of general control nonderepressible 2 (GCN2) in pulmonary veno-occlusive disease. *J Heart Lung Transplant* 2018;37: 647-655.

Analysis of a Novel *pncA* Mutation for Susceptibility to Pyrazinamide Therapy

To the Editor:

Pyrazinamide (PZA), which is an analog of nicotinamide, is an important first-line drug used in the short-course treatment of tuberculosis. PZA is a prodrug devoid of significant antibacterial activity. It is metabolized into its active form, pyrazinoic acid, by the amidase activity of the *Mycobacterium tuberculosis* nicotinamidase/pyrazinamidase, encoded by the *pncA* gene. Mutations in *pncA* that prevent activation of the prodrug represent the major mechanism of PZA resistance in *M. tuberculosis* (1). This antibiotic plays a key role in shortening the duration of antituberculous treatment because of its activity against the persisting tubercle bacilli at acidic pH.

Current phenotypic testing for PZA drug susceptibility is problematic. Culture-based methods such as Wayne's method are used as screening assays with confirmation of resistant strains via the BD BACTEC MGIT 960 system (Becton Dickinson) (2). Results obtained from phenotypic laboratory testing have poor reproducibility. Sequencing of the *pncA* gene to determine the presence of mutations may be a more reliable method for confirmation of phenotypic PZA resistance (3). International recommendations suggest continued usage of PZA irrespective of susceptibility results, particularly in the treatment of multidrug-resistant disease (4). This is despite the adverse effects associated with PZA treatment.

Case Report

In early 2017, a 42-year-old woman, originally from Vietnam, presented with right upper lobe pneumonia; she was diagnosed with pulmonary tuberculosis. Phenotypic drug susceptibility testing identified resistance to isoniazid, rifampicin, pyrazinamide, and ethambutol. Although drug susceptibility testing suggested the patient was phenotypically resistant to PZA, consistent with World Health Organization recommendations, PZA treatment was continued as part of a multidrug-resistant tuberculosis regimen. Amplicon sequencing identified a novel frameshift mutation in the *pncA* gene of *M. tuberculosis* (c.85_86insG). Given the uncertain impact of this mutation, we went on to consider whether computational analysis of protein structure (5) could provide insight into the potential efficacy of PZA.

Methods

We have developed an *in silico* mutational analysis platform that is able to characterize the molecular consequences of mutations on protein structure and function (5). This has been used to preemptively identify likely resistance mutations in drug targets (6, 7). Using these tools, we assessed the biophysical changes on mutation on the structure of PncA and drug activation.

A list of 617 nonsynonymous single-nucleotide variants (nsSNVs) of *pncA* was obtained from the GMTV (Genome-wide *Mycobacterium tuberculosis* Variation) Database Project, Tuberculosis Drug Resistance Mutation Database, and saturation mutagenesis (8). Mapping nsSNVs associated with resistance onto the crystal structure of PncA revealed that they were distributed throughout the entire protein structure (Figure 1A), complicating resistance inference from sequence analysis. The structural and functional effects of these mutations were assessed using our graph-based signature pipeline (5). This provided insight into how the curated nsSNVs altered protein folding, stability, conformation, and PZA-binding affinity. This information was used to train a Random Forest (machine-learning algorithm) binary classifier, using the Weka toolkit. Random Forest is an ensemble-learning robust classification algorithm, in which multiple decision trees are included over a random subset of features and decide the output via majority voting. The model was trained by 10-fold cross-validation and performance evaluated by area under the receiver operating characteristic curve, precision, and accuracy. Further validation of the models was performed using two subsets of 93 mutations, which were nonredundant at the position-level mutations in the training set. Analysis of the final model revealed a set of structural features that distinguished between susceptible and resistant *pncA* point mutations.

