

Manual cgMLST *M. abscessus*

Author: Margo Diricks ([mdiricks@fz-borstel.de](mailto:mdiricks@fz-borstel.de) or diricks\_margo@hotmail.com)  
🡪 Please let me know if you need any help or if you find any errors! 😊

Contents

[1 Abbreviations and Definitions 2](#_Toc103203539)

[2 Design and application of the cgMLST scheme 2](#_Toc103203540)

[3 Frequently asked questions 2](#_Toc103203541)

[3.1 How much good cgMLST targets do you expect/require 2](#_Toc103203542)

[3.2 Can you determine the subspecies of an unknown isolate with cgMLST analysis? 2](#_Toc103203543)

[3.3 Are their DCC specific alleles? 6](#_Toc103203544)

[3.4 Can this scheme also be used for other mycobacteria? 6](#_Toc103203545)

# Abbreviations and Definitions

**cgMLST:** core genome multi-locus sequence typing

**cgMLST targets:** loci, i.e. coding sequences (CDS) with known annotation

**Good quality** **cgMLST targets**: loci with (i) the same length as reference genes +/- 3 triplets, (ii) no ambiguities (e.g. N), (iii) no frame shifts compared to reference genes, (iv) at least 90% identity to reference sequence and (v) valid start and stop codons and no internal stop codons. Pairwise distances between two isolates were calculated as the amount of cgMLST loci with a different allele number, ignoring missing (bad quality or absent) cgMLST loci.

**DCC**: dominant circulating clone (Ruis et al. 2021; Bryant et al. 2016)

**MabA**: Mycobacterium abscessus subsp. abscessus

**MabB**: Mycobacterium abscessus subsp. bolletii

**MabM**: Mycobacterium abscessus subsp. massiliense

# Design and application of the cgMLST scheme for M. abscessus

All details can be found in Diricks et al. 2022 (preliminary doi: 10.21203/rs.3.rs-1482309/v1).

# Frequently asked questions

## How much good cgMLST targets do you expect/require

Several researchers have considered a cgMLST stable if at least 95% of the cgMLST genes are present in all or most strains (Neumann et al. 2019; Ruppitsch et al. 2015; Ghanem and El-Gazzar 2018). Preferably, this should be checked using a large strain collection spanning the whole diversity of a species, including strains that were not used to define the cgMLST scheme. In SeqSphere+, the amount of present cgMLST genes are indicated in the column “%Good cgMLST targets”. This value is colored yellow between 90 and 95% and red below 90%, as a warning for bad quality or diverged genomes. Note that including bad quality genomes can be problematic to reconstruct an accurate phylogeny. By default, SeqSphere+ suggest to exclude genomes with less than 90%.

We performed cgMLST analysis in SeqSphere+ on 1,797 isolates including 1,110 strains belonging to MabA, 563 to MabM and 124 to MabB (Diricks2022; Additional file 1: TableS1). For 1,786 out of 1,797 (99.4%) datasets, more than 95.0% good cgMLST targets were found and for 1,796 (99.9%) more than 90% of the cgMLST genes were present, indicating a stable core genome applicable for all Mab strains.

## Can you determine the subspecies of an unknown isolate with cgMLST analysis?

Definitely!   
Subspecies can be assigned based on the smallest genetic distance using reference genomes for each subspecies. For example, a sample should be assigned to subspecies massiliense if the genetic distance towards the reference genome of massiliense (e.g. type strain JCM 15300; accession number

NZ\_AP014547.1) is smaller than the genetic distances towards the reference genomes of bolletii (e.g. type strain BD; NZ\_AP018436.1) and abscessus (e.g. type strain ATCC19977; NC\_010397.1).

Protocol for SeqSphere+ users:

1. Download following reference genomes (assemblies; FastA) and determine cgMLST allele profiles in SeqSphere+

Table 1 Example of complete assemblies available for each M. abscessus subspecies type strain. 1 % Good cgMLST targets: as defined and determined in SeqSphere+ v.8.3.1.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Accession number*** | ***Strain*** | ***subspecies*** | ***DCC*** | ***ST*** | ***%Good targets (cgMLST)1*** | ***Source*** |
| NC\_010397.1  (GCF\_000069185.1) | ATCC19977T  (CIP 104536T) | abscessus | 1 | 5 | 100 | Ripoll\_2009 |
| NZ\_AP014547.1  (GCF\_000497265.2) | JCM 15300  (CIP 108297T) | massiliense | / | 63 | 98.3 | Sekizuka\_2014 |
| NZ\_AP018436.1  (GCF\_003609715.1) | BD  (CIP108541T) | bolletii | / | 71 | 98.9 | Yoshida\_2018 |

1. Open SeqSphere+ and determine cgMLST profiles of reference genomes using Pipeline Mode with the cgMLST scheme for M. abscessus (2904 target loci) in SeqSphere+ (Only needs to be done once).

