectively important mutations to include in an interpretative catalogue of resistance mutations. Catalogue-based approaches to interpretation gained a boost in 2021 after the World Health Organization (WHO) formally endorsed a catalogue of mutations associated with resistance in *M. tuberculosis* in 2021 (7). Although this catalogue may not provide more accurate results than other existing lists of mutations, it sets a standard in a field where until now none has been universally worked out.

The challenge with catalogue-based interpretation is that for most anti-tuberculosis drugs a small number of common mutations account for most global resistance, after which a potentially much larger number of rare mutations might collectively explain the remaining 5%–10% of drug resistance. Even in the dataset approaching 40,000 *Mycobacterium tuberculosis* (*M. tuberculosis*) isolates from which the WHO catalogue was generated, many of these mutations were seen too infrequently to confidently grade. Moreover, as some of these mutations may have only an incremental impact on the minimum inhibitory concentration (MIC), the resulting binary phenotype against which they are assessed is so inconsistent that just accumulating ever more data is unlikely to solve the problem. Recognizing that mutations do not all act in isolation but that some combine to have additive,

synergistic, or epistatic effects is part of the solution (8).

A number of machine learning techniques have been applied to the problem of drug susceptibility testing in *M. tuberculosis*. These avoid the need for a catalogue but do rely on training datasets. They have the potential to exploit features that have no mechanistic role in drug resistance, such as common drug resistance patterns (for example isoniazid resistance being common in the context of rifampicin resistance). The performance of some machine learning models has been excellent, and open-source tools exist for others to implement (9–10).

Although some machine learning approaches avoid the need for a bioinformatics pipeline, instead of operating more or less directly from the raw sequence data (11), a pipeline that maps sequencing reads to a reference genome and filters for variants remains a prerequisite for most of the above approaches. Again, although many different tools are used across academia, health service providers, and industry (12), no single global standard exists even though attempts have been made to outline what it may look like (13). Moreover, were opinions to coalesce around any one of the existing solutions, none would likely be capable of scaling to process tens of thousands of genomes a day to serve potential global demand. In the absence of a standardised solution, data comparability and data sharing are more burdensome, and progress is stalled.

The SARS-CoV-2 pandemic has led to global interest in genome sequencing as it has been the tool to identify the emergence of new variants and track their transmission locally, regionally, and globally (14). Huge efforts have gone into this work in countries around the world, with scientists facing many of the same problems described here: a need for bioinformatics pipelines; a need for global comparability and data sharing; and a need for scalability as samples are sequenced in their many thousands (15). The challenge has triggered a large philanthropic donation of cloud computing services and software engineering support from the tech giant ORACLE to the University of Oxford (16). The resulting Global Pathogen Analysis System (GPAS) delivered through the University of Oxford has been built in the first instance for an automated “turn-key” analysis service for SARS-CoV-2 genome sequence data agnostic to nucleic acid preparative methods and sequencing technology. Work is already in progress for an automated *M. tuberculosis* bioinformatics solution.

With the support of up to 100 ORACLE and University bioinformaticians and software engineers, GPAS is now a cloud based, industry standard, secure solution that can process around 1 million SARS-CoV-2 genomes per day. The solution provides the user with sovereign control of their own data, but also with the opportunity to share it with the global community (this is not a vehicle for either ORACLE or the University of Oxford to access data for themselves). GPAS will be accessible from anywhere in the world, by anyone with an internet connection, and will be a sequencing platform and able to process both WGS and tNGS data with the ability to accommodate bespoke primer sets. For users who wish to share data, options to upload data to existing global repositories will be offered. Drug susceptibility predictions will be catalogue based, but machine learning approaches are a future possibility. For WGS data, a user-friendly graphical interface will summarize the genomic relatedness to all other samples that have been globally shared to date.

A 10-year agreement with ORACLE provides GPAS free of charge for low- and middle-income countries, with services provided on a non-for-profit basis for high-income countries. GPAS is available for SARS-CoV-2 with a plan to release a parallel solution for *M. tuberculosis* in the third quarter of 2022, and thereafter for other microbial pathogens of importance to public health guided by a user group advising the development group led by the University of Oxford.

Amongst the many losses the fight against tuberculosis has suffered due to the SARS-CoV-2 pandemic, the promise of a secure, industry standard, scalable pipeline solution for *M. tuberculosis* diagnosis and control, accessible by anyone around the world at low or no cost for the global good is a definite silver lining. It opens up the potential for WGS or tNGS technology to be adopted more widely where it has hitherto been held back by the absence of sequencing analytic capability. Many challenges remain, but this unprecedented opportunity for data sharing in the interest of public health may play a role in the solution for some.

doi: 10.46234/ccdcw2022.063

<sup>1</sup> Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam; <sup>2</sup> Nuffield Department of Medicine University of Oxford, Oxford, UK.

