

*Bioinformatic tools for analysis of whole genome sequencing data
from non-tuberculous mycobacteria*

WHOLE GENOME SEQUENCING

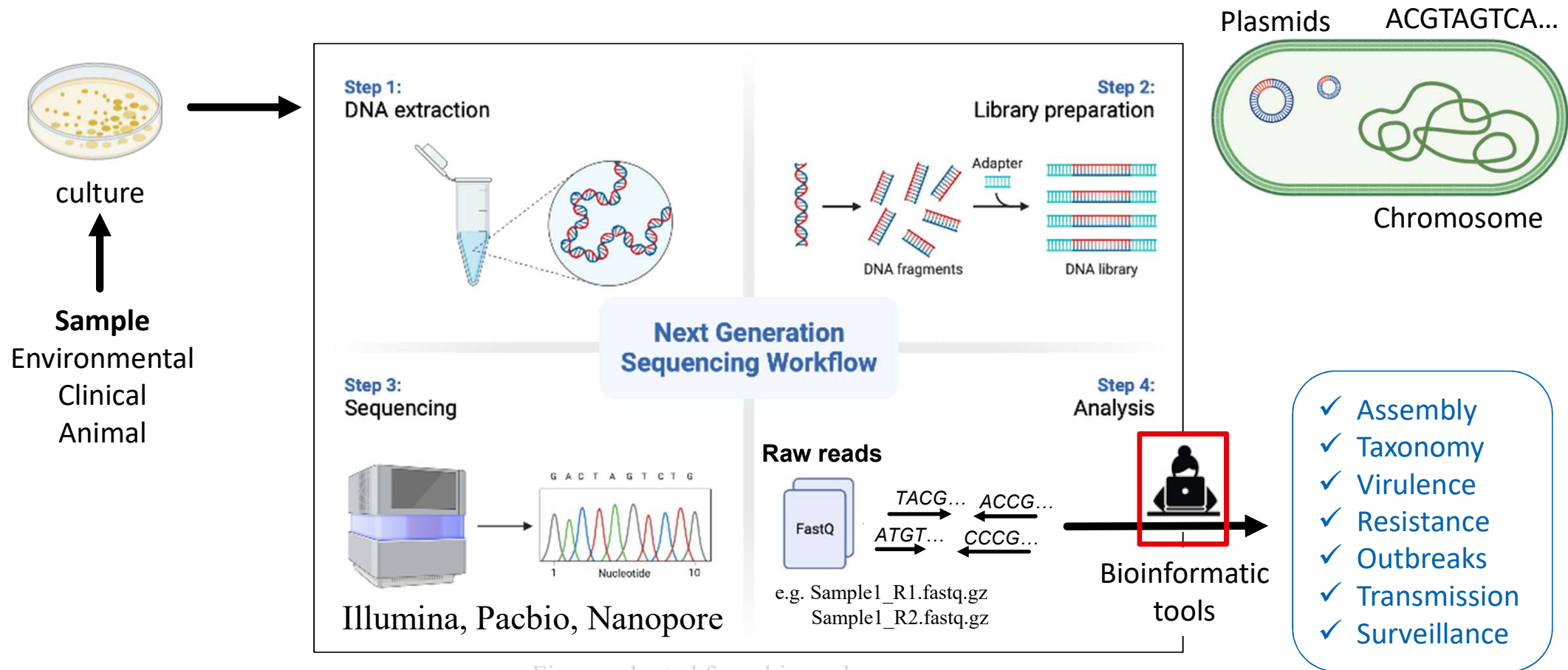
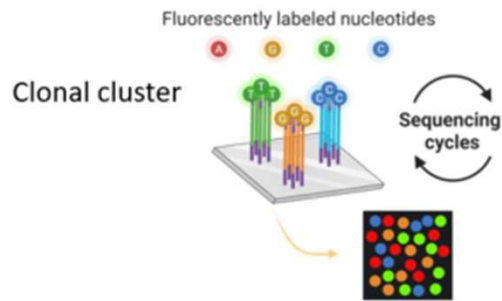


Figure adapted from biorender

WHOLE GENOME SEQUENCING TECHNOLOGIES

Illumina

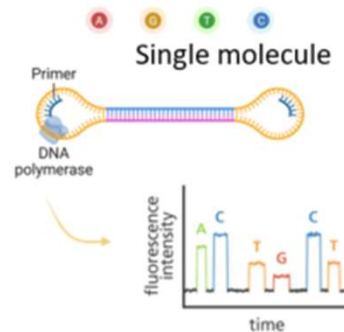


Short reads
(150-300 bp)

isolate1_R1.fastq.gz
isolate1_R2.fastq.gz

ACGTCGC...

