EMBnet.journal
Bioinformatics in Action

# Training workshop on *Mycobacterium* whole-genome sequence data analysis

**Yonas Kassahun Hirutu**[1] ✉, **Mesert D Bayeleygne**[2], **Adey F Desta**[3], **Tewodros Tariku**[1], **Markos Abebe**[1]

[1] Armauer Hansen Research Institute (AHRI), Ethiopia
[2] Ethiopian Biotechnology Institute (EBTI), Ethiopia
[3] Addis Ababa University (AAU), Ethiopia

Competing interests: YKH none; MDB none; AFD none; TT none; MA none

**Abstract**

Basic bioinformatics training workshop conducted at Armauer Hansen Research Institute (AHRI), Addis Ababa, Ethiopia. This report describes a bioinformatics training initiative started at AHRI aiming to support life science researchers and postgraduates in handling next-generation sequence data.

## Introduction

Institutional initiatives strengthening capacity at Armauer Hansen Research Institute (AHRI) focuses on building a bioinformatics training centre, a next-generation sequencing (NGS) facility and a computing platform to support researchers and postgraduates in Ethiopia. The faculties benefit both ongoing and new project initiatives such as on pathogen evolution, virulence determinants and epidemiology of important pathogens, including *M. tuberculosis*.

The workshop aimed at delivering a practical introduction to NGS data analysis of the M. tuberculosis complex (MTBC) genome.

Every workshop day included 40 minutes presentation, three hours hands-on practical and 20 minutes discussion. The presentation topics were on next-generation sequencing technologies, examples of

**Figure 1.** Participants and trainers of the Workshop

**Article history**
Received: 15 January 2019
Accepted: 21 September 2019
Published: 15 October 2019

sequence data file formats and stepwise description of each NGS data analysis bioinformatics workflow.

The workshop was held at the Armauer Hansen Research Institute (AHRI), Bioinformatics unit, 24-28 November 2018. There were 12 participants from health and biotechnology research institutes in Ethiopia (Figure 1).

## Implementation of the workshop

The workshop started with a lecture on Linux and command-line tools. AHRI's computing room with Bio-Linux workstations was used for demonstrations and practical sessions. Each workstation was assigned for two participants during the workshop. The first-day activity was focused on giving adequate practice time and guidance on command line environment.

Illumina paired-end reads of three Ethiopian *M. tuberculosis* complex (MTBC) strains were selected for the practical sessions. The strains were subsets of fastq files stored at AHRI for which NGS runs were outsourced as part of previous studies in the institute. Nine MTBC genome sequences were downloaded from NCBI[1]. *M. tuberculosis H37Rv* (NC_000962) was used as a reference genome for mapping, variant calling and annotation of MTBC strains.

The practical workflow analysis included quality control of reads with the FastQC tool (Andrews, 2010). Duplicates were removed using a custom Python script. The reads were aligned by Burrows-Wheeler Alignment Tool (BWA) with default parameters for each sample (Li and Durbin, 2009). The aligned results were piped to SAMtools for the conversion of BWA output format to BAM format (Li *et al.*, 2009). Finally, a consensus variant file (VCF) was generated with BCFtools for subsequent SNP analysis (Danecek *et al.*, 2011). The variants identified with SAMtools/ BCFtools were inspected using the Integrative Genomics Viewer (IGV) (Thorvaldsdottir *et al.*, 2013).The reads were also analysed with the MTBseq tool which is a comprehensive pipeline for whole-genome sequence analysis of M. tuberculosis complex isolates with a full workflow functionality implemented in Perl modules (Kohl *et al.*, 2018).

MTBseq analysis tool is designed for MTBC NGS data for reference mapping, variant detection, variant annotation for drug resistance and comparative analysis among the samples. MTBseq tool was used to demonstrate additional approaches and summarise a consolidated view of all the steps followed during the practical workshop. MTBseq annotation outputs of amino acid changes for a known association to antibiotic resistance were used to characterise the samples. Variants information file was used to construct phylogenetic tree using FastTree (Price *et al.*, 2010) (Figure 2).

Finally, a presentation was made on NGS analysis of metagenomics data. The workshop participants had an opportunity to reflect on their training experience and comment on different aspects. Participants' feedback highlighted the need to improve computational power of the workstations and address fluctuations of internet connectivity. The participants also expressed a satisfactory level for acquiring basic skills in handling NGS data.

**Figure 2.** Workshop participants during the hands-on session.

In conclusion, the training coordinators and leaders of the institute have appreciated the success of the workshop. The trainers acknowledged the participants for their active engagement throughout the training week.

## Acknowledgements

The workshop was supported by training and research directorate at AHRI with a funding source from the Swedish International Development Cooperation Agency (SIDA). The authors would also like to acknowledge support from the African Academy of Sciences (AESA) through Tuberculosis Genetics Network in Africa (TBGENAfrica) research project at AHRI.

## References

1. Andrews S. (2010)FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed 20 October 2018)
2. Li H and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform.Bioinformatics **25**: 1754-1760 http://dx.doi.org/10.1093/bioinformatics/btp324
3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078-2079. http://dx.doi.org/10.1093/bioinformatics/btp352
4. Danecek P, Auton A, Abecasis G, Albers CA, Banks E et al. (2011) The variant call format and VCFtools. Bioinformatics **27**: 2156-2158. http://dx.doi.org/10.1093/bioinformatics/btr330
5. Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. BriefBioinform. **14**: 178-192. http://dx.doi.org/10.1093/bib/bbs017
6. Kohl TA, Utpatel C, Schleusener V, De Filippo MR, Beckert P, Cirillo DM, Niemann S. (2018) MTBseq: a comprehensive pipeline for whole genome sequence analysis of Mycobacterium tuberculosis complex isolates. PeerJ6:e5895.http://dx.doi.org/10.7717/peerj.5895
7. Price MN, Dehal PS, Arkin AP (2010) FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. PLoSONE5:e9490. http://dx.doi.org/10.1371/journal.pone.0009490

[1] ftp://ftp.ncbi.nih.gov/genomes/Bacteria/