**aDNA Data Analysis II: analyzing *Mycobacterium leprae* genomes from NGS data**

**Scenario:** You have received NGS *M. leprae* capture data for three skeletons from Sant Llàtzer, a medieval leprosarium in Barcelona: **UF104**, **UF703**, **UF801**. The genetic diversity of *M. leprae* strains across medieval Europe and within leprosaria was high. No one has recovered ancient *M. leprae* genomes from Spain**,** so you would like to know if leprosy infections at Sant Llátzer were seeded by a diversity of lineages.

**Part 1: Using Qualimap to determine reference coverage and depth**

Since you have already assessed these raw data (aDNA Data Analysis I), you will begin this practical by using Qualimap to assess genome coverage and depth from .bam files that were generated by mapping reads to the *Mycobacterium leprae* reference genome. I have run the tool for you, and you have the outputs: a .pdf report and a .txt file for each sample.

* Using the .pdf reports in **1.Qualimap outputs,** fill out the following columns in the **M. leprae mapping report**: Q25 mapped reads after duplicate removal, Avg length of mapped Q25 reads, Mean Cov, and SD Cov

Now that you have read through the .pdf reports, open the Qualimap .txt files. Much but not all the information from the .pdf is reported here as well. Text file outputs are useful because you can write code that will loop through .txt files and pull out the information you need. Most geneticists automate their mapping reports so that they do not have to manually pull out each piece of information.

* Complete the next two columns in your **Sant Llátzer mapping report:** **%ref covered at >=1x** and **%ref covered at >=5x** using the Qualimap .txt files.

Some of the metrics you need to assess your data are not generated by a tool, so you need to generate them yourself. **Cluster factor**, also known as **library complexity**, is informative about the number of unique fragments in your library. It tells you on average how many times a read was sequenced. For example, cluster values over 2 indicate that on average, every fragment has been sequenced twice, so the library is probably exhausted of DNA from this organism. Depending on how deeply you sequence and if you’ve done targeted enrichment, cluster factors can be huge, between 20 and 30.

* Calculate cluster factor for each sample using the formula in the cell and the values in the columns you’ve already filled out.

Endogenous DNA frequency can be used in tandem with cluster factor to decide whether to sequence your library more deeply or determine how efficient your capture enrichment was. Since these data have been procured through targeted capture enrichment, endogenous DNA frequency will show us how efficient the capture experiment was.

* Calculate the frequency of endogenous DNA (*M. leprae* in this case) for each sample using the formula in the header cell and the values in the columns you’ve already filled out.

**Part 2: MapDamage plots**

There are several tools that produce iconic aDNA “damage plots,” but here we are using the tool MapDamage because it also rescales base qualities, which other damage profiling tools do not. Open **2.MapDamage outputs**.

Open the **fragment misincorporation plots** for each sample. The four plots at the top indicate base (A, C, T, and G) frequencies within the read (within the grey square) and outside of the read. **What are these plots illustrating?** Hint: depurination.

The bottom two plots are the most common damage plots that you may have seen before. The red line charts C to T substitutions, and the blue line charts G to A substitutions.

* From the bottom two plots, estimate the values for the last four columns in your **M. leprae mapping report**.
* Based on these plots, do all the samples have authentic aDNA?

**Part 3: Basics of phylogenetic inference**

Since we will be building phylogenies to determine which lineage *M. leprae* strains at Sant Llàtzer belong to, we must understand how phylogenies are made and how they represent evolutionary relationships among organisms.

The European Molecular Biology Laboratory (EMBL) offers introductory trainings on a variety of topics. Go to EMBL’s phylogenetics overview, read the specified sections, and take the short quiz to test your understanding.

**EMBL phylogenetics overview**: <https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics/what-is-phylogenetics/>

**Relevant sections:**

-What is phylogenetics?

-Why is phylogenetics important?

-What is a phylogeny?

-Major stages in phylogenetic analyses

-Quiz: test your knowledge

**Part 4: Multiple sequence alignments (MSA)**

1. Start with a question
2. Identify a model and parameters that could answer the question
3. Collect sequence data that would help to answer the question
4. Identify the orthologous sequences
5. Align sequences
6. Estimate tree and other parameters given the data and model
7. Estimate the error associated with the tree and/or parameter estimates
8. Does it answer your question?
This can all lead to new biological insight. 