Pacbio



Long Hifi reads
(~10-25 kb)

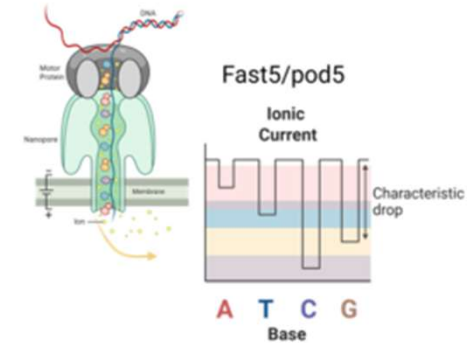
isolate1.fastq

ACGTCGC...

Nanopore



Single molecule



Long reads
(up to 1 Mbp)

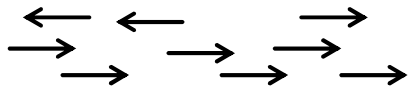
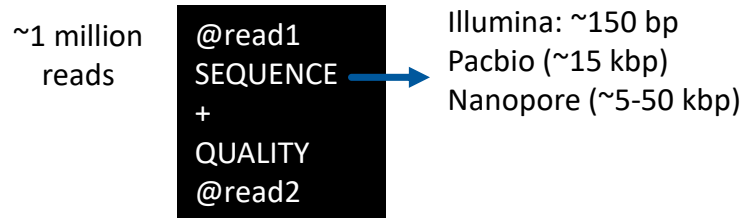
isolate1.fastq

Created with biorender

SEQUENCING TECHNOLOGIES

Feature	Illumina (NGS)	PacBio (HiFi / SMRT)	Oxford Nanopore (ONT)
Sequencing principle	Sequencing-by-synthesis (reversible terminator fluorescence)	Single-molecule real-time (SMRT) fluorescence	Ionic current changes through nanopore
Detection type	Optical (fluorescent base incorporation)	Optical (fluorescent pulses during incorporation)	Electrical signal (current disruption)
Amplification required	<input checked="" type="checkbox"/> Yes (PCR / bridge amplification)	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> No
Detection point	During synthesis on clustered DNA fragments	During synthesis in zero-mode waveguides real time	As DNA/RNA passes through pore in real time
Typical read length	~150–300 bp (paired-end)	~10–25 kb (HiFi)	10–100 kb typical; ultra-long >100 kb
Accuracy (raw / consensus)	Very high (~99.5%)	Very high (~99% HiFi)	Moderate raw (85–98%), improves after polishing
Run time	~12–48 h	~10–30 h	Real-time (minutes to ~48 h)
Library prep time	Moderate	Moderate	Fast
Throughput	Very high	Medium–high	Flexible
Epigenetics detection	<input checked="" type="checkbox"/> No (needs bisulfite, etc.)	<input checked="" type="checkbox"/> Yes (polymerase kinetics)	<input checked="" type="checkbox"/> Yes (direct signal)
Assembly quality (bacteria)	Fragmented	Excellent (often closed genomes)	Excellent (often closed genomes)
Estimated cost per NTM genome (5–8 Mbp)	~€50–150	~€100–300	~€50–200
Best use cases	SNPs, large cohorts, cheap resequencing	High-quality assemblies	Long reads, rapid sequencing

