Prediction of drug resistance using genotypic data

An argument for machine learning

CPTM Workshop
3/22/2017

Maha Farhat MD MSc
Department of Biomedical Informatics
Pulmonary and Critical Care Medicine
Harvard Medical School
TB & Resistance Diagnosis with Genomic Big Data

Drug Resistance:
- Rifampicin = R
- Isoniazid = R
- Pyrazinamide = S
- Streptomycin = S
- Ethambutol = R
- Amikacin = S
- Capreomycin = S
- Kanamycin = S
- Ofloxacin = R
- Moxifloxacin = S
- Ethionamide = R
- Bedaquiline = S
- Linezolid = S
- Cycloserine = S

Sputum collection → 0-1 day

Whole Genome Sequencing → Minutes

Treatment started → 0-1 day
Genetic basis of drug resistance

1:1 Mutation : Drug resistance
Clinical resistance is more complex

Evidence for epistasis/ gene-gene interaction

Farhat et al. JCM 2016
Nebenzahl-Guimaraes, Jacobson, Farhat, Murray. JAC 2014
The molecular diagnostic gap:

For Sensitivity

- Phenotypic overcall
- Heterogeneity
- Novel mutations or loci
- Epistasis

For Specificity

- Phenotypic undercall
- False positive associations
- Intermediate resistance mutations
- Epistasis
Geographic source and diversity of the TB isolates

Random Forest Classifier
Identifying the minimum set of mutations predictive of drug resistance

Farhat et al. AJRCCM 2016
Validated Predictive performance for drug resistance diagnosis

### Drug Selected Mutations

<table>
<thead>
<tr>
<th>Drug</th>
<th>Selected Mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isoniazid</td>
<td>19</td>
</tr>
<tr>
<td>Rifampicin</td>
<td>14</td>
</tr>
<tr>
<td>Pyrazinamide</td>
<td>124</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>18</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>39</td>
</tr>
<tr>
<td>Ethionamide</td>
<td>20</td>
</tr>
<tr>
<td>Kanamycin</td>
<td>3</td>
</tr>
<tr>
<td>Capreomycin</td>
<td>5</td>
</tr>
<tr>
<td>Amikacin</td>
<td>2</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>7</td>
</tr>
<tr>
<td>Levofoxacin</td>
<td>8</td>
</tr>
<tr>
<td>Ofloxacin</td>
<td>6</td>
</tr>
<tr>
<td>p-aminosalicylic acid</td>
<td>4</td>
</tr>
</tbody>
</table>

**Total** 250
Predict Upload your genetic data to make drug resistance prediction.

To examine the frequency of individual mutations by drug resistance you can consult the Map function for country specific data, and the Explore function to examine their frequency in all of the data.

New FastQ Pair-Ended Prediction
Create a prediction from a set of pair-ended FastQ genetic sequences. This option involves the largest files and takes more time to process that the vcf or manual options.

New FastQ Single-Ended Prediction
Create a prediction from a single-ended FastQ genetic sequence file. This option involves a large file and takes more time to process that the vcf or manual options.

New Variant Call Format Prediction
Minimal genotypic information for accurate resistance predictions are below. Genetic regions are listed in order of decreasing importance. For more detailed list of genetic variants see reference Farhat MR, Sultana R et al. Genetic Determinants of Drug Resistance in Mycobacterium tuberculosis and Their Diagnostic Value AJRCCM 2016 and get More information about the vcf format.

New Manually Entered Prediction
Create a prediction from a set of pair-ended FastQ genetic sequences. This option involves the largest files and takes more time to process that the vcf or manual options.
RF Validation using Reseq public dataset

2,429 Isolates:
- 348 INH R
- 129 RIF R
- 63 EMB R
- 91 PZA R
- 88 STR R
- 20 OFLX R

Comparison with canonical mutation +/-

Low INH sensitivity
- 83% (canonical mutations)
- 85% (RF predictor)
- INH-mono resistance?
Neural network model

Pyrazinamide (PZA): Predictive Model Performance

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Random Forest</td>
<td>0.640 (0.030)</td>
<td>0.920 (0.030)</td>
<td>1.560</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random Forest (1)</td>
<td>0.791 (0.019)</td>
<td>0.622 (0.029)</td>
<td>0.929 (0.019)</td>
<td>1.550</td>
<td>0.506 (0.061)</td>
</tr>
<tr>
<td>Neural Network (1)</td>
<td>0.856 (0.017)</td>
<td>0.718 (0.027)</td>
<td>0.946 (0.022)</td>
<td>1.664</td>
<td>0.573 (0.050)</td>
</tr>
<tr>
<td>Random Forest (2)</td>
<td>0.813 (0.018)</td>
<td>0.649 (0.029)</td>
<td>0.925 (0.020)</td>
<td>1.573</td>
<td>0.505 (0.056)</td>
</tr>
<tr>
<td>Neural Network (2)</td>
<td>0.883 (0.016)</td>
<td>0.752 (0.028)</td>
<td>0.943 (0.024)</td>
<td>1.695</td>
<td>0.577 (0.050)</td>
</tr>
</tbody>
</table>
With thanks

Collaborators
- Megan Murray, Department of Global Health and Social Medicine
- PIH/DGHSM (Molly Franke, Carole Mitnick)
- MSLI (Alex Sloutsky and Devinder Kaur)
- Stellenbosch University (Rob Warren, Tommie Victor and Lizma Streicher)
- RIVM (Dick van Soolingen, Hanna Guimaraes)
- US CDC (Bonnie Plikyatis, Jamie Posey)
- PHRI (Barry Kreiswirth, Natalia Kurepina)
- WHO - TDR (Leen Rigouts)
- Boston Medical Center (Karen Jacobson)
- IQSS (Gary King, Merce Crosas, Christine Choirat, Raman Prasad, James Honaker)

Funders

Partners In Health

Bill & Melinda Gates Foundation

American Lung Association

NIH

Big Data to Knowledge (BD2K)

CIFAR

Harvard University Center for AIDS Research