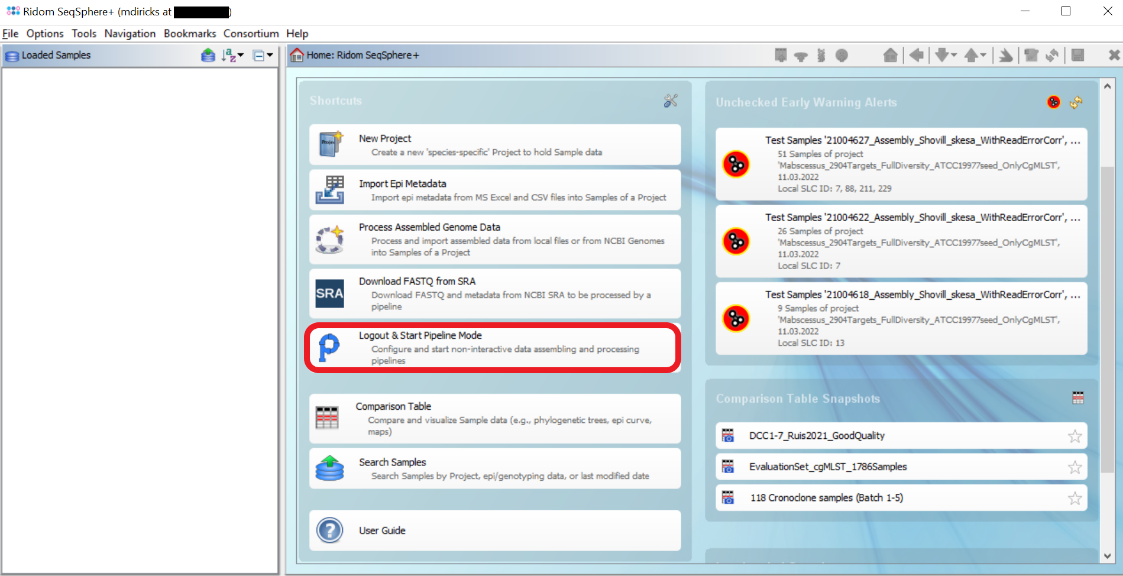


Figure 1 Main window of SeqSphere+.

1. Similarly, determine cgMLST allele profiles of your samples in SeqSphere+
2. Open a comparison table of your samples together with the reference genomes (Click on “Search Samples” in main window 🡪 Select the project in which you have processed the samples and copy e.g. an excel column containing all sample IDs including references in the search field of the “Field and Tag Criteria panel”, using the “In list” search option.

