


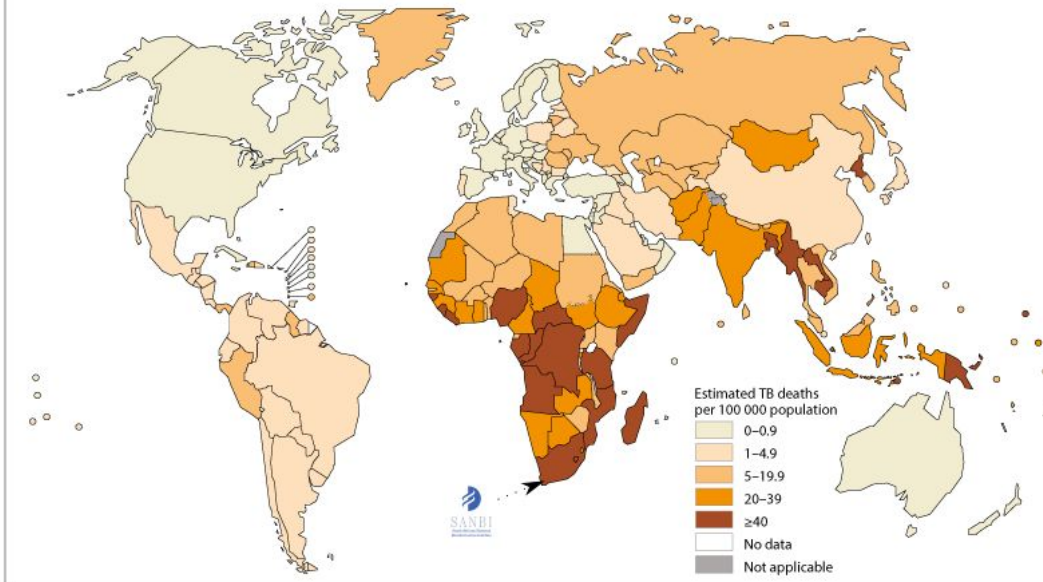


COMBAT TB

An integrated environment for
M. tuberculosis data analysis



Estimated TB mortality rates in HIV-negative people, 2015



The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

Data Source: *Global Tuberculosis Report 2016*. WHO, 2016.

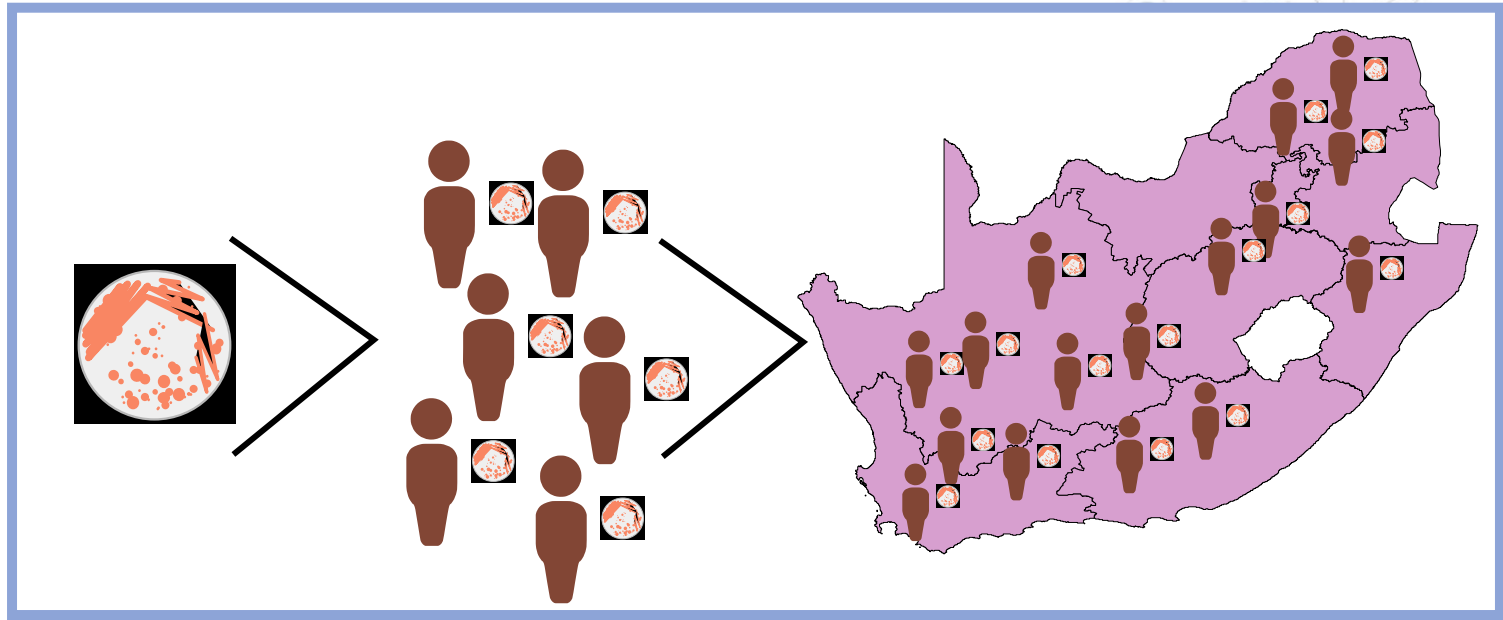
© WHO 2016. All rights reserved.



