Mycobacterium tuberculosis NGS made easy: data analysis step-by-step

We hope that you are enjoying the training, and that all the things that you are learning will be useful in your current and future research!

Please, read the questions, and provide the answers in the table below for **at least one** of the points in 1) 2) and 3)

1. **Mapping and Variant Calling**
    a. After SNP calling, we observe a genomic region that has almost double the sequencing depth than the rest of the genome. What is a possible explanation for this phenomenon?

2. **Clustering and DR prediction**
    a. We find two patients with rifampicin-resistant tuberculosis to be linked by recent transmission (in cluster). Based on our genomic analysis, we also know that rifampicin resistance evolved *de novo* in both patients, as opposed to one patient transmitting an already resistant strain to the other. What specific information did we need to look at to be able to know this?
    b. Would you be able to say whether there are imported cases of tuberculosis? Why?

3. **Phylogenetics**
    a. What is the difference between a rooted and an unrooted phylogenetic tree? Why is the former more informative?

Your Name	Question number	Answer

Afsana Akter Rupa	1(a)	This could indicate a duplicated region in the genome or the presence of a repetitive element. Reads from different locations might be mapping to the same region, artificially inflating the coverage.
	2(a)	We need to examine the specific resistance-conferring mutations in the *rpoB* gene. If both patients have different mutations conferring rifampicin resistance, it suggests independent (de novo) evolution.
	2(b)	Yes, by analyzing the phylogenetic relationships and genetic distances. If a strain is genetically distant from locally circulating strains, it may indicate an imported case.
	3(a)	A rooted tree shows direction of evolution and a common ancestor, while an unrooted tree only shows relationships without direction. Rooted trees are more informative for understanding transmission and ancestry.
	3(b)	Only if we have a well-rooted phylogenetic tree and robust historical data. Otherwise, it's difficult to definitively say a strain is ancestral just based on its position in the tree.

Adegun Oluwafemi Joseph	1a	**It is most likely due to duplication in the bacterial genome.** This region likely has two copies due to a gene duplication event, resulting in twice as many reads mapping to it compared to single-copy regions. The duplications are common in *M. tuberculosis* and can be confirmed by analyzing read-depth patterns
	2a.	If the two patients had *different mutations* in the *rpoB* gene or the *same mutation* on **divergent genetic backgrounds**, this indicates independent evolution(homoplasy) rather than Phylogenetic mutation (transmission of a resistant strain).
	2b.	Yes, I can do this by analyzing the phylogenetic relationships and genetic distances. If a strain is not related to the current transmitting strain in the area, we can classify it as an imported case. If the strain clusters with sequences from another geographic region and is genetically distant from local strains, it suggests an imported case.
	3a.	Rooted trees have a defined ancestral node (root) showing evolutionary direction and time. Unrooted trees only show topological relationships without direction. Rooted trees are more informative because they reveal evolutionary ancestry, direction of mutations, and relative timing of divergence

Mark Gutiérrez Pareja	1a	The most reasonable explanation for this phenomenon is the presence of a **duplicated region**
	2a	Thanks to increased resolution of WGS, looking at the specific mutations that are conferring drug resistance, if we see that strains have **different mutations**, we can conclude that, despite being in the same cluster, thus coming from the same susceptible strain, these strains have likely acquired resistance after being transmitted
	2b	You could by observing the phylogenetic placement of the isolates, that is if the SNP distance is big from the local samples, but I am not sure if you could tell if they are actually imported cases or maybe reactivations
	3a	The main difference is that rooted trees show the direction and sequence of branching events, which allows you to infer ancestry, evolutionary paths, and temporal direction

Marco Pardo-Freire	1a 2a 2b 3	It could indicate the presence of a duplication but also the presence of repetitive regions of the genome mapping to the same sequence. You need to look at the specific mutations conferring that resistance. If the mutation is not the same, they have probably emerged independently, possibly as a consequence of the drug treatment By clustering you can find strains that are not closely related to the ones in a certain region. This could suggest an imported case but you cannot be sure, it also could come from a reactivation or maybe your sampling is not good enough to represent all the variability in your population. The rooted tree contains a common ancestor of all the samples analyzed and thus it is more informative in the sense that it gives directionality to the evolution of the tree

Claudia Gutierrez	1. 2a. 2b. 3.	This could be explained by a duplicated region. You need to look at the specific mutation, these patients should have different mutations that confer resistance to rifampicin. An imported case could be suspected when phylogenetic analysis shows a large distance between the strain in question and the rest of the circulating strains. This could also be due to reactivation or sampling bias. Epidemiological data on whether the patient was residing in another country and matching with sequences from that country could confirm the suspicion. In a rooted tree the most recent common ancestor or an outgroup is included in the analysis. This allows to show the direction and sequence of branching events.

