Forschungszentrum Borstel
Leibniz Lungenzentrum

Manual cgMLST *M. abscessus*

Author: Margo Diricks (mdiricks@fz-borstel.de or diricks_margo@hotmail.com)
→ Please let me know if you need any help or if you find any errors! 😊

Last edit: 02.06.2022

Contents

1 Abbreviations and Definitions

Mab: *Mycobacterium abscessus*

MabA: *Mycobacterium abscessus* subsp. *abscessus*

MabB: *Mycobacterium abscessus* subsp. *bolletii*

MabM: *Mycobacterium abscessus* subsp. *massiliense*

cgMLST: core genome multi-locus sequence typing

cgMLST distance: pairwise distances between two isolates are calculated as the amount of cgMLST loci with a different allele number, ignoring missing (bad quality or absent) cgMLST loci.

cgMLST targets: loci, i.e. protein coding sequences (CDS) with known annotation. Non-coding genes (e.g. rRNA genes) and intergenic regions are not considered for cgMLST analysis.

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.

DCC: Dominant circulating clone. A group of closely related isolates found in multiple countries or on multiple continents. Three main *Mycobacterium abscessus* DCCs (i.e. MabA Cluster 1 and 2, and MabM Cluster 1) were first described by Bryant and coworkers (Bryant et al. 2013, 2016) based on cgSNP analysis. Later on, Ruis and coworkers have defined four additional DCCs based on cgSNP/hierbaps analysis (DCC1-7) with isolates from at least 20 individuals from multiple continents (DCC1-7) (Ruis et al. 2021). These DCC strains have been isolated from both CF and non-CF patients across the whole globe, are thought to have emerged around 1960 and have been associated with increased virulence, higher rates of resistance, and worse clinical outcomes compared to unclustered isolates (Bryant et al. 2013, 2016; Ruis et al. 2021). Bronson and coworkers, on the other hand, have identified 11 DCCs (A1-A7 and M1-M4) with 10 or more isolates from at least two countries using a 500 SNP threshold for hierarchical clustering (Bronson et al. 2021). To the best of our knowledge, there are no DCCs defined for MabB.

2 Design and application of the cgMLST scheme for *M. abscessus*

All details regarding scheme creation for Mab can be found in Diricks et al. 2022 - Delineating Mycobacterium abscessus population structure and transmission employing high-resolution core genome multilocus sequence typing (preliminary doi: 10.21203/rs.3.rs-1482309/v1).

In summary, we used MabA type strain ATCC19977 (NC_010397.1 (Ripoll et al. 2009)) and 96 additional publicly available assemblies from a genetically diverse and global set of Mab isolates to define a hard core genome with SeqSphere+ software (https://www.ridom.de/seqsphere/). The scheme creation set included representatives for all subspecies, for seven previously defined DCCs as well as non-DCC strains, which were collected in at least 11 different countries. The resulting cgMLST scheme consists of 2,904 loci, representing 59% of the gene set from MabA type strain ATCC19977.

The genes comprised in the cgMLST scheme can be identified in new samples using different algorithms (i.e. from assemblies using BLAST or from reads using kmer mapping), depending on the cgMLST software used (Jolley and Maiden 2010; Palma et al. 2022; Silva et al. 2018; Feijao et al. 2018). The cgMLST software will also automatically assign an allele number to each of the genes, using a fixed software-dependent translation table (i.e. each DNA sequence will correspond to only one allele number, but this allele number can differ between cgMLST softwares).

Currently, the cgMLST scheme is available in SeqSphere+ (a commercial suite) and will be made available in BIGSdb (Jolley and Maiden 2010) (open-source). These two software programs determine allele numbers using draft or full assemblies (fastA files). However, SeqSphere+ also allows to start from raw reads, perform pre-processing and de novo assembly using an integrated pipeline.

Instructions for BIGSdb can be found at https://bigsdb.readthedocs.io/en/latest/.

Assemblies (fastA files) from reference genomes (with known classification) can be downloaded from NCBI (https://www.ncbi.nlm.nih.gov/assembly/?term=mycobacterium%20abscessus) or from cgMLST.org server (e.g. seed genome https://www.cgmlst.org/ncs/strain/RID408249/).

A list of reference genomes can be found at https://github.com/ngs-fzb/NTMtools.

3 Perform cgMLST analysis and import metadata in SeqSphere+

A detailed SeqSphere+ tutorial explaining how to perform cgMLST analysis and import metadata is available at https://www.ridom.de/seqsphere/ug/latest/Tutorial_pipeline.html.

4 Set SLC definitions with allele thresholds in SeqSphere+

We will create new database entry (metadata) fields and make sure that SeqSphere+ assigns cluster types according to allele thresholds for DCC classification and transmission cluster analysis discussed in Diricks et al. 2022.

1. Open SeqSphere+
2. Click on Options > Projects
3. If you don´t have a project yet, make a project with task templates for M. abscessus cgMLST and MLST (and if desired, also for M. abscessus Accessory).