Building on this structural analysis, the functional consequence of the novel clinical frameshift mutation was analyzed in the context of the protein structure. The experimental crystal structure of holo-wild-type PncA (PDB ID: 3PL1) (9) was minimized in Prime, and PZA docked into the active site using Glide, two exclusive packages of the comprehensive homology modeling software Schrödinger Suites. The docking revealed that PZA formed key interactions within the pocket, including with the catalytic triad (Asp8, Lys96, and Cys138), substrate-binding residues (Trp68 and Phe13), and the iron center (Asp49, His51, His57, and Fe^{2+}) (10). The wild-type and mutant protein sequences were manually aligned and displayed with ESPript 3.0 (Figure 2A), and the structure of the mutant (Figure 2C) was generated by homology and *ab initio* modeling, using the experimental structure of the wild type (Figure 2B) (Schrödinger Suites).

Results

Using the structural and biophysical effects of the mutations on the protein structure, we were able to classify mutations as either susceptible or resistant with an accuracy of 77% (Figure 1C). This approach performed equally well in the identification of either class, correctly classifying all mutations previously associated with conferring PZA resistance at high confidence, and mutations not involved in PZA resistance (100% accuracy) (11).

Supported by a Melbourne Research Scholarship from the University of Melbourne (M.K.). D.B.A was funded by a Newton Fund RCUK-CONFAP Grant awarded by the Medical Research Council and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (MR/M026302/1), the Jack Brockhoff Foundation (JBF 4186, 2016), and a C. J. Martin Research Fellowship from the National Health and Medical Research Council of Australia (APP1072476).

Author Contributions: M.K., J.T.D., and D.B.A. were involved in the study design, execution, data analysis, and writing of all versions of this work. M.G., J.A.M.F., T.P.S., P.D.R.J., and N.E.H. were involved in clinical and laboratory aspects of investigation. All authors contributed to preparation of this manuscript and approve the final version.

Originally Published in Press as DOI: 10.1164/rccm.201712-2572LE on April 25, 2018