- Worldwide: more than 10 million infected
- 1.8 million deaths in 2015
- Majority of disease burden in Africa and Asia
- TB is a disease of poverty and (relatively) poor countries
- #1 cause of infectious disease mortality in SA

The COMBAT TB project

- Started at SANBI in 2012 to create tools and databases supporting TB research
 - Funded by the SA Medical Research Council for 5 years
- Building on more than a decade of SANBI work on TB



- Sequencing has moved from single isolates to hundreds of patients
- Whole genome sequencing promises to redraw map of *M. tuberculosis* species variation worldwide



This site can't be reached

www.tbdb.org took too long to respond.

Search Google for [tbdb org](#)

ERR_CONNECTION_TIMED_OUT

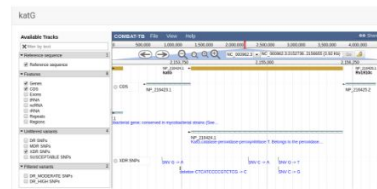
combatTB Explorer.

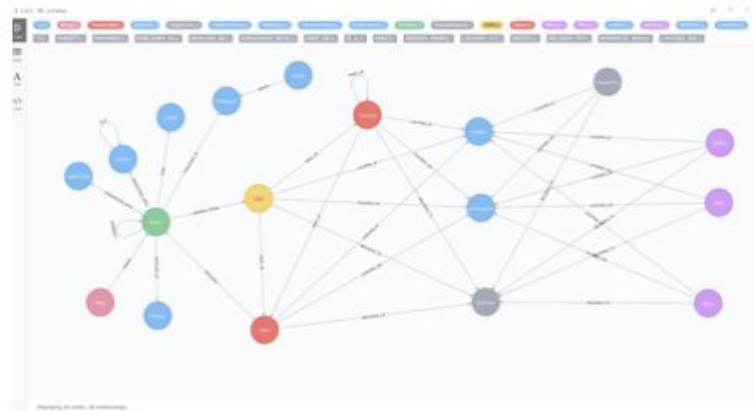
Interrogate data across domains in the context of tuberculosis research.

[Learn More](#)

Explore the TB genome with unparalleled speed.