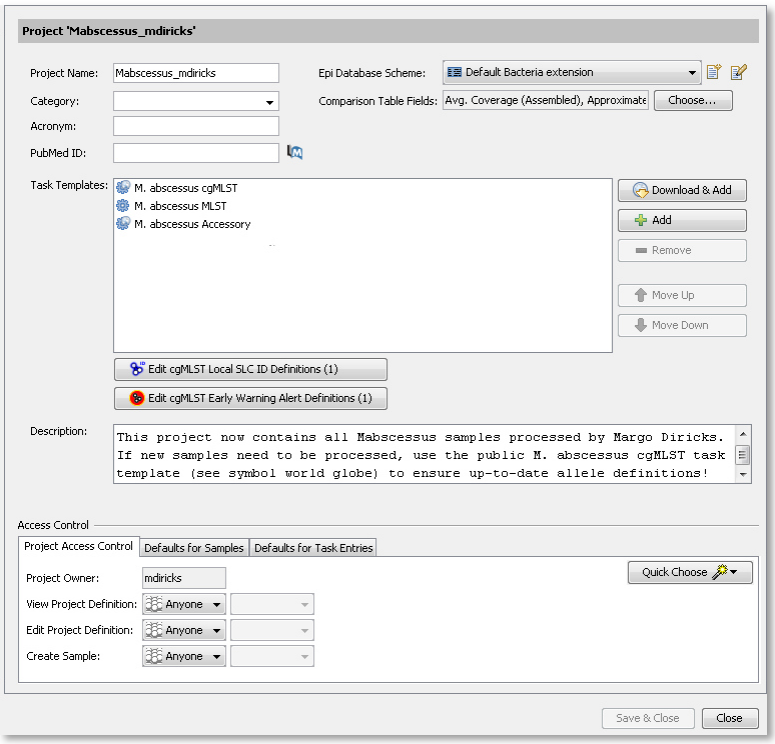

Figure 1 Project window in SeqSphere+.

4. Click on the most right icon next to Epi database scheme (“Edit database scheme”) > Click on Manage fields
5. Click on Create new Field > Enter field name (e.g. Cluster_10, Cluster_25 and Cluster_250), choose section (e.g. epi characteristic) and provide a Description if desired.

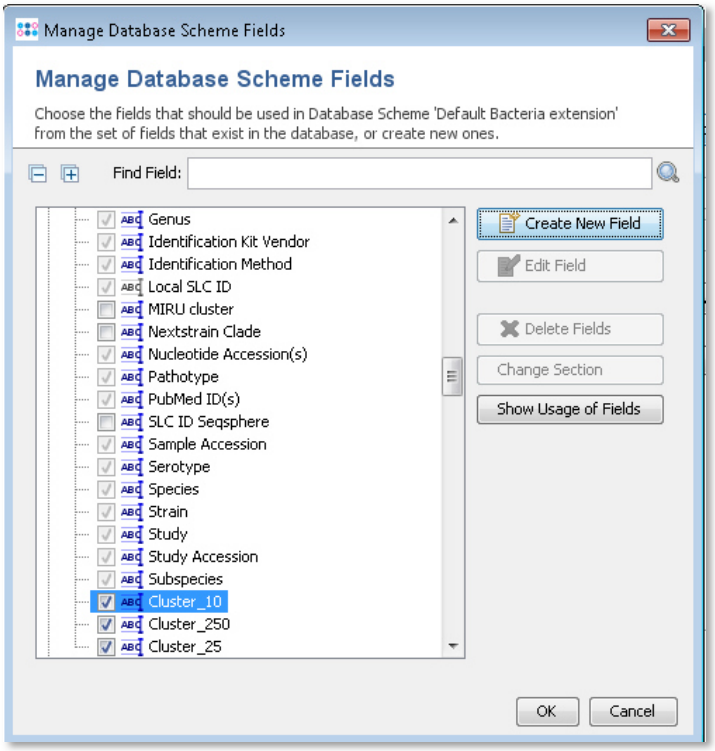

Figure 2 Manage database fields window in SeqSphere+.

6. Click on OK > OK until you are back in the project window.
7. Click on Edit cgMLST local SLC ID definitions: (i) change name (e.g. Cluster_25), (ii) Select Search Samples with a distance equal or less than (second option) e.g. 25 and (iii) change Store SLC ID in Field e.g. Cluster_25. Do this for all cluster tresholds that you want to use.