Submitted: March 21, 2022; Accepted: March 23, 2022

## REFERENCES

- World Health Organization. Impact of the COVID-19 pandemic on TB detection and mortality in 2020. Geneva: World Health Organization; 2021. <https://www.who.int/publications/m/item/impact-of-the-covid-19-pandemic-on-tb-detection-and-mortality-in-2020>.
- World Health Organization. Global tuberculosis report 2021. Geneva: World Health Organization. 2021. <https://apps.who.int/iris/handle/10665/346387>.
- World Health Organization. Global tuberculosis report 2020. Geneva: World Health Organization; 2020. <https://www.who.int/publications/item/9789240013131>.
- The CRyPTIC Consortium and the 100,000 Genomes Project. Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing. *N Engl J Med* 2018;379(15):1403 – 15. <http://dx.doi.org/10.1056/NEJMoa1800474>.
- Miotto P, Tessema B, Tagliani E, Chindelevitch L, Starks AM, Emerson C, et al. A standardised method for interpreting the association between mutations and phenotypic drug resistance in *Mycobacterium tuberculosis*. *Eur Respir J* 2017;50(6):1701354. <http://dx.doi.org/10.1183/13993003.01354-2017>.
- Sanchez-Padilla E, Merker M, Beckert P, Jochims F, Dlamini T, Kahn P, et al. Detection of drug-resistant tuberculosis by Xpert MTB/RIF in Swaziland. *N Engl J Med* 2015;372(12):1181 – 2. <http://dx.doi.org/10.1056/NEJMc1413930>.
- World Health Organization. Catalogue of mutations in *Mycobacterium tuberculosis* complex and their association with drug resistance. Geneva: World Health Organization. 2021. <https://apps.who.int/iris/handle/10665/341981>.
- Vargas R Jr, Freschi L, Spitaleri A, Tahseen S, Barilar I, Niemann S, et al. Role of epistasis in amikacin, kanamycin, bedaquiline, and clofazimine resistance in *Mycobacterium tuberculosis* complex. *Antimicrob Agents Chemother* 2021;65(11):e0116421. <http://dx.doi.org/10.1128/AAC.01164-21>.
- Gröschel MI, Owens M, Freschi L, Vargas R Jr, Marin MG, Phelan J, et al. GenTB: a user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. *Genome Med* 2021;13(1):138. <http://dx.doi.org/10.1186/s13073-021-00953-4>.
- Kouchaki S, Yang Y, Walker TM, Sarah Walker A, Wilson DJ, Peto TEA, et al. Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics* 2019;35(13):2276 – 82. <http://dx.doi.org/10.1093/bioinformatics/bty949>.
- The CRyPTIC Consortium, Lachapelle AS. A generalisable approach to drug susceptibility prediction for *M. tuberculosis* using machine learning and whole-genome sequencing. *Microbiology*. 2021. <http://biorxiv.org/lookup/doi/10.1101/2021.09.14.458035>. [2022-2-25].
- Hunt M, Bradley P, Lapierre SG, Heys S, Thomsit M, Hall MB, et al. Antibiotic resistance prediction for *Mycobacterium tuberculosis* from genome sequence data with Mykrobe. *Wellcome Open Res* 2019;4:191. <http://dx.doi.org/10.12688/wellcomeopenres.15603.1>.
- Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, et al. Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. *Nat Rev Microbiol* 2019;17(9):533 – 45. <http://dx.doi.org/10.1038/s41579-019-0214-5>.
- World Health Organization. Guidance for surveillance of SARS-CoV-2 variants: interim guidance. 2021. [https://www.who.int/publications/item/WHO\\_2019-nCoV\\_surveillance\\_variants](https://www.who.int/publications/item/WHO_2019-nCoV_surveillance_variants). [2022-2-25].
- Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* 2021;592(7854):438 – 43. <http://dx.doi.org/10.1038/s41586-021-03402-9>.
- Porter S. Oxford University partners with Oracle for COVID-19 variant identification. *Healthcare IT News*. 2021. <https://www.healthcareitnews.com/news/emea/oxford-university-partner-oracle-covid-variant-identification>. [2022-2-25].



Associate Professor Timothy M Walker  
Wellcome Trust Clinical Career Development Fellow  
Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam



Professor Derrick W Crook  
Professor of Microbiology, Nuffield Department of Medicine  
University of Oxford, Oxford, UK