FastQ



```
@HWI-ST911:111:C0N4WACXX:5:1101:2249:2216 1:N:0:TTAGGC CGATC:@@FF
NATGGCACCATTAAAAAGAAATGTTTATATGGTGTGAGAAGGACAAAGCTGAAGAAGAAATTTAGTCTGCACCTTGATGTTGCAATGCAAGAAAA
+
#2A2<CCFHIIIIIIIIIGCCHIIIGIIIFPHIIDGHIGIIIIITCHGIIIGGCECEGICFHCECEFFFFDEEEEEEDDDDDCDDDDDDBC
@HWI-ST911:111:C0N4WACXX:5:1101:2509:2197 1:N:0:TTAGGC CGATC:1+4=B
NATGAGATAAATCAATGTCCTTAATGAAGTACAGCTTTGAATAATGAGTTTTGAACTCTTCTGCAACTTTTTGAAACTTTAAAGTTTGAATG
+
#4A2<AADHIIIIIIIIHIIIIIIIFGIIIGIIIFIIIGIIIIIDHEHIIIIIIIIIIICHIIIHHEEDFFFECEEEADDFFC
@HWI-ST911:111:C0N4WACXX:5:1101:3746:2179 1:N:0:TTAGGC CCATC:+11+A
NATGTCATCCATCTTTCTATCTAAAAAAGAATCAAAAAAGGGATAGTACAGAGGAAAGTTCAATCCAGAGGACGATGAAACACTGATTGATGG
+
```

e.g. Isolate1_R1.fastq.gz
200 MB

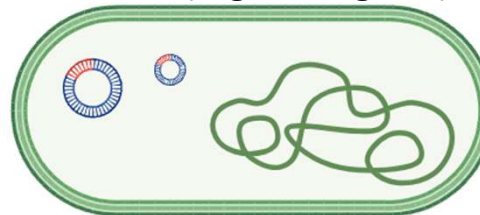


Short reads:
SPAdes, skesa,...

De novo assembly

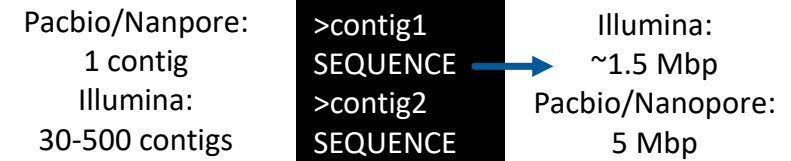
Long reads:
flye, canu,...

Plasmids (e.g. contig 2-3)



Chromosome (e.g. contig 1)

FastA



TACGACGTCGTCGTCG GTCTA GTC
Contig 1 Contig 2 Contig 3

Draft genome

```
>contig00001 len=1551651 cov=62.3 corr=0 origname=Contig_11_6
CGCCACCGCCGCCCTCCGCCGGTGACGACGCATGAGGCGCCGCCCGGGTGACTACCA
CGACGACACCGCCGCCCTCCGCCCGACGACCAGACACCGGGCCGGGCACGGCAGA
TCACGTACTCGGTGACCGGTTCCAAGGCTCCGCTCGATCGCATCTCGATCACCTGGACCG
ACGGTTCGGGACGCACCCGGGTGAACCCGAATGTGTACATCCCGTGGTTCGATCACGGTTA
CCCCATTTGCAATTCGGAGATCGGATCGGTGTCGGCGAGTAGCTTCTGCGGCTGAGTC
AGCTCAACTGCACGATCACCACCAGCGACGGTCAGGTTTTGTCTCCAACAACAATT
CGGCGCAGGCAACCTGCTGATGCCAGAGGTCGATTGGCCAAGCCACTGGAGGCAATCGC
GGGCCGCTGCGGCGGTATCGCCGGAGTCCGTTGACCGGATGCTCATCGGCTTGTGCG
```

e.g. Isolate1.fasta

5 MB

SUMMARY OF BIOINFORMATIC TOOLS FOR NTM WGS



Tool Name	Interface	License Type	Key Functionalities for NTM	Input Formats	Specificity
NTM-Profiler	Command-line	Open-source	(sub)species identification, drug resistance prediction	FASTQ, BAM/CRAM, FASTA, VCF	NTM-specific
MyCodentifier	Command-line	Open-source	Species identification	FASTQ	TB and NTM
NTMseq	Command-line	Open-source	Quality control, contamination detection, (sub)species identification, resistance prediction, plasmid prediction, assembly, MLST, phylogenetics	FASTQ, FASTA	NTM-specific
Mycobacteria Explorer	Web-based	Free	Subspecies and drug resistance prediction	FASTA	NTM-specific MAC
SAM-TB	Web-based	Free	(sub)species identification	FASTQ	TB and NTM
GenoMycAnalyzer	Web-based	Free	Quality control, species identification	FASTQ	TB and NTM
Pathogenwatch	Web-based	Free	Species identification, MLST, phylogenetics	FASTA, FASTQ	Microbes
TYGS	Web-based	Free	Species identification, phylogenetics	FASTA	Microbes
SeqSphere+	Desktop-based	Commercial	(cg)MLST, SNP analysis, phylogenetics, cluster analysis	FASTQ, FASTA	Microbes
pubMLST	Web-based	Free	MLST, phylogenetics, cluster analysis	FASTA	Microbes

For research use only!!