**Figure 1**. Outline of stages in phylogenetic analysis from EMBL phylogenetics overview

This is how our research maps on to EMBL’s stages in phylogenetic analysis (Fig. 1):

1. Was *M. leprae* genetic diversity high at Sant Llàtzer?
2. Our inference models:
   1. **Maximum parsimony (MP)**: character-based method that prioritizes the phylogeny with the fewest state transformation (substitutions)
   2. **Maximum likelihood (ML)**: generalized time reversible (GTR) model of nucleotide substitution; most general model that allows all transitions and transversions to occur with independent frequencies
3. Sequencing data from Sant Llàtzer residents
4. Since *M. leprae* has a relatively short genome (~3.3 million bp), is haploid (only one copy of the genome that does not recombine), and exhibits little horizontal gene transfer (trading of genomic elements with other bacteria), we are using variant sites across the entire *M. leprae* genome.
5. Multiple-sequence SNP alignments (explained in Part 4)
6. Building MP and ML trees in MEGA 11
7. What is the bootstrap support? Do the tree topologies match?
8. To be determined
9. The goal!!!

**Part 4: What data do we use to build phylogenies?**

The file type we will use to build phylogenies is called *.fasta*. We discussed this file type during aDNA Data Analysis I. It is a one-dimensional file that only contains the nucleotide bases assigned to a sample. A .fasta file can contain sequences from many different organisms/samples, but each new sequence must be signified with a header that begins with >.

Open the folder **4.Multi-sequence SNP alignments.**

Open **M.leprae Sant Llatzer MSA.fasta** in NotePad++.

* How many *M. leprae* samples are represented in this file? (hint: use ctrl + f to search for the special character that beings each header)

Each base pair stands for a **single nucleotide variant** (aka SNV/SNP) among the *M. leprae* samples. Our .fasta file comprises SNVs/SNPs from across the whole genome of our samples. *N* means that position was not sufficiently covered by reads, so the variant caller could not tell if the site was the **reference allele** or a **mutation**.

A **multiple sequence alignment (MSA)** contains aligned positions for all the variant sites among the samples. Since samples will not have SNVs/SNPs at every position of their genome, many of the base calls in our MSA.fasta are reference calls.

Pick a sample and use your cursor to highlight all the base pairs for that sample.

* How many SNV/SNPs differentiate these samples?

Let’s visualize the same SNV/SNP across three different data types.

Open the following files:

1. **M.leprae SNV/SNP table.xlsx**
2. **M.leprae Sant Llatzer MSA.fasta** (should already be open)
3. **UF801\_leprae.bam** in IGV (make sure to open the *M. leprae* reference genome: Genomes -> Load Genome from File -> M\_leprae\_TN.fasta)

**M.leprae SNV/SNP table.xlsx** is a table of every variant position called among the samples.

* What is the genomic position of the first variant?
* What is the reference allele at this position?
* What base does sample UF801 have at this position?

Let’s view this same variant in **M.leprae Sant Llatzer MSA.fasta.**

* What is the first base in the SNP/SNV alignment of sample UK801?
* Is it what you expect?

Finally, let’s look at the data we have about this variant before it was called and “flattened” into a consensus format in the .fasta file. Look at **UF801\_leprae.bam** in IGV, and zoom all the way in using the + button in the top right corner.

* What are you visualizing in IGV?
* What variant is identifiable at base 73?
* How many reads are covering base 73?
* Do you think there is strong evidence that this is true variant?

Repeat the steps above for the second variant in **M.leprae SNV/SNP table.xlsx** (the variant in row 3). A “.” indicates the sample has the reference allele at that position.

**Part 5: Building MP trees in MEGA 11**

Now let’s build a phylogeny using MEGA 11. First we will build a **maximum parsimony tree** of comparative *M. leprae* samples I have chosen to represent the known worldwide genetic diversity.

Open the folder **5.Maximum parsimony.** You will be using the file inside.

1. Load the data into MEGA 11 by clicking on “Data” as shown below and loading the file **M.leprae comparative.fasta**

Graphical user interface, text, application

Description automatically generated

1. You will have a series of prompts. Respond as follows:
   1. Analyze -> Nucleotide Sequences -> Ok -> No (to protein coding questions)
2. Once the data are loaded, view the alignments by clicking on the “TA” square as highlighted below

Graphical user interface, text, application

Description automatically generated

1. Each of the 40 samples corresponds to a row, and each column is a position in the SNP/SNV alignment. Just like the SNP/SNV table, “.” Indicates reference allele.
2. In the bottom left corner, you can see how many variant sites are included in this alignment (1,336).
3. You can close the data window or leave it open.
4. Go back to the main window and select “Phylogeny” and “Construct/Test Maximum Parsimony Tree(s)” as shown below

Graphical user interface, text, application

Description automatically generated

1. Select “Yes” that you would like to use the currently active data.
2. You will see a lot of parameter options, but you only need to change one field. Change “Test of Phylogeny” to “Bootstrap method” as shown below and select “Ok”. The run should be pretty fast. You’ve (probably) made your first phylogenetic tree!