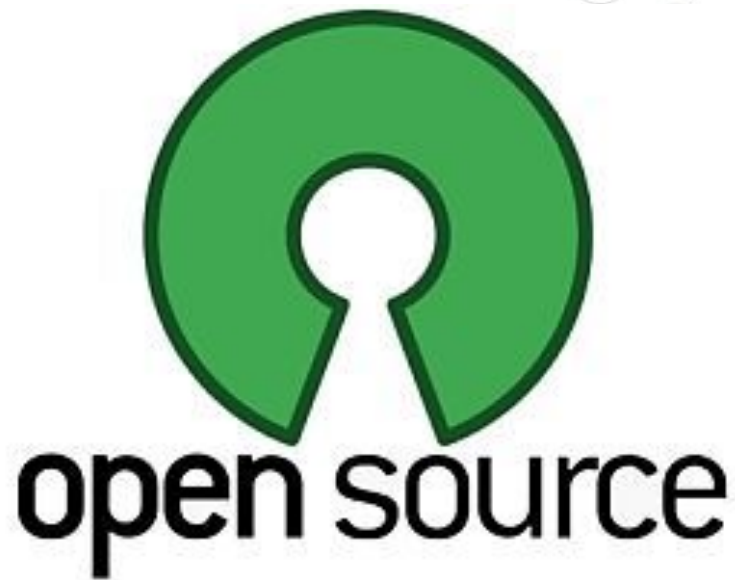
Using a genome browser with a fully dynamic interface. It is very fast and scales well to large datasets.





Queries are served from a graph database.

Understanding the complex relationships that exist among heterogeneous biological data.



Yeah, right few scripts!



Joining Two Worlds of TB Research

Enhance TB research in two ways:

1. Easier and more reproducible computation for the bioinformaticist
2. Accessible informatics and computational resources for the lab biologist