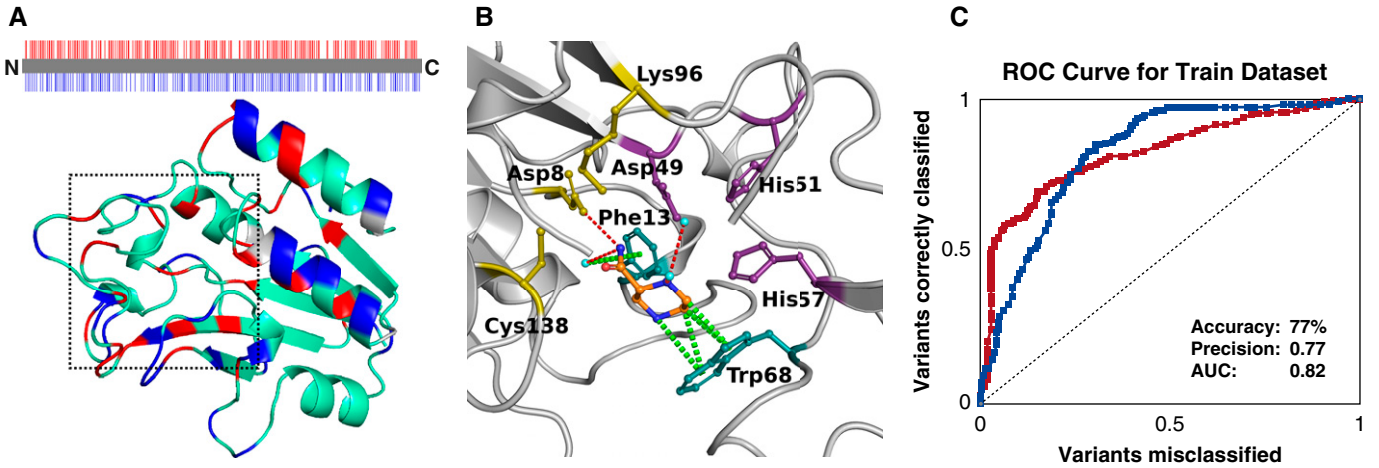

Figure 1. Identification of resistant and susceptible missense mutations in *pncA*. (*A*) The protein sequence and structure of PncA is colored by whether resistant (red) or susceptible (blue) variants have been observed at that location. Highlighting the difficulty of genomic analysis of *pncA*, both resistant and susceptible variants have been observed across many residue positions (cyan). The catalytic site in which pyrazinamide (PZA) was docked is located inside the dashed box. (*B*) The key molecular interactions between PZA (orange segments) and the catalytic triad (yellow), substrate-binding site (teal), and iron center (purple). Hydrogen bonds are shown as red dashes, and π interactions as green dashes. (*C*) The ROC curve shows that, using the structural and functional consequences of the variants, we were able to accurately identify resistant (red) and susceptible (blue) variants. AUC = area under the curve; ROC = receiver operating characteristic.

Strain-specific differences in variants with conflicting experimental data (12) could be detected by our tool, using homology models of the corresponding strain's PncA protein.

Analysis of our model revealed that PncA-resistant mutations were associated with large changes in protein folding and stability (mCSM-Stability scores ≥ -1.72 kcal/mol) ($P < 0.0001$) or located in close proximity to the catalytic triad and substrate-binding site (<8.54 Å) ($P < 0.0001$). Therefore, these freely available biophysical measurements could provide useful information to help guide genomic analysis of novel *pncA* variants.

We next considered the patient's frameshift mutation in light of these structural insights. As shown in Figure 2, the frameshift mutation resulted in the generation of a truncated and incomplete protein that lacked the active site pocket, including most of the catalytic residues and iron coordination residues necessary for activity. This strongly suggests that the *pncA* c.85_86insG frameshift mutation would lead to a total loss of catalytic activity of the protein, and hence PZA treatment would be completely ineffective in this case, as the mutant PncA could not activate the prodrug. This is reflected in the structure of the mutant protein, which is incomplete and would lack any activity (Figure 2). This result was consistent with phenotypic testing, and accordingly, pyrazinamide treatment was ceased.

Discussion

This case study demonstrates the power of using structural information to quantitatively evaluate novel variants in real time, providing invaluable insight to help guide therapy. Although existing recommendations may suggest continuing treatment of multidrug-resistant tuberculosis with pyrazinamide irrespective of phenotype testing, our approach suggests that using structural information to guide analysis of genomic sequencing may offer useful tools for clinicians to consider. These structural insights also assist in informing the mechanisms for drug activity and the development of resistance. Our approach is not limited only to analysis of variants in *pncA* but could be applied to any protein associated with resistance for infectious and noninfectious disease treatment. ■

Author disclosures are available with the text of this letter at www.atsjournals.org.

Malancha Karmakar, B.Sc., M.Sc.
University of Melbourne
Melbourne, Victoria, Australia

Maria Globan, B.Sc.
Janet A. M. Fyfe, B.Sc. (Hons.), Ph.D.
University of Melbourne
Melbourne, Victoria, Australia
and
Melbourne Health
Melbourne, Victoria, Australia

Timothy P. Stinear, B.Sc. (Hons.), Ph.D.
University of Melbourne
Melbourne, Victoria, Australia

Paul D. R. Johnson, M.B. B.S., Ph.D., F.R.A.C.P.
University of Melbourne
Melbourne, Victoria, Australia
and
World Health Organization Collaborating Centre for Mycobacterium ulcerans
Melbourne, Victoria, Australia

Natasha E. Holmes, M.B. B.S., Ph.D., F.R.A.C.P., Grad.Cert.Clin.Teach., Grad.Cert.Clin.Ed.
University of Melbourne
Melbourne, Victoria, Australia
and
Austin Health
Heidelberg, Victoria, Australia

