

GENOME WATCH

Lean, mean, learning machines

Nicole E. Wheeler, Leonor Sánchez-Busó, Silvia Argimón and Benjamin Jeffrey



This month's Genome Watch examines how novel machine learning-enabled molecular diagnostic approaches can predict antibiotic resistance when genetic variation falls short.

Antimicrobial resistance (AMR) is a major public health threat, and we need advanced diagnostics to better inform the course of treatment and ensure that the right antibiotics are prescribed. Culture-based antimicrobial susceptibility testing (AST) usually takes between 24–48 hours for fast-growing microorganisms and up to months for *Mycobacterium tuberculosis*, leaving a time window during which the patient could be given harmful antibiotics unnecessarily or face a delay in receiving effective treatment. Whole-genome sequencing could provide information within hours of clinical sample collection, including insights into antibiotic resistance and virulence traits of the pathogen as well as patient-to-patient transmission patterns. For some pathogens, known genetic mechanisms explain most of the observed resistance with accuracy in line with laboratory testing¹; however, for other pathogen–antibiotic combinations, the effect of rare mutations or mutations that alter gene expression makes the prediction of the resistance phenotype difficult. Recent machine learning approaches might overcome these barriers.

In *M. tuberculosis*, rare variants make a substantial contribution to some resistance phenotypes, but the small sample sizes currently available lack the power to assess the individual effects of the mutations. Chen et al.² modelled resistance to 11 drugs in 3,601 *M. tuberculosis* isolates. They found that including pooled rare variants across genes in 28 targeted genomic regions substantially improved predictions compared with including common mutations only. Previously

uncharacterized mutations in the *embA* gene were important predictors of ethambutol resistance, pointing to the potential for machine learning to identify novel candidate resistance mechanisms for further investigation. Accuracy also improved slightly when the authors built models to predict resistance to all drugs simultaneously, which enables the sharing of information across phenotypes.

In *Pseudomonas aeruginosa*, differences in the expression of resistance genes account for some resistance phenotypes that currently cannot be explained by genetic variation alone³. Using 414 *P. aeruginosa* clinical isolates, Khaledi et al.⁴ examined the ability of machine learning methods to predict resistance to four common anti-pseudomonal antimicrobials using data on gene presence or absence (GPA), genetic variation and gene expression under standard culture conditions. The relative importance of the three data types differed markedly for the different antibiotics, reflecting the different mechanisms of resistance. Whereas ciprofloxacin resistance could be accurately predicted using genetic variation alone, including gene expression data substantially improved prediction of ceftazidime, tobramycin and meropenem resistance over GPA alone. The importance of expression of the multidrug efflux pump *oprAB–oprM* was detected for meropenem, but not ciprofloxacin and ceftazidime, which could potentially have been remedied by a multi-drug modelling approach or the inclusion of more samples. The authors built equally accurate models using <100 features (out of >80,000), potentially enabling the development of simpler molecular tests. However, the requirement of expression data limits the practicality of this approach as a rapid point-of-care test without further experimental work to dissect expression regulation and link genetic variation to gene expression⁵.

Both studies also considered the impact of population structure on their models: Chen et al.² examined the reliance of their models on known lineage markers and found that they were more predictive of sensitivity than resistance; and Khaledi et al.⁵ identified differences in the reliance of different model types on genetic background by testing their models on withheld sequence types. The type of data required, and the right method for modelling the data, differs depending on the resistance mechanism. In both studies the simplest model tested, a logistic regression, performed best for some drugs, but more complex methods were superior for some hard-to-predict phenotypes. Algorithms for predicting AMR when there is a straightforward link between genotype and phenotype currently report high accuracy, but more biologically informed approaches are needed for cases for which standard predictive methods still struggle. These studies illustrate progress in bridging these gaps.

Nicole E. Wheeler✉, Leonor Sánchez-Busó,
Silvia Argimón and Benjamin Jeffrey
Centre for Genomic Pathogen Surveillance, Wellcome
Sanger Institute, Wellcome Genome Campus,
Hinxton, UK.

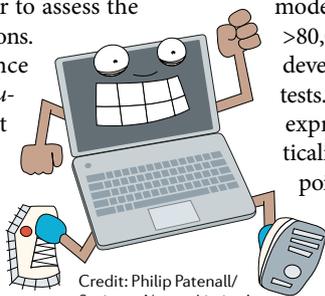
✉e-mail: microbes@sanger.ac.uk

<https://doi.org/10.1038/s41579-020-0357-4>

- Bradley, P. et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat. Commun.* **6**, 10063 (2015).
- Chen, M. L. et al. Beyond multidrug resistance: leveraging rare variants with machine and statistical learning models in *Mycobacterium tuberculosis* resistance prediction. *EBioMedicine* **43**, 356–369 (2019).
- Martin, L. W. et al. Expression of *Pseudomonas aeruginosa* antibiotic resistance genes varies greatly during infections in cystic fibrosis patients. *Antimicrob. Agents Chemother.* **62**, e01789–18 (2018).
- Khaledi, A. et al. Predicting antimicrobial resistance in *Pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics. *EMBO Mol. Med.* **12**, e10264 (2020).
- Belliveau, N. M. et al. Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proc. Natl Acad. Sci. USA* **115**, E4796–E4805 (2018).

Competing interests

The authors declare no competing interests.



Credit: Philip Patenall/
Springer Nature Limited