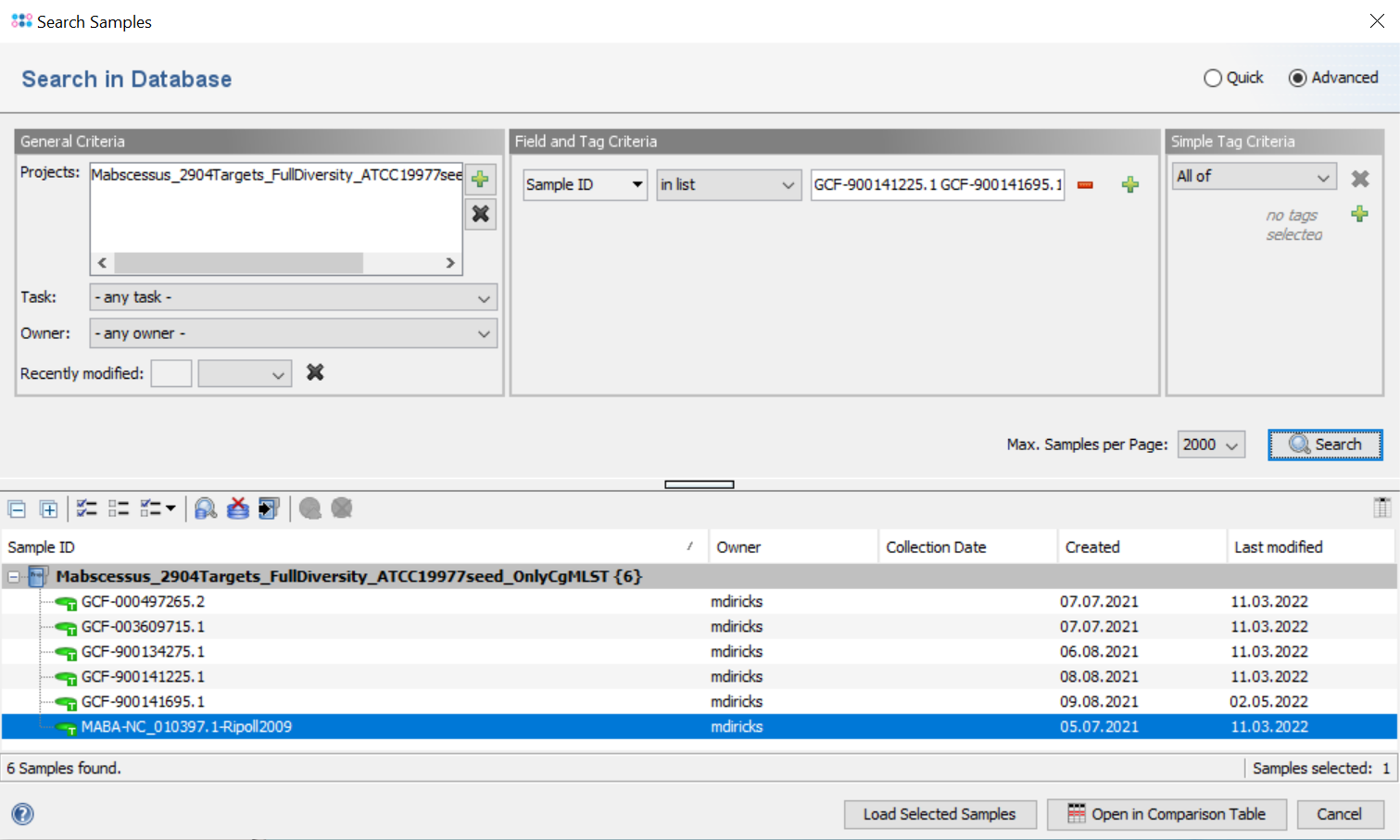


Figure 2 Search Samples window in SeqSphere+.

1. Select all samples and click on „Open in Comparison Table” at bottom of the window.

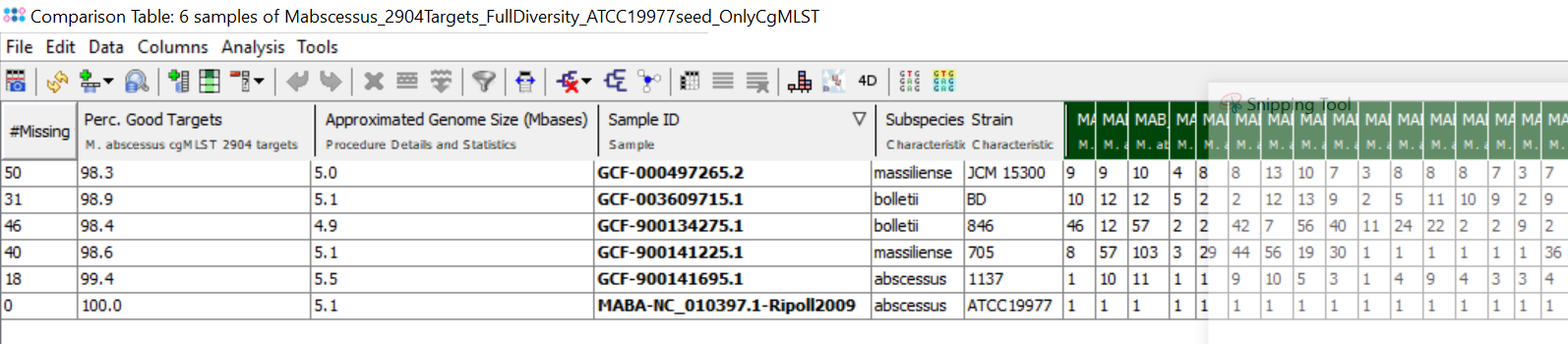


Figure 3 Comparison table window in SeqSphere+.