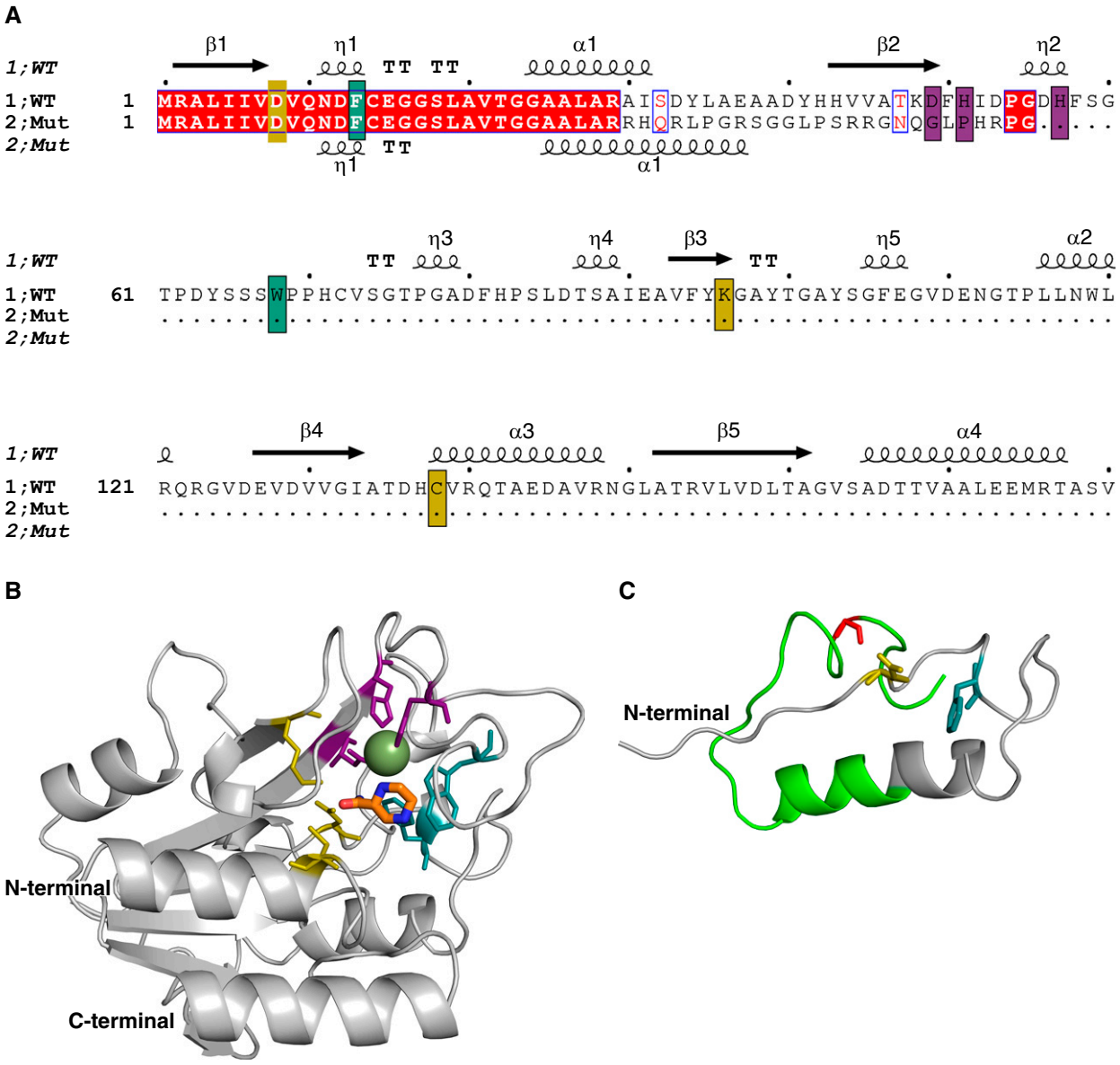

Figure 2. Structural analysis of a novel *pncA* frameshift mutation. (*A*) The sequence alignment between the wild-type and mutant protein sequences shows that only the first 29 residues are conserved (red), and that the frameshift leads to the introduction of a premature stop codon. The catalytic triad, substrate-binding site, and iron center are highlighted in yellow, teal, and purple, respectively. The secondary structure of the wild-type PncA protein is shown above the sequence (β = β sheet, α = α helix, and η = loop). (*B*) The structure of the wild-type PncA protein is represented as a ribbon (gray), bound to the drug pyrazinamide (in orange segments). The Fe^{2+} ion is shown as a green sphere. (*C*) The modeled structure of the mutant PncA protein highlights that most of the catalytic site and structure of the wild-type protein is absent in the mutant. The region not conserved with the wild-type sequence is shown in green. Both wild-type and mutant structures are shown from the same perspective. Mut = mutant; WT = wild type.