Graphical user interface, text, application

Description automatically generated

Let’s investigate the phylogeny. Select “Layout” on the left side and adjust the “Tree Width” to make the branches and names easier to see. The samples are named in this construct: **[ID] [Country/state of isolation] [Year isolated in CE] [host if not human]**

Above the tree select the “Bootstrap Tree” view. This will add bootstrap values to the tree. To make this tree, MEGA 11 just iterated through 100 different replications to calculate parsimony (a value set in the screenshot above). A **bootstrap** value is the percentage of replicates that resulted in this organization. For example, *M. lepromatosis* shared a node with 516, 518, CM1, Kanazawa, Kyoto-1, S10, and Jorgen-507 in 100% of replicates (see blue dot below).

Diagram

Description automatically generated

Bootstraps are used to determine the probability of the final tree. Values of >90 offer very strong support.

The maximum parsimony algorithm, however, doesn’t account for everything we know about our data. We know the *direction of evolution* that the maximum parsimony analysis doesn’t account for. We need to manually identify the outgroup *M. lepromatosis* to root the tree so that we can see the evolutionary history of each lineage.

To root the tree, right click around the tree and select “Root Tree”. Next click on the branch leading to *M. lepromatosis*.

Diagram, schematic

Description automatically generated

This is the tree topology that we will add our samples UF104, UF703, and UF801 to. You can minimize this tree window. Let’s add our samples and see what happens!

**Part 6: Building ML trees in MEGA 11: garbage in, garbage out**

Now we will build **maximum likelihood** (**ML**) trees in MEGA 11. ML is considered a more robust technique for phylogenetic analysis than maximum parsimony. ML approach searches for the tree that has the highest probability of giving rise to the observed data.

A ML approach is better at accommodating branches (lineages) that evolve at different rates. This is relevant for pathogens, because their genomes evolve relatively quickly depending on infectivity, transmissibility, and access to hosts. Further, using a ML approach, branch lengths are scaled by the number of substitutions, so you can visualize heterogeneity in evolutionary rates.

Let’s build a ML tree using our samples. Open folder **6.Maximum likelihood**. You will be using the files inside.

Like before, go to Data -> Open a File/Session -> **M.leprae SantLlatzer MSA.fasta** -> Yes -> No -> Analyze -> Ok -> No

To make the maximum likelihood tree:

Phylogeny -> Construct/Test Maximum Likelihood Tree -> Yes use the active data

Graphical user interface, text, application, chat or text message

Description automatically generated

Set up your run to look like the following:

Table

Description automatically generated with medium confidence

This will take about 10-15 minutes to run. While you are waiting, you can look over a version of the MP tree you made that has branch labels added: **M.leprae MP tree with branch labels.pdf.**

* Which branches do the ancient samples belong to?

Once the ML run finishes, look at the resulting topology.

* Does it look strange to you? In what way?

Let’s root the tree and see how that changes things. Right click on the whitespace -> Root Tree -> click on *M. lepromatosis*.

UF104 is still behaving strangely. It has a very long branch length. Let’s inspect the alignment data to see if we can find out why.

Go to the main MEGA 11 window and select the TA square to bring up the alignments. Scroll down to sample UF104 (it should be row #40).

* What do you notice about the variant calls for UF104? What does it mean?
* Why do you think the branch length is so long?

Go back to the **M. leprae mapping report.**

* What information from the mapping report suggests that UF104 does not contain authentic *M. leprae* aDNA?

Let’s remove UF104 from our MSA and make another ML tree.

Open **M.leprae\_SantLlatzer\_MSA.fasta** and find sample UF104. Highlight the header and all the variant calls and erase them (back space). Save the edited file.

Upload the new MSA into MEGA 11 as before: Data -> Open File/Session -> modified **M.leprae\_SantLlatzer\_MSA.fasta** -> Yes -> No -> Analyze -> Ok -> No

On the main MEGA 11 window, select Phylogeny -> Construct/Test Maximum Likelihood Tree -> Yes

Your ML run setup should like this (same as above):

Table

Description automatically generated with medium confidence

This will take another 10-15 minutes to run.

Once it is finished, don’t forget to root it. Adjust Layout -> Tree Width to make the tree easier to interpret.

It is a good indicator that your phylogenetic inference reflects the true evolutionary relationships when the unrooted and rooted trees have the same topology.

* Do the rooted and unrooted ML tree have the same topology?
* In which branch do the Sant Llátzer M. leprae genomes fall?
* What other samples are in this branch? How old are they and where were they sampled?

More information on interpreting phylogenetic trees of pathogens -> https://artic.network/how-to-read-a-tree.html