Bruke Gezhagene	1a	The reason for this extra data could be contamination from other bacteria or errors in aligning the sequences from similar genes and a duplication region.
	2a	We conclude that one is the rifampicin resistance gene.
	b	Yes, it also needs to compare local isolates to global reference datasets or known lineage distributions, depending on the epidemiological context.
	3	Identifying imported cases requires comparing local isolates to RefSeq, and rooted trees help us understand how different strains of TB are related, how they evolved, and how the disease spreads. This makes them very useful for studying TB transmission and its development.

Pacome ACHIMI ABDUL	1a 2a 2b 3.	The increased sequencing depth in the specific region can be linked to the presence of duplicated or repetitive elements in the genome We can focus on the specific resistance conferring mutation in the rpoB gene and look at if the two patients have different mutations associated with resistance. The SNP distance within the cluster can also provide the evidence of events Yes we can use the phylogenetic analysis and make a comparative analysis with the local and global strain A rooted tree includes a defined root, representing the most recent common ancestor, and thus shows the direction of evolutionary change from the ancestor to the present An unrooted tree shows the relationships among sequences but does not indicate the direction of evolution — it doesn't identify a common ancestor or the evolutionary path. This allow us to identify ancestral and derived traits is also able to indicate the direction of the evolution

Imen Bouzouita	1a	it could be a repetitive region or an insertion sequence
	2a	Just we need to identify the mutation in rpoB gene conferring resistance : rif resistance evolved de novo means that the two patients had different mutations in rpo B gene conferring RIF-R
	2b	Yes, by performing a phylogenetic analysis: imported cases appear as ' unclustered cases' (but they could be clustered in their countries and thus involved in a recent transmission chains)
	3	In the unrooted tree , the position of the common ancestor was not shown in the tree and the evolution direction is not indicated (it could become rooted if we add an outgroup) The rooted tree gives information about the position of the ancestor , the direction of evolution, distances...

Ma. Lyka Padiernos	1 (a) 2 (a) 3	This may be due to a duplicated or repetitive region where reads map more than once. We looked at the specific mutations in the rpoB gene. Different mutations in both patients mean resistance developed independently. A rooted tree shows the direction of evolution and common ancestor; unrooted trees show topological relationships but do not show direction of evolution. Rooted trees are better for tracing transmission paths.

Nabila Ismail	1a	A genomic region with double the sequencing depth could be due to a gene duplication or a repetitive region (especially with short read sequencing), e.g. PE/PPE genes, where shorter reads are more difficult to map for these genes because of the repetitive sequence and more reads may map to this portion of the gene incorrectly.
	2a	You have to look at three different types of information to confirm this: 1. The two patients must have the same strain, i.e. same lineage 2. The two patients must have strains with a SNP distance falling within the threshold that you have chosen for your study to indicate recent transmission, e.g. 5/10/12 3. The variant calling for rifampicin, must have shown two different SNPs
	2b	We would be able to say if there were any imported cases, if there was no sampling bias, which is difficult to determine for a specific population. If there is no sampling bias, we could say that there was “likely importation” if a case was unclustered and did not represent the lineages circulating in a particular region.
	3a	The additional information that a rooted tree provides is both direction and sequence of branching events, whereas an unrooted tree can not show when a particular clade or tip evolved or arose from a diverged early or later compared to other clades.

Ameenah Salihu	1a	A duplication event may have occurred in that particular region. Another possible reason is contamination by another organism.
	2a	The gene mutations should be looked out specifically especially the rpoB gene which is linked to rifampicin resistance. If the mutations differ in both patients, it could be suggestive of de Novo resistance in each patient.
	2b	Yes, but you will need both epidemiology and genomic data to conclude this.
	3a	Rooted tree shows how evolution occurred from ancestors to descendants structurally. The unrooted tree does not indicate the direction of evolution, has no common ancestor but instead relates taxa. The rooted is more informative as it enables us to track evolutionary changes (including mutations), supports understanding genomic transmissions and outbreak investigations. History sheds a lot more light and enables us to track patterns when easily

Lilian Nwagbara	1a)	It is possible that there are repetitions in the reference genome i.e. duplicate regions, and the reads are aligned >1 times i.e. multi mapped reads.
	2a)	We need to determine: i) that the rifampicin-resistant tuberculosis in both patients belong to the same strain through strain typing. ii) the specific mutations causing rifampicin-resistance in both patients. If the same mutations are responsible for the drug-resistance effect, then the tuberculosis could be linked to a recent transmission. But if different mutations are responsible for the rifampicin-resistance effect, then the resistance could be treated as *de novo* evolution.
	2b)	Yes. If there is no sampling bias and genetic distance is greater than the SNPs threshold for determining genotype, then I would say there are imported cases of tuberculosis.
	3b)	Rooted phylogenetic trees show the direction and sequence branching of evolutionary events from a common ancestry but unrooted phylogenetic trees do not. Rooted phylogenetic trees are more informative because they provide a historical relationship and can show the evolutionary history of an organism.