Justin T. Denholm, B.Med., M.Bioethics, M.P.H.+T.M., Ph.D., F.R.A.C.P.
University of Melbourne
Melbourne, Victoria, Australia

David B. Ascher, B.Biotech., B.Sc. (Hons.), L.L.B., Ph.D., M.R.A.C.I. C.Chem.
University of Melbourne
Melbourne, Victoria, Australia
and
University of Cambridge
Cambridge, United Kingdom

ORCID ID: 0000-0003-2948-2413 (D.B.A.).

References

1. Zhang Y, Shi W, Zhang W, Mitchison D. Mechanisms of pyrazinamide action and resistance. *Microbiol Spectr* 2013;2:1–12.
2. Cui Z, Wang J, Lu J, Huang X, Zheng R, Hu Z. Evaluation of methods for testing the susceptibility of clinical *Mycobacterium tuberculosis* isolates to pyrazinamide. *J Clin Microbiol* 2013;51:1374–1380.
3. Chang KC, Yew WW, Zhang Y. Pyrazinamide susceptibility testing in *Mycobacterium tuberculosis*: a systematic review with meta-analyses. *Antimicrob Agents Chemother* 2011;55:4499–4505.
4. World Health Organization. WHO treatment guidelines for drug-resistant tuberculosis: 2016 update. Geneva, Switzerland: World Health Organization; 2016.

5. Pires DE, Chen J, Blundell TL, Ascher DB. *In silico* functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci Rep* 2016;6:19848.
6. Park Y, Pacitto A, Bayliss T, Cleghorn LA, Wang Z, Hartman T, *et al*. Essential but not vulnerable: indazole sulfonamides targeting inosine monophosphate dehydrogenase as potential leads against *Mycobacterium tuberculosis*. *ACS Infect Dis* 2017;3:18–33.
7. Singh V, Donini S, Pacitto A, Sala C, Hartkoorn RC, Dhar N, *et al*. The inosine monophosphate dehydrogenase, GuaB2, is a vulnerable new bactericidal drug target for tuberculosis. *ACS Infect Dis* 2017;3:5–17.
8. Yadon AN, Maharaj K, Adamson JH, Lai YP, Sacchettini JC, Ioerger TR, *et al*. A comprehensive characterization of PncA polymorphisms that confer resistance to pyrazinamide. *Nat Commun* 2017;8:588.
9. Petrella S, Gelus-Ziental N, Maudry A, Laurans C, Boudjelloul R, Sougakoff W. Crystal structure of the pyrazinamidase of *Mycobacterium tuberculosis*: insights into natural and acquired resistance to pyrazinamide. *PLoS One* 2011;6:e15785.
10. Jubb HC, Higueruelo AP, Ochoa-Montaño B, Pitt WR, Ascher DB, Blundell TL. Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J Mol Biol* 2017;429: 365–371.
11. Miotto P, Cabibbe AM, Feuerriegel S, Casali N, Drobniewski F, Rodionova Y, *et al*. *Mycobacterium tuberculosis* pyrazinamide resistance determinants: a multicenter study. *MBio* 2014;5: e01819-14.
12. Baddam R, Kumar N, Wieler LH, Lankapalli AK, Ahmed N, Peacock SJ, *et al*. Analysis of mutations in *pncA* reveals non-overlapping patterns among various lineages of *Mycobacterium tuberculosis*. *Sci Rep* 2018;8:4628.