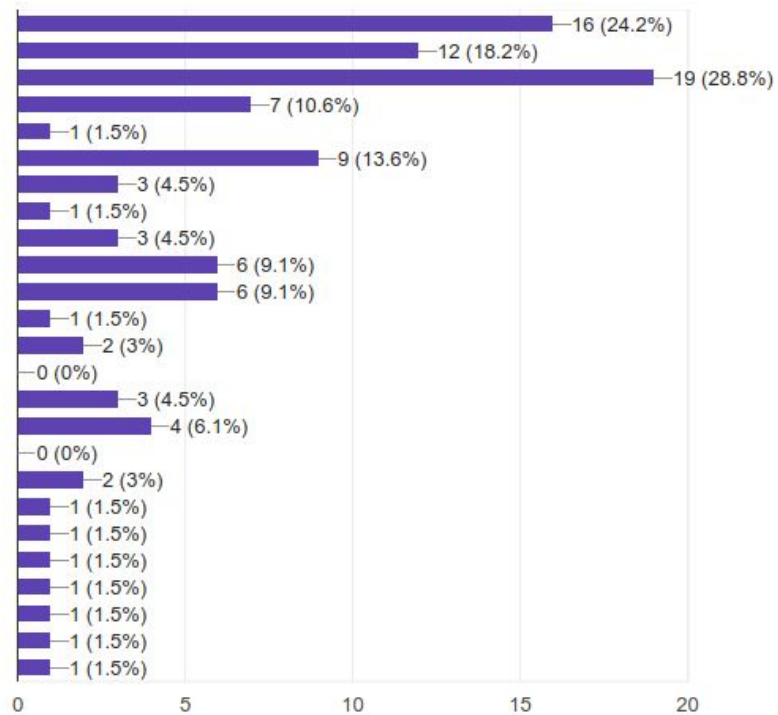
Galaxy

The screenshot displays the Galaxy web interface at ctbqx.sanbi.ac.za. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, and User. The left sidebar contains a 'Tools' section with a search bar and a list of tool categories: Get Data, Send Data, Collection Operations, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Sequences, Fetch Alignments, Statistics, Graph/Display Data, Trimmomatic flexible read trimming tool for Illumina NGS data, QualiMap BamQC Tool to facilitate the quality control of alignment sequencing data and its derivatives like feature counts, SnpEff Available Databases, SnpEff Download Download a new database, SnpEff Variant effect and annotation, Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome, Map with BWA - map short reads (< 100 bp) against reference genome, FreeBayes - bayesian genetic variant detector, and BamLeftAlign Indels in BAM. The main content area shows a 'Hello, Galaxy is running!' message with a 'static/welcome.html' link and buttons for 'Configuring Galaxy' and 'Installing Tools'. Below this, a text block states: 'Take an interactive tour: Galaxy UI History Scratchbook'. The right sidebar shows a 'History' section with a search bar and a list of datasets: tb2neo (8 shown, 12 deleted, 1.09 GB), 20: Mycobacterium tuberculosis_h37rv.GCA_000195955.2.30.gff3, 19: Import SnpEff produced VCF files to a Neo4j Graph database, on data 12, data 13, and others, 16: SnpEff refvct (a list of 3 datasets), 15: Intersect all vcf_Scols_Final_XDR_annotation.vcf, 14: Intersect all vcf_Scols_Final_susceptible_annotation.vcf, and 13: Intersect all vcf. The bottom section shows a workflow diagram with steps: Input dataset collection, Trimmomatic (Trimming FASTQ dataset collection with BBUZ pair), NOVO ALIGN (Aligning FASTQ dataset collection with BBUZ pair), NOVO SORT (Sorting file), and a final output step.

- Converted workflow from “Detecting novel variants associated with *Mycobacterium tuberculosis* drug resistance” (Jalali et al. forthcoming) to Galaxy workflow
- Align using *novobalign*, call variants using *GATK v3* and *FreeBayes*, annotate using *SnpEff* and build phylogeny from variants using *FastTree*

Which microbial SNP pipeline do you use?

66 responses



Source: @AnitaSchurch



Galaxy

PROJECT

COMBAT TB
Applications

Bioinformatics
Tools & Libraries

Scientific
Libraries

General Purpose
Computing

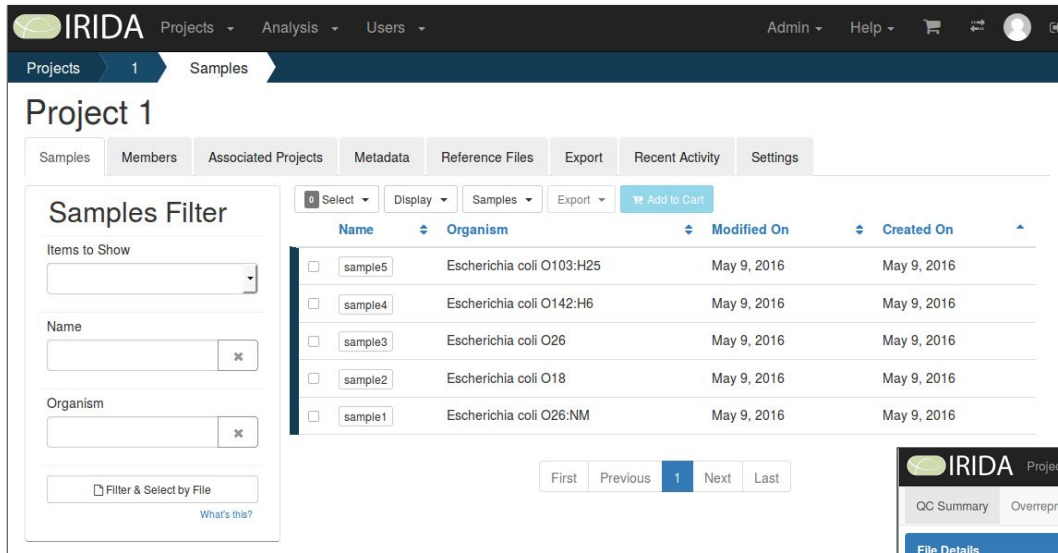
 **Galaxy**
PROJECT

 **IRIDA**

BIOCONDA



 **openstack.**



IRIDA Projects Analysis Users Admin Help

Projects 1 Samples

Project 1

Samples Members Associated Projects Metadata Reference Files Export Recent Activity Settings