Mycobacteria/NTM-specific tools

• Access

- **Web-based:** <https://bioinformatics.lshtm.ac.uk/ntm-profiler/>
- **CLI:** <https://github.com/jodyphelan/NTM-Profiler>
- **Resistance mutation catalogues:**
 - <https://pathogen-profiler.github.io/ntm-db>
 - <https://github.com/pathogen-profiler/ntm-db>
- **Author:** Jody Phelan (London School of Hygiene and Tropical Medicine)
- **Input:** FastQ, BAM/CRAM, FastA, VCF
- **Output:** (Sub)species, drug resistance
- **Principle for NTM ID:** Sequences matched against mycobacterial ref. genomes from GTDB (sourmash)
- **Comments**
 - Still under development – new versions released often



Jody Phelan

<https://gtdb.ecogenomic.org/>



- **Web-based:** <https://bioinformatics.lshtm.ac.uk/ntm-profiler/>



Jody Phelan

NTM-Profiler Home Upload

Welcome to the webserver of NTM-Profiler - a pipeline which allows users to analyse *Mycobacterial* whole genome sequencing data to predict lineage and drug resistance. Follow the instructions below to upload a new sample or view analysed runs.

How does it work?

The pipeline searches for small variants and big deletions associated with drug resistance. It will also report the lineage. By default it uses Trimmomatic to trim the reads, BWA (or minimap2 for nanopore) to align to the reference genome and GATK (open source v4) to call variants.

Step 1

Profile your sample

Upload your next generation sequencing data in **fastQ** format. You can upload one or two (forward and reverse) fastq files. When you upload your data, the run will be assigned a unique ID. Please take a note of this ID as you will need to find your results later. Batch upload of samples is also possible.

Upload

Step 2

View the results

Find your results by entering you unique run ID directly into the search box below.

Sample ID

Submit

- **Web-based:** <https://bioinformatics.lshtm.ac.uk/ntm-profiler/>

Summary QC Species Drug resistance Genome browser Log

Species found

Species assignment is performed using a kmer-based approach using the taxonomy as defined by GTDB. The closest match in the database is shown along with the average nucleotide identity (ANI). The relative abundance is calculated by normalising the abundance of species-specific kmers found.

Species	Closest DB match	ANI	Relative abundance
Mycobacterium abscessus	GCA_015355655.1	99.98	0.9796340397556688
Mycobacterium abscessus	GCF_900136315.1	97.86	0.020365960244331297

Subspecies taxonomy

Subspecies taxonomic placement is performed using a SNP barcode.

Cluster ID	Frequency
subsp. bolletii	0.21359223300970873

Summary QC Species Drug resistance Genome browser Log

Overview

The drug resistance analysis is performed using the genome variants found in the sample. The resistance is determined by the presence of known resistance to the drugs tested and the supporting genetic determinants.

Drug	Resistance	Supporting variants
macrolides		
amikacin		

Validated only for *M. abscessus*

Other variants

Variant ID	Position	Reference	Variant	Category	Frequency
rfl	1464318	n.111C>T		non_coding_transcript_exon_variant	0.9859154929577465
rfl	1464484	n.277C>T		non_coding_transcript_exon_variant	1.0
rfl	1464540	n.333C>T		non_coding_transcript_exon_variant	1.0
rfl	1464557	n.350G>A		non_coding_transcript_exon_variant	1.0
rfl	1464841	n.634C>T		non_coding_transcript_exon_variant	1.0
rfl	1465142	n.935T>C		non_coding_transcript_exon_variant	0.9807692307692307
erm(41)	2345735	c.-220T>C		upstream_gene_variant	0.9705882352941176
erm(41)	2345982	p.ttp10Arg		loss_of_function_variant	0.9824561403508771
erm(41)	2346113	c.159T>C		synonymous_variant	1.0
erm(41)	2346192	p.lle80Val		missense_variant	1.0
erm(41)	2346284	c.330A>C		synonymous_variant	1.0

Quality control

The quality control metrics are calculated using the reads mapped to the reference genome.