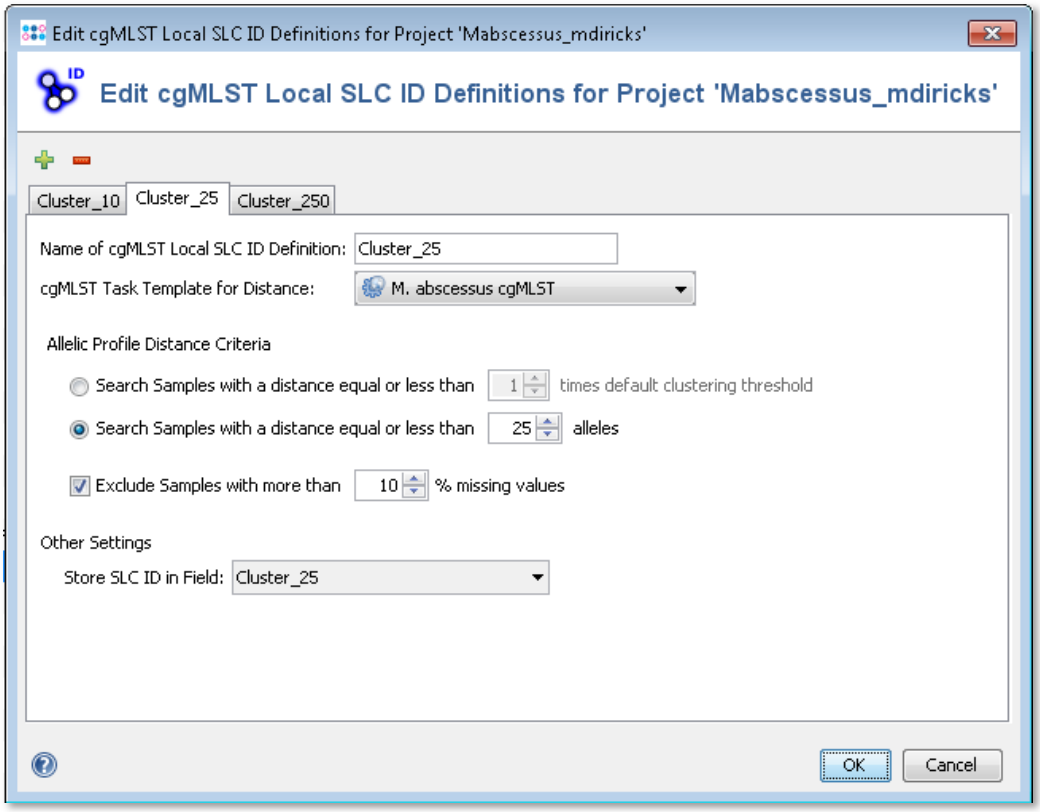

Figure 3 Edit cgMLST local SLC ID definition window in SeqSphere+.

To determine DCC classification of your samples using this approach, see section Using a 250 allele treshold compared to a minimal set of DCC representatives.

5 Early warning alerts in SeqSphere+

Early Warning Alerts (EWA) are defined per project. They are used to automatically detect for newly processed samples (query samples) samples with similar allelic cgMLST profiles (hit samples) that are already stored in the project. If similar samples within the defined thresholds are found, an EWA is triggered and stored in the database. Multiple EWAs can be defined per project.

To set an EWA in SeqSphere+, see https://www.ridom.de/seqsphere/u/Early_Warning_Alerts.html.

6 Determine the subspecies of an unknown isolate with cgMLST analysis

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), e.g. from NCBI.

Table 1 M. abscessus type strains. [1] % Good cgMLST targets: as defined and determined in SeqSphere+ v.8.3.1.

Accession number	***Strain***	***subspecies***	***DCC***	***ST***	***%Good targets [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