1. Check if all samples have more than 95% good cgMLST targets (or at least more than 90%)
2. If only few samples, you can determine the subspecies manually/visually using a minimum spanning tree: In Comparison Table click on Analysis > Minimum spanning tree.

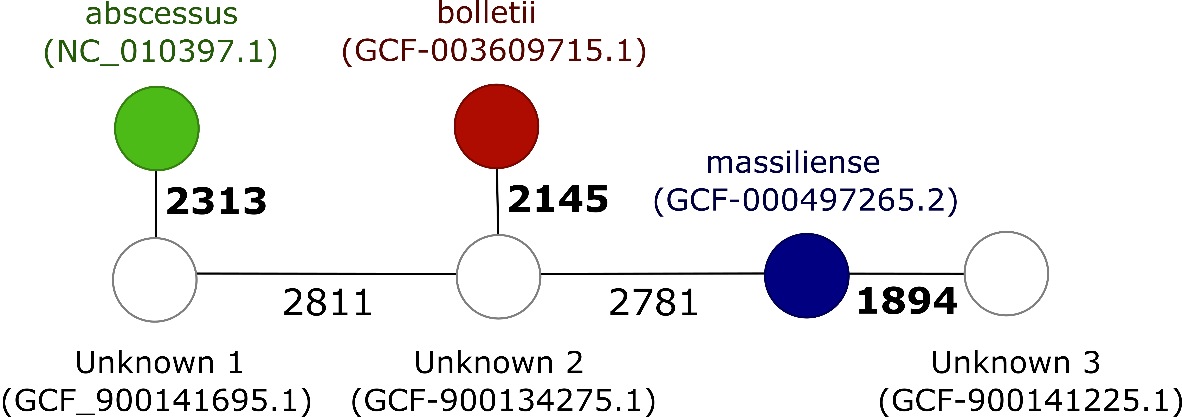


Figure 4 Minimum spanning tree in SeqSphere+. Pairwise distances (allele differences) are displayed on the branches. Subspecies is known for three type strains (colored). Unknown 1 would be classified as abscessus as the distance is minimal towards the abscessus type strain (highlighted in bold). Similarly, unknown 2 and 3 would be classified as bolletii and massiliense, respectively.

1. Alternatively, and more suitable for larger sets of new samples, you can export the distance matrix from within SeqSphere+ as excel file: In the comparison Table window click on Analysis > Distance matrix > Leave Default values > OK > Export matrix .