Overview

Metric	Value	Description
Number of reads mapped	1511887	Number of reads mapped to the reference genome
Percent of reads mapped	92.26	Percentage of reads mapped to the reference genome
Median depth	61.0	Median depth of the target regions

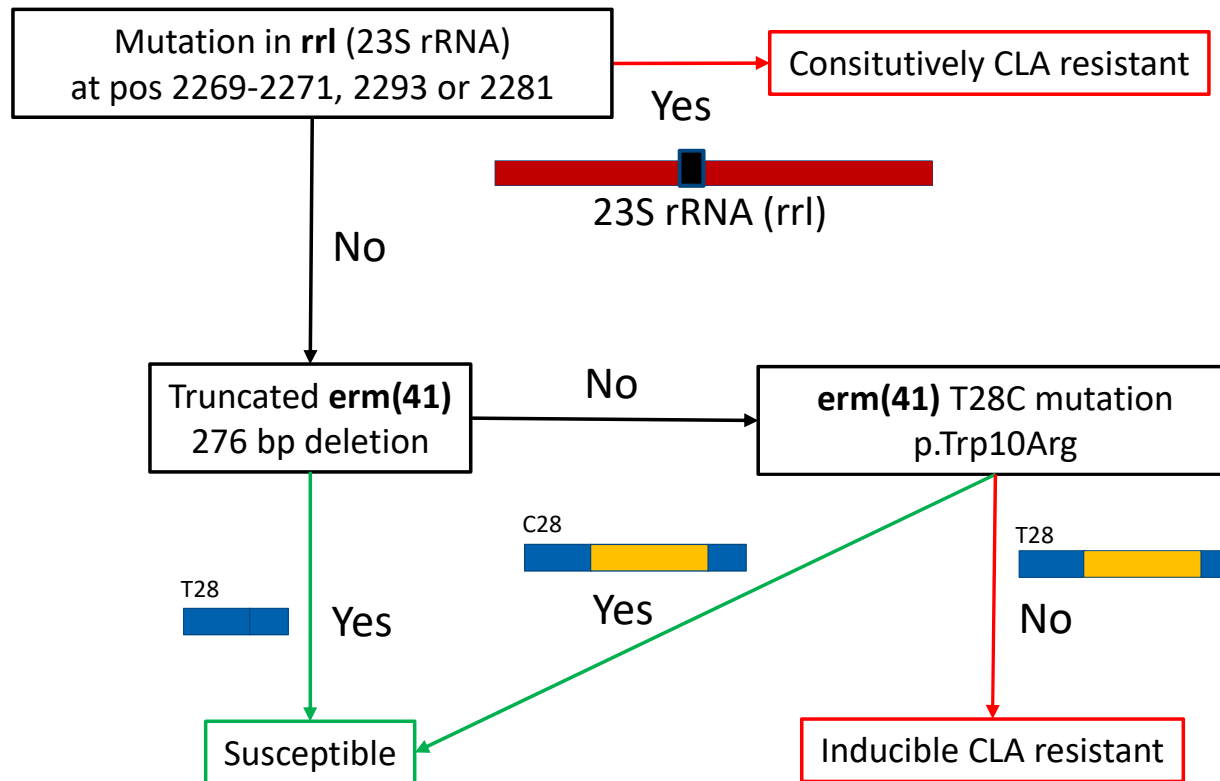
Target-specific coverage

The target-specific coverage is calculated using the reads mapped to the reference genome. The percent depth pass is the percentage of the target covered at a depth greater than 10.