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

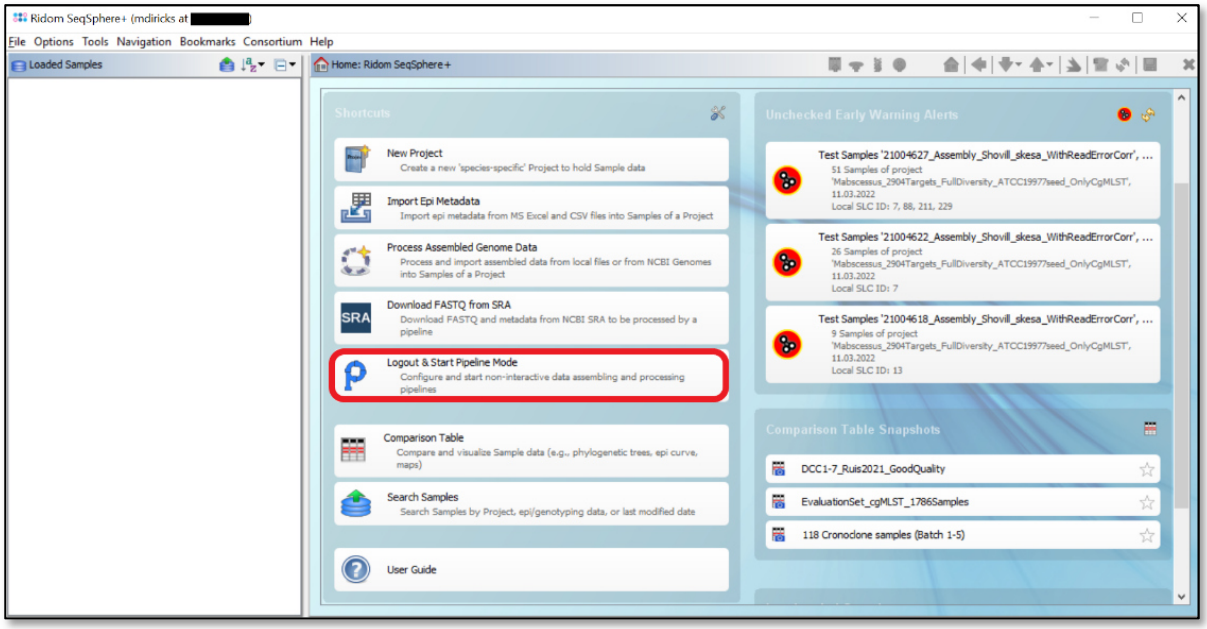

Figure 4 Main window of SeqSphere+.

3. Similarly, determine cgMLST allele profiles of your samples in SeqSphere+
4. 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.

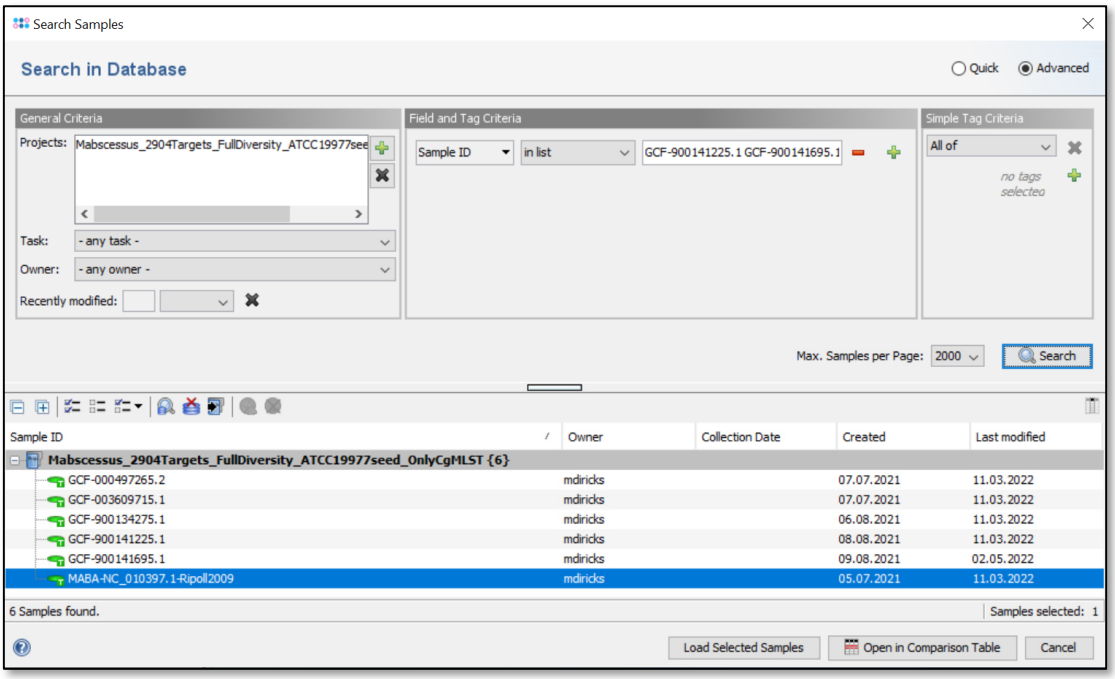

Figure 5 Search Samples window in SeqSphere+.

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

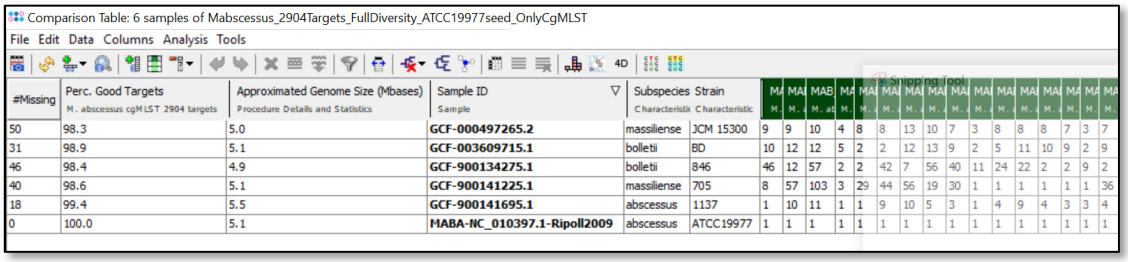

Figure 6 Comparison table window in SeqSphere+.

6. Check if all samples have more than 95% good cgMLST targets (or at least more than 90%)
7. 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.

