Understanding *M. tuberculosis* variants using Galaxy and the COMBAT TB Explorer

Thoba Lose
Software Developer
### Top 10 leading underlying causes of death in South Africa

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cause</th>
<th>2015 (%)</th>
<th>2016 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tuberculosis</td>
<td>7.2%</td>
<td>6.5%</td>
</tr>
<tr>
<td>2</td>
<td>Diabetes mellitus</td>
<td>5.4%</td>
<td>5.5%</td>
</tr>
<tr>
<td>3</td>
<td>Cerebrovascular diseases</td>
<td>5.0%</td>
<td>5.1%</td>
</tr>
<tr>
<td>4</td>
<td>Other forms of heart disease</td>
<td>4.8%</td>
<td>5.1%</td>
</tr>
<tr>
<td>5</td>
<td>HIV disease</td>
<td>4.8%</td>
<td>4.8%</td>
</tr>
<tr>
<td>6</td>
<td>Influenza and pneumonia</td>
<td>4.5%</td>
<td>4.4%</td>
</tr>
<tr>
<td>7</td>
<td>Hypertensive diseases</td>
<td>4.2%</td>
<td>4.3%</td>
</tr>
<tr>
<td>8</td>
<td>Other viral diseases</td>
<td>3.5%</td>
<td>3.6%</td>
</tr>
<tr>
<td>9</td>
<td>Chronic lower respiratory diseases</td>
<td>2.8%</td>
<td>2.8%</td>
</tr>
<tr>
<td>10</td>
<td>Ischaemic heart diseases</td>
<td>2.7%</td>
<td>2.8%</td>
</tr>
</tbody>
</table>

*Stats SA and NDOF*
Whole Genome Sequencing (WGS)

- WGS has provided ways of uncovering disease related variants in both rare and non-rare diseases.
- The workflows/pipelines for variant calling/analysis are quite standard.
- These workflows generate data on a large number of variants.
While workflows and associated software to process raw data and produce high-confidence variant calls have significantly improved, filtering of candidates to identify a subset relevant to a specific study is still a complex exercise...

A practical guide to filtering and prioritizing genetic variants

Mahjoubeh Jalali Sefid Dashti and Junaid Gamieldien
South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa

Keywords: whole-genome sequencing; whole-exome sequencing; functional variants; variant prioritization
Questions a biologist might ask

What part of the gene is this variant located in?

What is the predicted effect of the variant?

Is the variant in a gene associated with a essential pathway?

Is the variant in a gene that encodes a protein that interacts with a known disease protein?

Is the variant in a gene that has a function that coincides with the pathology of TB?

Is the variant in a gene that encodes a target protein (drug target)?

There is a possibility that genes associated with IPR000873 activate fatty acids during dormancy.
Variant Annotation

- Involves adding metadata and knowledge in effort to enhance assessment of variants likely to impact function.

- SnpEff is a variant annotation and effect prediction tool.
### SnpEff

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL FILTER</th>
<th>INFO</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>1290</td>
<td>rs6875309</td>
<td>T</td>
<td>C</td>
<td>56</td>
<td>PASS NS=4;DP=20;AF=0.045</td>
</tr>
<tr>
<td>9</td>
<td>680</td>
<td>rs44309</td>
<td>G</td>
<td>A</td>
<td>2</td>
<td>q15 NS=1;DP=3;AF=0.78</td>
</tr>
</tbody>
</table>

- Variant nucleotide information
- VCF or TXT file
- SnpEff
- Annotated and characterized variants
- Modified INFO column and tables
Variant Prioritization

- It is possible to rank variants based on predicted impact. However, identifying the strongest candidates is not obvious.

- Assessing candidate genes in the context of extant knowledge is vital.
Graph Databases

- Storing biological data involves modeling and storing thousands of interrelated entities.
- Relational databases meet very specific needs.
- Graph databases which focus on relationships between entities and are seen as a natural fit.
Neo4j

- A highly scalable graph database with a declarative query language called Cypher
  - Property Graph Model
  - Whiteboard friendly
  - Schema-less
  - Exposes a RESTful API
GA4GH Variant Model
COMBAT TB Explorer

An integrated knowledge management system that allows researchers to interrogate data in the context to TB research and data available in public repositories.
Neo4j Galaxy IE

Based on a customized Neo4j Docker image and utilises the neostore datatype.
tbvcfreport

A tool that utilises the Neo4j database to generate an interactive HTML-based VCF report, from SnpEff produced VCF files, with links to the COMBAT TB Explorer
Conclusion

- The integration of extant biological resources in the context of TB, the Combat TB Explorer combined with Galaxy provide a single, easy to use, integrated approach to prioritize candidate variants.
The Team

Christoffels Lab:
- L. Abdoll
- T. Calvert-Joshua
- A. Christoffels
- K. Direko
- Z. Mashologu
- J. Southgate
- P. van Heusden
This project is funded by:

- South African Medical Research Council
- South African Research Chairs Initiative of the Department of Science and Technology and
- National Research Foundation
THANKS!

Any questions?

You can find me at: thoba@sanbi.ac.za