Scaling up sequencing in order to determine sequence diversity worldwide and for the identification of genetic variants that confer resistance to all old and new anti-mycobacterial drugs

Derrick Crook
Public Health England
University of Oxford
Oxford Universities Hospitals
Can the all/most genomic variation conferring anti-tuberculosis drug resistance be discovered?

- Many resistance determinants have been discovered, are there many more?

- How best to discover more?

- Can we dispense with routine phenotyping
  - Need a complete knowledge base
  - Requires a software to process sequences
Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study


Lancet Infect Dis 2015; 15: 1193-1202
Published Online
June 24, 2015
http://dx.doi.org/10.1016/S1473-3099(15)00062-6
Can we discover more determinants

• Resistance is conferred by genomic variation:
  – Non-synonymous mutations, deletions and insertions in relevant genes – 23 genes
  – Arises mostly de-novo in a non-recombining genome leading to homoplasy

• Investigation of 3651 isolates:
  – Using a heuristic method of predicting resistance

• divided into
  – a 2099 derivation set
  – a 1552 validation set
Resistance prediction in a validation set

<table>
<thead>
<tr>
<th>Phenotypically Resistant Genotype</th>
<th>Phenotypically Sensitive Genotype</th>
<th>All</th>
<th>Excluding Unclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>S₀</td>
<td>Sₛ</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Isoniazid</td>
<td>310</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>Rifampicin</td>
<td>275</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>158</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Pyrazinamide</td>
<td>43</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>284</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Ofloxacin</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Amikacin</td>
<td>52</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1127</td>
<td>75</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 1: Genotypic predictions in the validation-set based on: R (resistance-determinant); S₀ (zero non-synonymous variants/SNPs present); Sₛ (only sensitive variants present); U (unclassified variants present). Weighted mean sensitivity and specificity given for all phenotypes, and with the 10.8% of phenotypes associated with previously unclassified variation (U) excluded.
Phenotypic resistance occurring for each genetic variant

Figure 1: The number of resistant and sensitive phenotypes associated with each ‘resistance-determinant’ in the derivation- and validation sets: Black and red respectively for the derivation-set, and gray and orange respectively for the validation set. Variants probed by a line-probe assay are highlighted with red labels. Variants that are only seen once in the derivation-set and not in the validation-set (i.e. with no additional information) are not shown.
Variation conferring resistance
Can we do better?

Yes we can, but will need very large numbers

Comprehensive Resistance Prediction for Tuberculosis: an International Consortium (CRyPTIC)
What needs to be done?

• Improve DST – working on optimising microtitre plate produced by Thermo Fisher
  – Gates Foundation funded

• Increase the sample size of isolates
  – Gates Foundation funded
    • 21,000 isolates (5,000 with extensive DST)
  – Aspiration to get Wellcome Funding
    • 80,000 isolates (37,000 with extensive DST)
  – Potential total 100,000 (43,000 with extensive DST)
Optimise a microtitre DST with ~ 16 drugs

* remove cycloserine and made significant changes
A large global investigation

Approaches to analysis

• Heuristic approach described above

• Statistical genetic approached namesl GWAS

• Machine learning

• Data store shared with CPTR ReSeqTB to build a sustainable data store
Results of pilot kmer genome wide association studies in *Mycobacterium tuberculosis* searching for genetic variants associated with resistance to first-line drugs: isoniazid (INH), rifampicin (RIF), ethambutol (EMB) pyrazinamide (PZA). Each variant is plotted to show its position in the H37Rv reference genome (horizontal axis) and the significance of its association with resistance to each drug (vertical axis), where significances is measured by the \(-\log_{10}(P\text{-value})\). Points are colour-coded to show the association between each variant and resistance to INH (maroon), RIF (blue), EMB (orange) and PZA (yellow). Genes which have mutations that are known to most commonly confer resistance to each drug are indicated by vertical bars, colour-coded to match the corresponding drug: *katG* (INH, maroon), *rpoB* (RIF, blue), *embB* (EMB orange) and *pncA* (PZA, yellow).
Machine Learning approaches

Figure 3: (left) The probability of an isolate being resistant to isoniazid (INH) is shown, depending on whether or not a particular SNP is present (fabG1_*-15*); i.e., the figure shows $P(\text{resistant} \mid \text{SNP})$. As more sequences are observed (up to $N = 1808$), the certainty of resistance to INH, given the presence of this SNP, increases. The converse situation is shown (right), which plots the probability of having the same particular SNP (fabG1_*-15*) if the isolate is resistant; i.e., $P(\text{SNP} \mid \text{resistant})$. Again, observing more TB sequences improves the certainty of the estimation (up to $N = 1808$ isolates). These plots show that if this SNP is present, then resistance is almost guaranteed; however, if the isolate is resistant, then there is only a 25% chance that this particular SNP will be present. The link between resistance and this particular SNP has been previously described in the literature.
Outcomes

• Probability a variant confers resistant for most clinically used drugs
• Frequency distribution of resistance conferring variants
• Need to assess the clinical impact of vary rare variants
• Extent to which other mechanisms other than chromosomal genetic variation contribute to resistance
Wider participation

- Subject to securing Wellcome Trust funding, the CRyPTIC consortium will be seeking participation from the wider TB community.