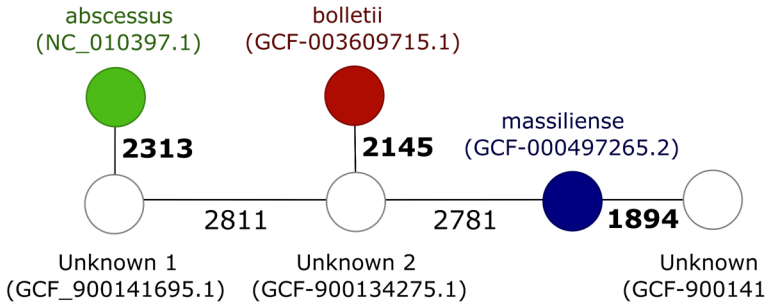

Figure 7 Minimum spanning tree in SeqSphere+. Pairwise distances (allele differences) are displayed on the branches. Subspecies is known for three type strains (colored). Unknown 1 would be classified as abscessus as the distance is minimal towards the abscessus type strain (highlighted in bold). Similarly, unknown 2 and 3 would be classified as bolletii and massiliense, respectively.

8. 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 > Choose folder where you want to save file and choose excel file as file of type (bottom of window) > Save

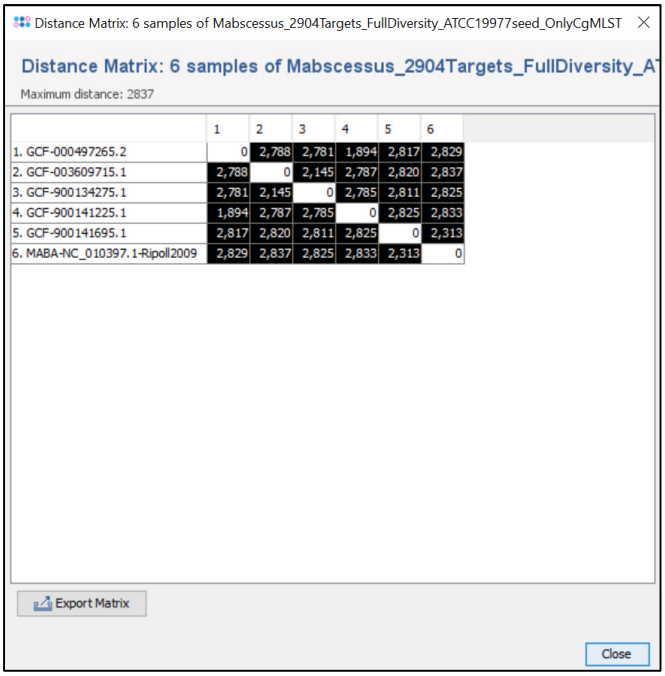

Figure 8 Distance matrix window in SeqSphere+.

9. 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 columns (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*

7 Determine whether an unknown isolate with cgMLST analysis belongs to a known DCC

To know whether your isolate belongs to one of the known DCCs you can (i) calculate a cgMLST based phylogenetic tree comprising your sample(s) and a large set of isolates with known DCC classification

(i.e. principle of phylogenetic positioning) or (ii) calculate the cgMLST distance of your samples towards one reference of each DCC and assign your sample to a DCC if the distance is not more than 250 alleles (Diricks et al., 2022).

7.1 By phylogenetic positioning in tree

Protocol for SeqSphere+ users:

1. Download reference genomes (assemblies; FastA) with known DCC classification and determine cgMLST allele profiles in SeqSphere+. Note: You can e.g. use the reference genomes mentioned in Table S1.
2. Make sure you import the DCC classification in SeqSphere+ too (https://www.ridom.de/seqsphere/ug/latest/Tutorial_pipeline.html).
3. Determine cgMLST allele profiles of your samples in SeqSphere+
4. 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.
5. Select all samples and click on „Open in Comparison Table" at bottom of the window.
6. Check if all samples have more than 95% good cgMLST targets (or at least more than 90%)
7. You can make e.g. a NJ tree by clicking in the comparison window on Analysis > Neighbor joining tree > Leave default settings and click OK. Click on view > circular view
8. If your sample falls within a DCC clade (see e.g. Unknown_1 in Figure 9), you can left click on it (keep also the CTRL button pressed if you want to select multiple samples). Go back to the comparison table window. The sample(s) you have selected in the tree will also be highlighted in this window. You can now right click on the column where you want to store the DCC classification result and choose "Set value and store to database ...". You can now enter the correct DCC classification.
9. Data from the comparison table window can be exported from the comparison table window by clicking on File > Export Table Data.