Overfitting and Use of Mismatched Cohorts in Deep Learning Models: Preventable Design Limitations

To the Editor:

We read with great interest the study by González and colleagues (1) in which they used deep learning models to learn from the computed tomography (CT) scans of 7,983 participants in the COPDGene (Genetic Epidemiology of COPD) study (2). Their objective was to learn from visual data present in these CT scans and subsequently study the model's ability to diagnose chronic obstructive pulmonary disease (COPD) and predict respiratory events and mortality in a validation cohort (1,000 COPDGene scans) and a test cohort (1,672 ECLIPSE [Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points] [3] scans). The validation and test cohorts differed significantly in terms of their COPD severity (lower FEV_1% predicted and higher Global Initiative for Chronic Obstructive Lung Disease [GOLD] stage in ECLIPSE [3]).

The model performed very well in terms of COPD detection as well as prediction of acute respiratory events in the validation cohort of 1,000 COPDGene participants (i.e., it correctly identified COPD in 773/1,000 scans, and there was a strong correlation between the actual FEV_1 and the predicted FEV_1 [1]). However, the model's performance in the test cohort was inferior in both detection of COPD and prediction of acute respiratory events (only 29% of individuals were correctly staged, and the model was unable to identify patients at higher risk of respiratory events [1, 4]).

We believe that there are two significant design limitations in the authors' approach toward execution of the deep learning process and selection of the cohorts.

A significant proportion of the CT scan data (7,983 out of a total of 8,983 COPDGene scans, 88.9%) were used for training the deep learning model (1). This leads to a potential overfitting of the learning model. Overfitting is the consequence of the model learning from a high volume of details that incorporate both noise and signal existing in the training datasets. This leads to a superior performance in the internal validation cohort and inferior performance in an external test dataset (5). In other words, such models do not explain test cohorts, but they explain the training data very well (5). This suspicion is supported by the study's superior results in the smaller internal validation cohort and inferior performance in the external test cohort. This is particularly relevant because the validation cohort (n = 1,000 scans, 10% of the COPDGene cohort) likely does not represent most of the variance existing in the COPDGene cohort (a cohort of smokers with and without COPD [2]).

The second limitation arises as an indirect consequence of inherent differences between the COPDGene and ECLIPSE cohorts. The authors do acknowledge in their discussion that there are significant differences between the validation and test cohorts (1). In this study, a predominantly GOLD stage 0–1 cohort (1, 2) served as the training set for a model that was tested in a GOLD stage $\geqslant$2 cohort (3). In our opinion, selecting COPDGene scans with established GOLD stages of $\geqslant$2 (representing 36% of the COPDGene cohort [1], n = 3,600 scans) for teaching and internal validation purposes would have improved the external performance. An ideal deep learning strategy would have allocated 50–70% (n = 1,800–2,500) of these scans to the learning model and the remaining 30–50% scans (n = 1,080–1,800) to the internal validation effort. This would have resulted in a true enumeration of the model's performance in the internal validation phase.

In conclusion, the findings could simply represent the performance of a potentially overfitted model (5) and likely do not reflect the suggested superior performance of the tool in the COPDGene validation dataset. The lack of use of appropriate cohorts for training and validation is another significant limitation and can explain the inferior performance in the test cohort (ECLIPSE). ■

Author disclosures are available with the text of this letter at www.atsjournals.org.

Srinivas R. Mummadi, M.D., M.B.I.
Metro Health-University of Michigan Health
Wyoming, Michigan

Akrum Al-Zubaidi, D.O.
National Jewish Health
Denver, Colorado

Peter Y. Hahn, M.D., M.B.A.
Metro Health-University of Michigan Health
Wyoming, Michigan

ORCID IDs: 0000-0002-8806-445X (S.R.M.); 0000-0001-6143-1144 (A.A.-Z.); 0000-0003-2410-6152 (P.Y.H.).

Originally Published in Press as DOI: 10.1164/rccm.201802-0350LE on April 11, 2018