Day2 session 1.1- Mycobacterium tuberculosis NGS made easy: data analysis step-by-step -

Trainers: Andrea Cabibbe, Andrea Spitaleri, Arash Ghodousi, Christoph Stritt, Daniela Brites, Liliana Rutaihwa

We would like you to think for a couple of minutes about what you have heard. Also, we would like to assess the quality of our teaching in order to improve the quality of this training.Please name 1 to 3 things that you have learned so far and 1 to 3 that you have not fully understood. Thank you!

Day 2 - Webinar "Mapping and Variant Calling in MTBC"

Your name (optional)	Name 1 to 3 things that you have learned so far	Name 1 to 3 things that you have not fully understood
Afsana Akter Rupa	**From this session, I have learned:** 1. How to use Galaxy. 2. How to analyze sequencing data from FASTQ files to variant calling.	I have a question: In Galaxy, most workflows are focused on whole genome data analysis. Can I use FASTQ files generated from targeted sequencing? If so, can I use the same reference genome that I use for whole genome sequencing? Also, can I generate a phylogenetic tree from targeted sequencing (tNGS) data? (Although phylogeny is not included in this session.)
Guy Arnault R MFOUMBI	1. Usage of Galaxy 2. Key points of QC 3. FAstQ analysis	Galaxy vs Seqspher

	4. Concept of SNPs 5. Illumina Sequencing	
Oluwafemi Joseph	1. Use of Galaxy. 2. Concept of Illumina sequencing 3. Getting knowledge about de novo and re-assembly	1. I would like more explanation on understanding vcf (variant calling) 2. And I would like to understand the phylogenetic tree better, i mean the part that trace them back to ancestor
Annisa Meliana Shani	1. The use of Galaxy for analyzing MTBC sequencing data, i understand how to quality check our fastqc data and evaluating its clean reads also doing the variant calling 2. The state of the art of Illumina Sequencing (Short read)	
Imen Bouzouita	Steps of WGS bioinformatic analysis:data cleaning, mapping , variant calling and annotation	Is genomic assembly used in the absence of a Refseq genome?
Micheska EPOLA	1- Illumina sequencing technology 2- How to use galaxy 3- Fastq analysis	
Abraham Ali	Steps for biionformatics	
Marco Pardo Freire	1- How Illumina sequencing technology works 2- The basic workflow to detect variants from sequencing data	1-How would you deal with repetitive zones and variant calling if you do not have a reference genome?. Instead you would have scaffolds of a bacterial genome that you want to compare with new sequencing data. 2-The quality histogram

		obtained with multiQC represents the mean quality of all reads in every position from 1 to 100?
Mark Gutiérrez Pareja	1. Basic training on the Galaxy platform with clear and concise webinars 2. Principles behind Illumina sequencing technology 3. Workflow of the main 4 steps in the WGS analysis of Illumina short-reads (data cleaning, mapping, variant calling, annotation)	
Abraham Ali	1. Illumina Sequencing Technology 2. Sequence analysis steps (Workflow) 3. How to use Galaxy	
Pacome ABDUL ACHIMI	1- The illumina technology 2- How to setup a workflow 3-	
Naphatcha Thawong	1- The use of Galaxy and its trips 2- Short read technology, illumina 3- Bioinformatics analysis of MTB, more detail to update my understanding	
Desmond	I have learned that bioinformatics, i.e., sequencing pipelines, consists of four main steps. These steps include data cleaning, which is performed to assess the quality of the sequence; mapping, where we align our sequence to the reference genome; variant calling, where we identify variants by pointing out differences; and variant annotation.	Kindly throw more light on the ancestral reconstruction. I do not understand

Claudia Gutierrez	1-Training on Galaxy, I will need a lot of practice. 2- Understood fundamentals of illumina platform: Sequencing by synthesis. 3- How to assess the quality of a sequence. 4- Data cleaning, mapping and variant calling.	Please instruct with more detail how to view data on IGV.
Ma. Lyka Padiernos	1. I learned how to use the Galaxy as a simple and user-friendly bioinformatics tool. 2. I learned the general overview of the step-by-step bioinformatics for short-reads sequence specifically for illumina fastq to mapping and variant calling. The tools used in the analysis were also mentioned, which is helpful if we would like to analyze the same data. 3. It's really helpful to tackle the concept of variable and fixed SNPs, because in the concept of infectious disease transmission and monitoring such as in M. tuberculosis, these SNPs will help in identifying lineages, sublineages, which will help in tracking outbreak sources. Fixed SNPs may also be helpful in identifying drug resistance.	
Nabila Ismail	1. How to look at the coverage for large	Is the quality that we used in the tutorial equal to the Phred