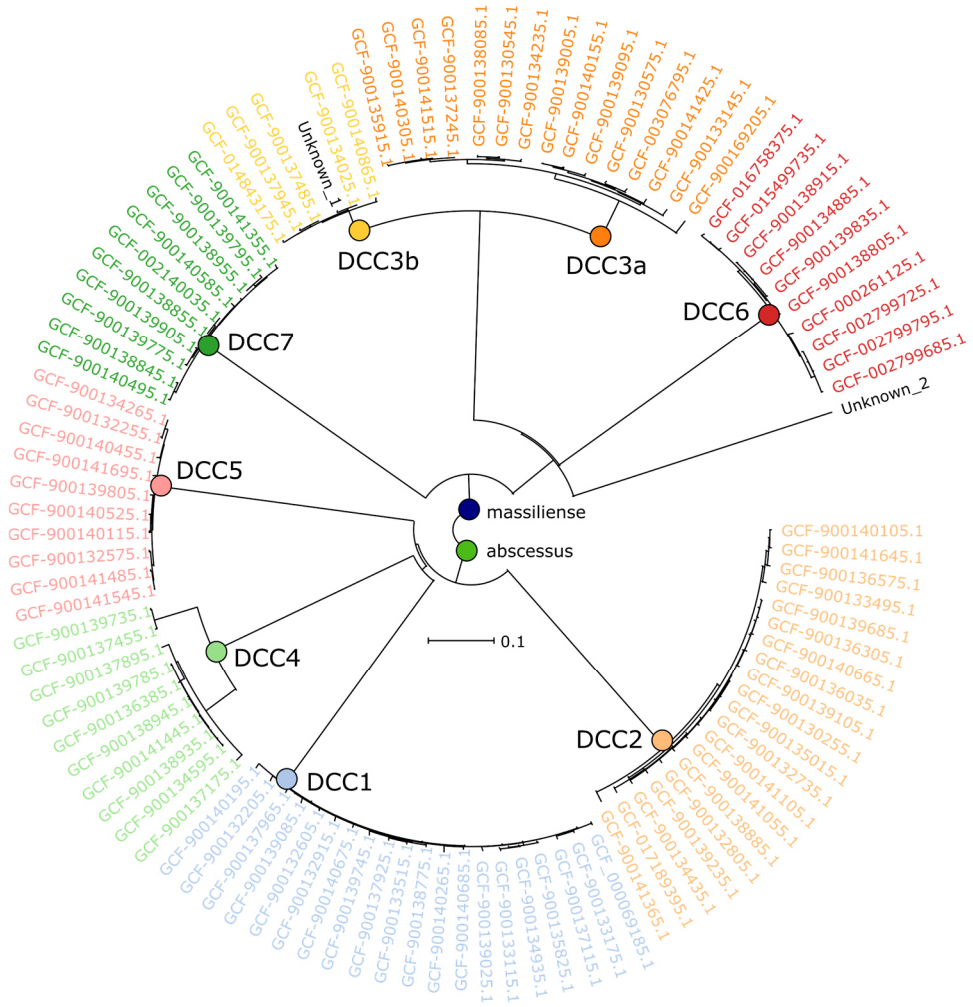

Figure 9 Classification of unknown samples using phylogenetic positioning in a tree with known references (see Table S1). Sample Unknown 1 would be classified as DCC3b and Sample Unknown_2 as unclustered. Note that the classification of GCF-900139735.1 and GCF-900137455.1 into DCC4 is under debate (as they have another ST type and more than 250 allele differences compared to the other cluster members.

7.2 Using a 250 allele treshold compared to a minimal set of DCC representatives

1. Download reference genomes (assemblies; FastA) with known DCC classification and determine cgMLST allele profiles in SeqSphere+. Note: You can use the reference genomes listed in Table 3.

Table 3 Reference genomes for dominant circulating clones (DCC). [1] DCC defined by Ruis et al 2021. [2] MLST ST defined by pubMLST

Subspecies	DCC[1]	ST[2]	Accession number	Strain	Country	Genome length (Mbp)
abscessus	DCC1	5	NC_010397.1 (GCF_000069185.1)	ATCC19977T (CIP 104536T)	Unknown	5.1
abscessus	DCC2	9	NZ_CP065276.1 (GCF_017189395.1)	GD54	Australia	5.2
massiliense	DCC3a	33	NZ_CP029072.1 (GCF_003076795.1)	G220	China	4.8

massiliense	DCC3b	37	NZ_CP062132.1 (GCF_014843175.1)	JHN_AB_0023_1	USA	4.9
abscessus	DCC4	101	FSFB01 (GCF_900134595.1)	976	United Kingdom	4.9
abscessus	DCC5	23	FVYR01 (GCF_900141545.1)	1100	Australia	5.1
massiliense	DCC6	39	NQPA01 (GCF_002799795.1)	A47	China	4.9
massiliense	DCC7	42	NZ_CP021122.1 (GCF_002140035.1)	FLAC047	USA	4.9

2. Similarly, determine cgMLST allele profiles of your samples in SeqSphere+
3. 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.
4. Select all samples and click on „Open in Comparison Table" at bottom of the window.
5. Check if all samples have more than 95% good cgMLST targets (or at least more than 90%)
6. The column "Cluster_250" should by default be present if you have added it according to section Set SLC definitions with allele thresholds. If this column is not present even though you have followed the steps from section Set SLC definitions with allele thresholds, you can add this column in the comparison window with Columns > add additional database fields as columns > choose Cluster_250 column.
7. If your samples have the same number in the Cluster_250 column as a strain with known DCC classification, you know your samples belong to the same DCC.

