

COMBAT TB Explorer, a TB data exploration workbench.

Peter van Heusden (pvh@sanbi.ac.za), Ziphozake Mashologu, Alan Christoffels (alan@sanbi.ac.za)
South African National Bioinformatics Institute, University of the Western Cape



Introduction

Tuberculosis (TB), an infectious disease caused by the *Mycobacterium tuberculosis*, ranks as one of the leading causes of death worldwide, with WHO recording 9.6 million people falling ill with 1.5 million deaths in 2014. This disease burden has arguably been matched with continued increase in genomic, transcriptomic and proteomic data for *M. tuberculosis* as a result of NGS technologies. This continued expansion of data is exemplified by the growth of data repositories such as the tuberculosis database (TBDB) and the pathosystems resource integrated center (PATRICBRC). Unfortunately, these resources only present pre-computed data and do not provide the computational toolkit for biomedical researchers to analyze their own data. We have created the COMBAT TB Explorer, a Galaxy-based environment for annotating and exploring *M. tuberculosis* sequence data.

Methods and Results

We implemented a bacterial genomic variant calling workflow in Galaxy based on previous SANBI work on analysing mutations in drug resistance strains of *M. tuberculosis* and a downstream tool that combines variant calls with a reference database of *M. tuberculosis* genome annotation and a JBrowse based genome browser to produce a new Galaxy datatype (ctb_report). To visualise the resulting datasets we implemented COMBAT TB Explorer, a Galaxy Interactive Environment. The result is that variants gleaned from novel sequence data can be explored in the context of known annotation, both from genome-centric and annotation-centric perspectives. The COMBAT TB Explorer also contains a module for geneset enrichment analysis (GSEA).

- Trimming
- Novo Align
- Novo Sort
- Mark Duplicates
- Realigner Target Creator
- Indel Realigner
- Depth of Coverage
- Snpeff

COMBAT-TB Pipeline

168: MPileup on data 1 and data 55 (log)

167: MPileup on data 1 and data 55

166: NOVO ALIGN on data 165 and data 164

165: Trim Galore on data 2 and data 3: trimmed reads pair 2

164: Trim Galore on data 2 and data 3: trimmed reads pair 1

163: Trim Galore on data 3: trimmed reads

162: Trim Galore on data 2: trimmed reads

161: QualiMap: Multi BamQC PDF output

65: Haplotype Caller on data 1 and data 55 (VCF)

55: Indel Realigner on data 53, data 1, and data 51 (log)

54: Realigner Target Creator on data 1 and data 51 (log)

53: Realigner Target Creator on data 1 and data 51 (GATK interval)

52: Collect Alignment Summary Metrics on data 1 and data 51: Summary stats

51: MarkDuplicates on data 48: MarkDuplicates BAM output

50: MarkDuplicates on...

History

COMBAT-TB Pipeline

9.92 GB

168: MPileup on data 1 and data 55 (log)

167: MPileup on data 1 and data 55

166: NOVO ALIGN on data 165 and data 164

165: Trim Galore on data 2 and data 3: trimmed reads pair 2

164: Trim Galore on data 2 and data 3: trimmed reads pair 1

163: Trim Galore on data 3: trimmed reads

162: Trim Galore on data 2: trimmed reads

161: QualiMap: Multi BamQC PDF output

65: Haplotype Caller on data 1 and data 55 (VCF)

55: Indel Realigner on data 53, data 1, and data 51 (log)

54: Realigner Target Creator on data 1 and data 51 (log)

53: Realigner Target Creator on data 1 and data 51 (GATK interval)

52: Collect Alignment Summary Metrics on data 1 and data 51: Summary stats

51: MarkDuplicates on data 48: MarkDuplicates BAM output

50: MarkDuplicates on...

COMBAT-TB

geneset enrichment analysis

GO Term ID	GO Term Name	Raw p-value	Corrected p-value
GO:0045454	cell redox homeostasis	0.0016287817582708746	0.016075395698978433
GO:0004803	transposase activity	0.003152393480534542	0.016075395698978433
GO:0006313	transposition, DNA-mediated	0.003314455982926706	0.016075395698978433
GO:0006739	NADP metabolic process	0.004465387694160676	0.016075395698978433
GO:0003957	NAD(P)+ transhydrogenase (B-specific) activity	0.004465387694160676	0.016075395698978433
GO:0016655	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	0.008913042420826633	0.022919251939268484

Build ctb_gene_tool on data 2

ncbi database (multiple files)

format: ncbi database: 2

2: h37rv.gtf3

16,729 lines, 4,116 comments

format: gtf3 database: 2

uploaded gtf3 file

display with VGV local

hg37 version

2 source > Type > 4: 11

Conclusions and acknowledgements

- Integrating novel sequencing results with existing annotation will be key to researching M. tuberculosis and ultimately fighting tuberculosis.
- Our data exploration environment allows biomedical researchers to analyse, visualise and explore their own data in the context of existing annotation.
- We intend to build on the foundation we have created by extending the COMBAT TB Explorer with additional analysis modules for interrogation molecular pathway relationships, human - pathogen and drug interaction data.

The COMBAT TB project is funded by the South African Medical Research Council (MRC). We would like to acknowledge the support of Zahra Jalali (SANBI) the Galaxy Team and community.