David Adeoye Adedokun	1a	Gene duplication, insertion or repetitive elements. All of these increase the read depth artificially.
	2a	We would look at the nature of the resistance mutation. If they have the same resistance mutation and similar genomes, it would suggest transmission. Different mutations would suggest independent resistance development.
	2b	Yes, imported cases can be suggested if some strains are not part of local transmission clusters.
	3a	A rooted tree shows the direction of evolution from a common ancestor. An unrooted tree shows relationships but no direction of evolution. A rooted tree is more informative because it helps us understand how evolution progressed over time.

Veronica Medrano		1a. A possible explanation for the double sequencing depth in a particular region is that the region is duplicated within the genome. Two copies will cause twice as many reads that are mapped which in turn increases the depth. 2a. It is necessary to evaluate the specific mutations that cause the resistance to rifampicin. Since we know that the two patients are in the same cluster, it is expected that the two strains have different mutations which means a de novo evolution of rifampicin resistance. 2b. It is possible through analysis of the phylogenetic placement of the isolates and comparing it with the cluster of known local circulating strains. 3a. The difference is in the representation of evolutionary relationships and directionality. An unrooted tree shows relationships among sequences but not the direction of the evolution, it does not indicate a common ancestor or the order of divergence. Therefore, an unrooted tree is useful for visualizing similarity among sequences but cannot infer ancestry. On the opposite, a rooted tree will have a defined root that represents the most common ancestor, showing the direction of evolution. It allows an estimation of divergence times and evolutionary pathways. This type of tree is more informative because it shows ancestral relationships, timing and order of divergence events and the evolutionary history of mutations or traits.

Dania Saeed	1a	This could be indicative of a repetitive region, or gene duplication or insertion within the genome that leads to the sequencing depth being doubled.
	2a	We would analyze the mutation in the rpoB region of the two patients, different variants in the RIF- resistance conferring gene would be indicative of denovo resistance in strains from the same cluster.
	2b	To conclude that these are imported cases, we would need to know the travel history of the patients and analyze the SNP distance between these strains and others from our sample. As these strains are indicated to be part of a cluster I do not think they are imported but temporal spatial history should rule out any sample bias or bias due to allele threshold used.
	3a	A rooted tree will give us information about the sequence and direction of evolutionary/mutation events. Therefore becomes very important when studying or investigating outbreak or transmission events.

Olga Shavuka	1. a)	-It could be due to duplication, the same sequence appears more than once in the genome, so reads map to both places.
	2. a)	-We need to look at the exact mutations in the rpoB gene , different mutations mean separate events.
	b)	-Yes, we could tell if there are imported cases by looking at strain lineages. If a strain is genetically very different from local strains, it may be imported.
	3. a)	-A rooted tree shows the direction of evolution (from ancestor to present), while an unrooted tree only shows relationships. Rooted trees are more informative because they help us understand transmission and mutation order.

Bernice Fumilayor Sawyerr	1a.	This could result from the presence of duplicated regions in the genome. During read alignment, the reference genome has only one copy. If a region in the genome is duplicated, both copies map to the same location in the reference genome, causing increased coverage.
	2a.	The information needed to determine that rifampicin resistance evolved de novo in both patients includes the whole genome sequence of Mtb isolates from both patients. Phylogenetic analysis to verify that the isolates from patients fall within the same transmission events and identify resistance mutations in the rpoB gene in each patient isolate. If the patients have different mutations in the rpoB conferring resistance, this supports independent(de novo) evolution of resistance.
	2b.	Yes. Imported TB cases can be predicted if there are WGS data, global phylogenetic data, and relevant epidemiological data. These can allow us to distinguish local transmission and importation with high confidence.
	3a.	A rooted phylogenetic tree has a common ancestor from which all other species evolved, while unrooted phylogenetic trees show the relationship between species. A rooted phylogenetic tree is more informative because it reveals the direction of evolution.