Alternative:

1. Perform step 1-5 as before.
2. In the comparison Table window click on Analysis > Distance matrix > Leave Default values > OK > Export matrix > Choose folder where you want to save file and choose excel file as file of type (bottom of window) > Save
3. Open this file and copy the columns corresponding to the 8 reference sequences to another excel file/sheet. Add an additional row to indicate which colums (i.e. references) are which DCC (row one in Table 4). Determine the DCC in column 9, row three (cell J3) using following formula: =IF(SMALL(B3:I3;1) <250; INDEX($B$1:$I$1;1;MATCH(SMALL(B3:I3;1);B3:I3;0)); "Non-DCC"). Drag the bottom right corner (or double click) down to other cells in column 9 (E) to determine the DCC of the other samples.

*Table 4 Exported distance matrix and adjustments to let formula work. *formula for this cell would be =IF(SMALL(B3:I3;1) <250; INDEX($B$1:$I$1;1;MATCH(SMALL(B3:I3;1);B3:I3;0)); "Non-DCC")*

	DCC1	DCC2	DCC3a	DCC3b	DCC4	DCC5	DCC6	DCC7	Decision
	GCF_000069185	GCF_017189395	GCF_003076795	GCF_014843175	GCF_900134595	GCF_900141545	GCF_002799795	GCF_002140035	
GCF_017189395	2244	0	2837	2801	2343	2247	2842	2524	DCC2*
GCF_003076795	2839	2837	0	342	2831	2829	2217	2574	DCC3a
GCF_014843175	2589	2801	342	0	2786	2787	2294	2598	DCC3b
GCF_000069185	0	2244	2839	2589	2297	2312	2845	2732	DCC1
GCF_002140035	2732	2524	2574	2598	2706	2717	2535	0	DCC7
GCF_002799795	2845	2842	2217	2294	2838	2837	0	2535	DCC6
GCF_900141545	2312	2247	2829	2787	2333	0	2837	2717	DCC5
GCF_900134595	2297	2343	2831	2786	0	2333	2838	2706	DCC4
Unknown_1	2593	2803	341	16	2788	2789	2296	2600	**DCC3b**
Unknown_2	2852	2848	2281	2327	2841	2839	2338	2587	**Non-DCC**

8 Frequently asked questions

8.1 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 (Diricks et al. 2022; Additional file 1: Table S1). 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.

8.2 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 DCC6) we could not find any.

To be continued.

8.3 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 5 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 immunogenum	20.1	5.6
Mycobacterium tuberculosis	0.1	4.4

9 Supplementary materials

Table S1 Reference set of genomes with known DCC classification (Diricks 2022). Metadata is largely derived from Ruis et al 2021. DCC4?: isolates were classified as DCC4 by Ruis and coworkers, but differ in ST and in more than 250 cgMLST alleles compared to the DCC4 representative.

Sample ID	cgMLST cluster (Diricks 2021)	ST	CF status	Country	Sample_date
GCF_000069185.1	DCC1	5	Non-CF	France	Unknown
GCF_900138775.1	DCC1	5	Non-CF	UK	Unknown
GCF_900134935.1	DCC1	5	CF	UK	2011
GCF_900137925.1	DCC1	5	CF	UK	2013
GCF_900139025.1	DCC1	5	CF	USA	2005
GCF_900139085.1	DCC1	5	Non-CF	USA	2006
GCF_900137965.1	DCC1	5	CF	UK	2013
GCF_900139745.1	DCC1	5	Non-CF	USA	2012
GCF_900140195.1	DCC1	5	CF	Denmark	Unknown
GCF_900140265.1	DCC1	5	Non-CF	USA	2011
GCF_900140675.1	DCC1	5	CF	Denmark	Unknown
GCF_900140685.1	DCC1	5	Non-CF	USA	2006
GCF_900133515.1	DCC1	5	CF	Holland	2009
GCF_900132205.1	DCC1	5	CF	Australia	2006
GCF_900132915.1	DCC1	5	Non-CF	Australia	2008
GCF_900135825.1	DCC1	5	Non-CF	Australia	2011
GCF_900137115.1	DCC1	5	CF	Australia	2012
GCF_900132605.1	DCC1	5	Non-CF	Australia	2007
GCF_900133115.1	DCC1	5	CF	UK	2008
GCF_900133175.1	DCC1	5	Non-CF	Australia	2008
GCF_017189395.1	DCC2	9	CF	Australia	2019
GCF_900130255.1	DCC2	9	CF	UK	2001
GCF_900135015.1	DCC2	9	Non-CF	UK	2011
GCF_900138885.1	DCC2	9	CF	UK	Unknown
GCF_900136575.1	DCC2	9	CF	UK	2012
GCF_900132805.1	DCC2	9	Non-CF	UK	2008
GCF_900132735.1	DCC2	9	CF	UK	2008
GCF_900139105.1	DCC2	9	CF	USA	2008
GCF_900139235.1	DCC2	9	CF	USA	2011
GCF_900139685.1	DCC2	9	CF	USA	2007
GCF_900140105.1	DCC2	9	CF	Denmark	Unknown
GCF_900141055.1	DCC2	9	CF	USA	2013
GCF_900140665.1	DCC2	9	Non-CF	USA	2006
GCF_900141105.1	DCC2	9	CF	USA	2010
GCF_900133495.1	DCC2	9	CF	Australia	2009
GCF_900141365.1	DCC2	9	CF	Ireland	Unknown
GCF_900134435.1	DCC2	9	CF	UK	2010
GCF_900136035.1	DCC2	9	CF	UK	2011
GCF_900136305.1	DCC2	9	CF	UK	2011
GCF_900141645.1	DCC2	9	CF	Sweden	Unknown