Select Display Samples Export Add to Cart

Samples Filter

Items to Show

Name

Organism

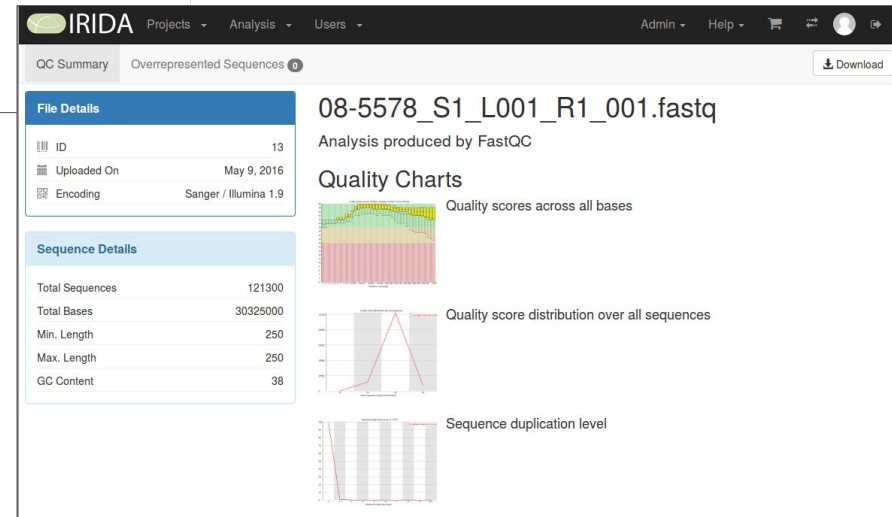
Filter & Select by File

What's this?

Name	Organism	Modified On	Created On
<input type="checkbox"/> sample5	Escherichia coli O103:H25	May 9, 2016	May 9, 2016
<input type="checkbox"/> sample4	Escherichia coli O142:H6	May 9, 2016	May 9, 2016
<input type="checkbox"/> sample3	Escherichia coli O26	May 9, 2016	May 9, 2016
<input type="checkbox"/> sample2	Escherichia coli O18	May 9, 2016	May 9, 2016
<input type="checkbox"/> sample1	Escherichia coli O26:NM	May 9, 2016	May 9, 2016

First Previous 1 Next Last

- Project of Public Health Canada
- Metadata aware sequence storage
- Automated quality control



IRIDA Projects Analysis Users Admin Help

QC Summary Overrepresented Sequences 0 Download

File Details

ID 13
 Uploaded On May 9, 2016
 Encoding Sanger / Illumina 1.9

Sequence Details

Total Sequences	121300
Total Bases	30325000
Min. Length	250
Max. Length	250
GC Content	38

08-5578_S1_L001_R1_001.fastq

Analysis produced by FastQC

Quality Charts

Quality scores across all bases

Quality score distribution over all sequences

Sequence duplication level



IRIDA

- combine analysis
metadata



In Conclusion



combat**TB**



**Galaxy
Africa**

Thanks!

PI: Alan Christoffels

Students:

- Zahra Jalali
- Tracey Calvert-Joshua
- Mmakamohelo Direko
- Lynley Abdoll

IT crew:

- Thoba Lose
- Ziphosakhe Mashologu
- Peter van Heusden
- Eugene de Beste

Reach us at:

combat-tb-help@sanbi.ac.za

@SANBI_SA



**National
Research
Foundation**



**science
& technology**

Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA



**UNIVERSITY of the
WESTERN CAPE**

slide styling: Cordelia from [SlidesCarnival](#)