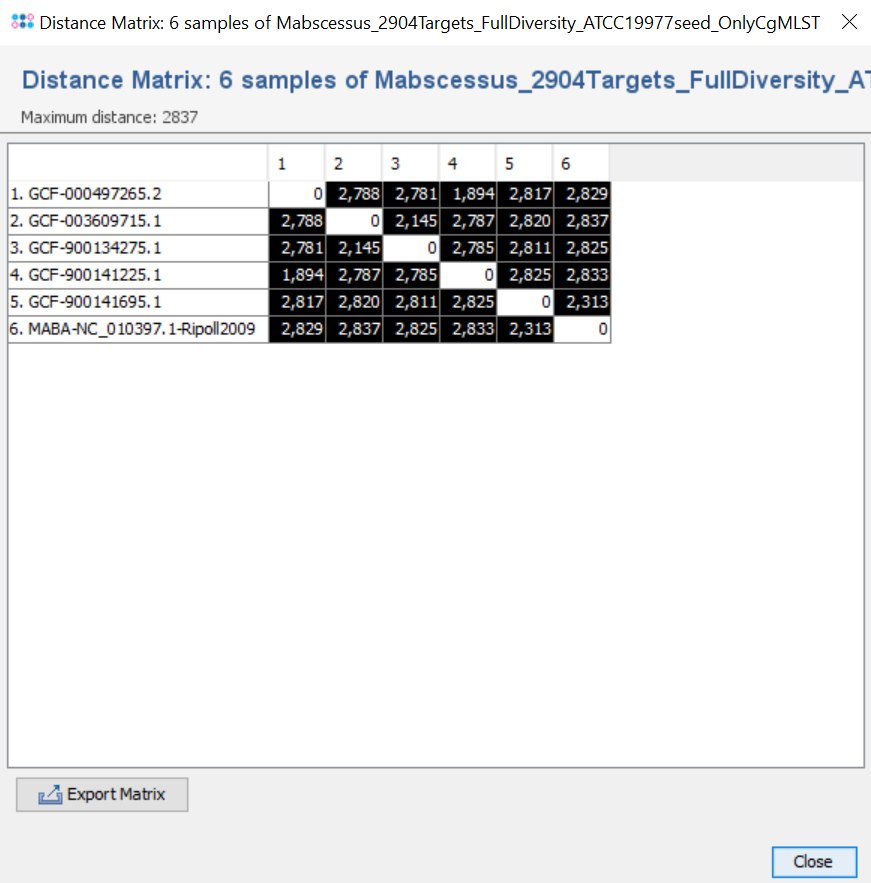


Figure 5 Distance matrix window in SeqSphere+.

1. Open this file and copy the columns corresponding to the three reference sequences to another excel file/sheet. Add an additional row to indicate which colums (i.e. references) are which subspecies (row one in Table 2). Determine the subspecies in column 5, row three (cell E3) based on the minimum genetic distance using following formula: =IF(COUNTIF(B3:D3;MIN(B3:D3)) =1; INDEX($B$1:$D$1;1;MATCH(MIN(B3:D3);B3:D3;0)); "Undecided"). Drag the bottom right corner (or double click) down to other cells in column 5 (E) to determine the subspecies of the other samples.

Table 2 Exported distance matrix and adjustments to let formula work. \*formula for this cell would be =IF(COUNTIF(B3:D3;MIN(B3:D3)) =1; INDEX($B$1:$D$1;1;MATCH(MIN(B3:D3);B3:D3;0)); "Undecided")

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *abscessus* | *massiliense* | *bolletii* |  |
| Sample ID | NC\_010397.1 | NZ\_AP014547.1 | NZ\_AP018436.1 | Subspecies |
| GCF-900141225.1 | 2833 | **1894** | 2787 | *massiliense\** |
| GCF-900141695.1 | **2313** | 2817 | 2820 | *abscessus* |
| GCF-900134275.1 | 2825 | 2781 | **2145** | *bolletii* |

## Are their DCC specific alleles?

I.e. are there loci for which all members of a DCC have the same allele number and for which this allele number is not found in isolates not belonging to this DCC?

Using the evaluation set (n=1,786) samples with >95% good cgMLST targets, we searched for DCC specific alleles. However, for two clades (DCC5 and 6) we could not find any.

To be continued.

## Can this scheme also be used for other mycobacteria?

No, this scheme can only be used to type isolates belonging to *Mycobacterium abscessus.*

For other mycobacterial species, not enough good cgMLST targets are found (because most loci are not present or too diverged compared to *M. abscessus*)

Table 3 cgMLST results for other mycobacteria

|  |  |  |
| --- | --- | --- |
| **Sample ID** | **Perc. Good Targets** | **Approximated Genome Size (Mbases)** |
| Mycobacterium\_avium\_subsp.\_avium | 0.1 | 4.9 |
| Mycobacterium\_celatum | 0.1 | 4.7 |
| Mycobacterium\_chelonae | 8.2 | 5.0 |
| Mycobacterium\_chelonae\_subsp.\_gwanakae | 8.3 | 5.1 |
| Mycobacterium\_fortuitum\_subsp.\_fortuitum | 0.3 | 6.3 |
| Mycobacterium\_intracellulare\_subsp.\_chimaera | 0.1 | 6.1 |
| Mycobacterium\_intracellulare\_subsp.\_intracellulare | 0.1 | 5.4 |
| Mycobacterium\_intracellulare\_subsp.\_paraintracellulare | 0.1 | 5.5 |
| Mycobacterium\_intracellulare\_subsp.\_yongonense | 0.1 | 5.7 |
| Mycobacterium\_kansasii | 0.1 | 6.6 |
| Mycobacterium\_malmoense | 0.1 | 5.3 |
| Mycobacterium\_simiae | 0.1 | 5.8 |
| Mycobacterium\_smegmatis | 0.2 | 7.0 |
| Mycobacterium\_tuberculosis\_L1 | 0.1 | 4.4 |
| Mycobacterium\_tuberculosis\_L2 | 0.1 | 4.4 |
| Mycobacterium\_tuberculosis\_L3 | 0.1 | 4.4 |
| Mycobacterium\_tuberculosis\_L4 | 0.1 | 4.4 |
| Mycobacterium\_tuberculosis\_L6 | 0.1 | 4.4 |