GCF_003076795.1	DCC3a	33	Unknown	China	Unknown
GCF_900169205.1	DCC3a	33	Non-CF	Brazil	2004
GCF_900130575.1	DCC3a	33	CF	UK	2002
GCF_900139005.1	DCC3a	33	CF	Denmark	Unknown
GCF_900139095.1	DCC3a	33	CF	USA	2006
GCF_900140155.1	DCC3a	33	CF	Denmark	Unknown
GCF_900140305.1	DCC3a	33	Environmental	UK	Unknown
GCF_900130545.1	DCC3a	33	CF	Australia	2001
GCF_900134235.1	DCC3a	33	Non-CF	Australia	2010
GCF_900138085.1	DCC3a	33	CF	Australia	2013
GCF_900141425.1	DCC3a	33	CF	Ireland	Unknown
GCF_900141515.1	DCC3a	33	CF	UK	Unknown
GCF_900135915.1	DCC3a	33	Non-CF	UK	2011
GCF_900133145.1	DCC3a	33	CF	UK	2008
GCF_900137245.1	DCC3a	33	CF	UK	2012
GCF_900137485.1	DCC3b	37	CF	UK	2012
GCF_014843175.1	DCC3b	37	Unknown	USA	Unknown
GCF_900137945.1	DCC3b	37	CF	UK	2012
GCF_900134025.1	DCC3b	37	Unknown	UK	2010
GCF_900140865.1	DCC3b	37	CF	USA	2007
GCF_900134595.1	DCC4	101	CF	UK	2010
GCF_900138935.1	DCC4	101	CF	UK	Unknown
GCF_900138945.1	DCC4	101	CF	UK	Unknown
GCF_900139735.1	DCC4?	97	CF	USA	2006
GCF_900139785.1	DCC4	101	CF	Denmark	Unknown
GCF_900141445.1	DCC4	101	CF	UK	Unknown
GCF_900137175.1	DCC4	101	CF	UK	2012
GCF_900137455.1	DCC4?	97	CF	UK	2012
GCF_900136385.1	DCC4	101	CF	UK	2011
GCF_900137895.1	DCC4	101	CF	UK	2012
GCF_900141545.1	DCC5	23	Unknown	Australia	Unknown
GCF_900139805.1	DCC5	23	CF	USA	2008
GCF_900140115.1	DCC5	23	CF	Denmark	Unknown
GCF_900140455.1	DCC5	23	CF	USA	2005
GCF_900140525.1	DCC5	23	CF	USA	2005
GCF_900134265.1	DCC5	23	Non-CF	Australia	2010
GCF_900132255.1	DCC5	23	Non-CF	Australia	2006
GCF_900132575.1	DCC5	124	Non-CF	Australia	2007
GCF_900141695.1	DCC5	154	CF	Sweden	Unknown
GCF_900141485.1	DCC5	23	Unknown	Australia	
GCF_900134885.1	DCC6	39	Non-CF	UK	2011
GCF_900138805.1	DCC6	39	CF	UK	Unknown
GCF_900138915.1	DCC6	39	CF	UK	Unknown
GCF_900139835.1	DCC6	39	Non-CF	UK	Unknown
GCF_002799795.1	DCC6	39	Non-CF	China	2013
GCF_002799725.1	DCC6	39	Non-CF	China	2012
GCF_002799685.1	DCC6	39	Non-CF	China	2014

GCF_000261125.1	DCC6	39	Unknown	Malaysia	Unknown
GCF_015499735.1	DCC6	39	Unknown	Mexico	Unknown
GCF_016758375.1	DCC6	39	Unknown	Brazil	Unknown
GCF_002140035.1	DCC7	42	Unknown	USA	Unknown
GCF_900138845.1	DCC7	42	CF	UK	Unknown
GCF_900138855.1	DCC7	42	CF	UK	Unknown
GCF_900138955.1	DCC7	42	CF	UK	Unknown
GCF_900139795.1	DCC7	42	Non-CF	USA	2011
GCF_900139775.1	DCC7	42	Non-CF	USA	2008
GCF_900139905.1	DCC7	42	CF	USA	2007
GCF_900140495.1	DCC7	42	CF	USA	2005
GCF_900140585.1	DCC7	42	Non-CF	USA	2006
GCF_900141355.1	DCC7	42	CF	Netherlands	Unknown