	deletions and duplications	score- if we change from 35 to 36, all reads were discarded? This is from the day 1 tutorials linked to what we learnt in day 2.
Ameenah Salihu	1. Invaluable workflow for DNA sequencing 2. Choosing reference genome (H37Rv Vs ancestor sequence) 3. Variant calling	1. Ancestry reconstruction is not very clear to me
Catherine Sacopon	1. I learned how to use Galaxy as a bioinformatics tool, and it is quite easy to move around and use it in our analyses. 2. The sequencing workflow was discussed in a matter that is easy to follow. I learned about what tools to use in data cleaning such as FastQC, how to check the Q scores, and trimming 3. I also learned what ancestral genome reconstruction IS and how it is done. Also how to evaluate if our mapping is good, and how to do variant calling.	
David Adeoye Adedokun	1. Understand the flow of the bioinformatics pipeline from quality control to mapping with reference genome (for re-sequencing), variant calling and annotation. 2. I now better understand how annotation works. Segments of the	1. I am still thinking in my head, how is a phylogenetic tree constructed? 2. How could there possibly be a mutation, say for drug resistance, but without phenotypically showing drug resistance?

	genome are known for a particular function, and thus, mutations in those segments can be linked to those functions. 3. The file formats at different stages of the analysis, including FASTQ, BAM, and VCF	3. How is bioinformatics applied to molecular epidemiology? What are the endpoints, and what could be achieved?
Lilian Nwagbara	1. A good understanding of bioinformatics workflow for Mtb analysis. 2. What the different parts of a fastq file really mean and the difference between single reads and paired end reads. 3. A good understanding of the different file formats and what they contain.	1. How is ancestral reconstruction carried out practically? 2. If you are working with reads of several isolates from different lineages, can you map all reads at once to the reference genome? How many reads can be mapped to a reference genome at once? Are there guidelines for what kind of reads can be mapped?
MUSANA HABIMANA Arsene	1. **How to perform quality control** on raw sequencing reads using tools like FastQC and Trimmomatic, including what quality metrics to focus on (e.g., per base sequence quality, adapter content). 2. **The process of aligning TB reads** to the	**Details behind variant calling tools** (e.g., how GATK or Samtools differ in their algorithms and output quality for *M. tuberculosis*). **Best practices for filtering variants** specifically, how to set thresholds for coverage depth or quality score in a TB context. **How phylogenetic trees are constructed from variant data**, especially how the SNP matrix is generated and cleaned before tree inference.

	reference genome using BWA, and why accurate alignment is critical before variant calling. 3. **Using TB-Profiler** to generate drug resistance profiles and lineage predictions	
Bernice Fumilayor Sawyerr	1. The main difference between exons and SNPs, and the use of Galaxy to access and analyse them from UCSC. 2. The difference between de novo assembly and re-sequencing, as well as their pros, cons, and significance in computational analysis. 3. The processes used in re-sequencing, their tools, pros, and cons in Mycobacterium TB analysis.	
Dania Saeed	1. Significance of using a reconstructed ancestral genome as a reference for studying past mutation events. 2. Understanding the different formats of files generated during sequence analysis e.g. aligned sequences are stored in bam file while variants are stored in vcf format 3. Choosing the right thresholds for allele frequency to avoid calling a sequence error as a variant and	Can allele frequency and depth coverage of a base be used interchangeably?

	computational challenges with sequencing repetitive reads.	
Olga Shavuka	1. You map your short-read sequencing data to a reference genome and then call variants based on differences between your reads and the reference. 2. Certain repetitive genomes are often filtered out to improve mapping and variant accuracy. 3. After variant calling each variant is linked to genes and functions, including potential drug resistance markers or phylogenetically relevant mutations.	1. I got the concept of identifying mixed strains, but I didn't catch how they set cutoffs for calling a minor variant.
Lorraine Boois	1. Annotated variants help detect drug resistance and reveal genetic relationships between *Mycobacterium tuberculosis* strains, allowing researchers to trace transmission patterns and cluster related cases.	2. What is a bit unclear is What defines a SNP as "fixed" versus "variable," and why both categories matter in epidemiological and resistance analyses
Veronica Medrano	1. How to use Galaxy for MTB sequence analysis. 2. Workflow for the WGS bioinformatic analysis (data cleaning, mapping, variant calling, annotation). 3. The concept of variable and	The workflow for Illumina sequences was clear. However, I wonder if the same can be applied for other technologies that also produce short reads, like MGI.

	fixed SNPs.	
		-