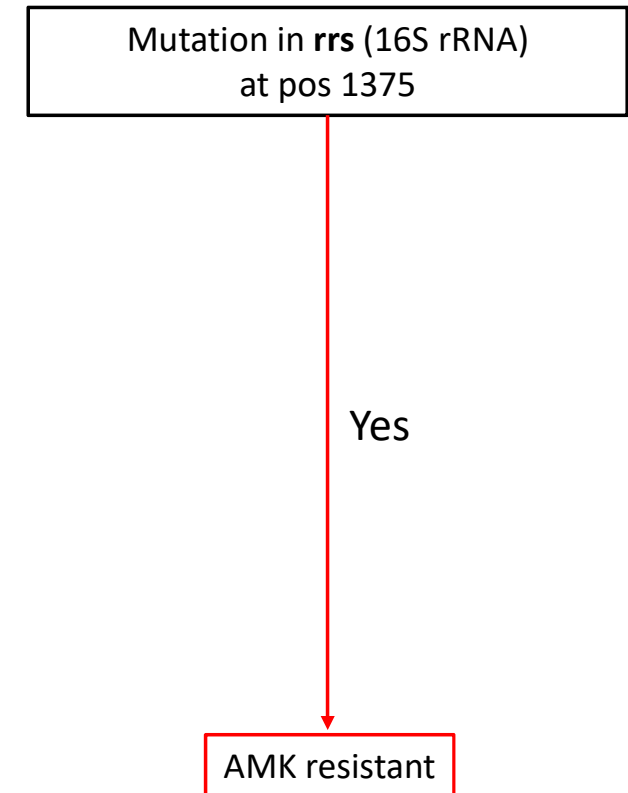
Target	Percent depth pass	Median depth
MAB_0006	95.71	61.0
MAB_0019	100.0	62.0
MAB_0185c	100.0	54.0
rrs	100.0	61.0
rfl	100.0	64.0
erm(41)	100.0	43.0

M. abscessus

Macrolides



Aminoglycosides



NTM PROFILER



- CLI: <https://github.com/jodyphelan/NTM-Profiler>

README GPL-3.0 license

NTM-Profiler

This repository hosts the code for **NTM-Profiler**. A tool to predict species and drug resistance from NTM WGS data.

Please beware that this tools is in alpha testing and should not yet be considered for production use. If you would like to get involved and help out with testing or development please drop me a line through the issues tab.

Installation is available through conda:

```
conda install bioconda::ntm-profiler
```

After installing, the relevant species and resistance databases can be downloaded by running:

```
ntm-profiler update_db
```

pathogen-profiler / ntm-db

<> Code Issues Pull requests

Files

main

Go to file

- > .github
- ✓ db
 - > Mycobacterium_abscessus
 - > Mycobacterium_avium
 - > Mycobacterium_fortuitum
 - > Mycobacterium_intracellulare
 - > Mycobacterium_leprae
 - > Mycobacterium_malmoense
 - > Mycobacterium_marinum
 - ✓ species
 - ➔ ntm-sylph-db
 - ✓ sketches

Set of hashes =
fingerprint for each
genome)

- ✓ sketches
 - GCA_000026685.1.sig
 - GCA_000165695.1.sig
 - GCA_000230935.2.sig
 - GCA_000240345.2.sig
 - GCA_000240365.2.sig
 - GCA_000240505.2.sig
 - GCA_000240525.2.sig
 - GCA_000330785.1.sig

<https://github.com/pathogen-profiler/ntm-db/tree/main/db>

NTM PROFILER



- CLI: <https://github.com/jodyphelan/NTM-Profiler>

README GPL-3.0 license

NTM-Profiler

This repository hosts the code for **NTM-Profiler**. A tool to predict species and drug resistance from NTM WGS data.

Please beware that this tools is in alpha testing and should not yet be considered for production use. If you would like to get involved and help out with testing or development please drop me a line through the issues tab.

Installation is available through conda:

```
conda install bioconda::ntm-profiler
```

After installing, the relevant species and resistance databases can be downloaded by running:

```
ntm-profiler update_db
```

pathogen-profiler / ntm-db

Code

Files

main

Go to file

.github

db

- Mycobacterium_abscessus
- Mycobacterium_avium
- Mycobacterium_fortuitum
- Mycobacterium_intracellulare
- Mycobacterium_leprae
- Mycobacterium_malmoense
- Mycobacterium_marinum
- species
 - ntm-sylph-db
- sketches

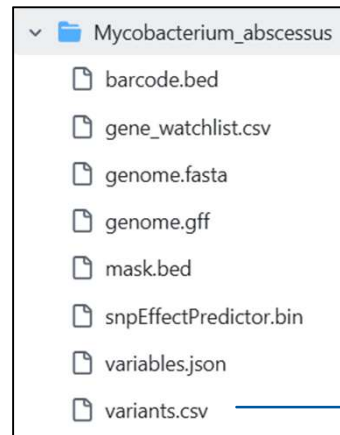
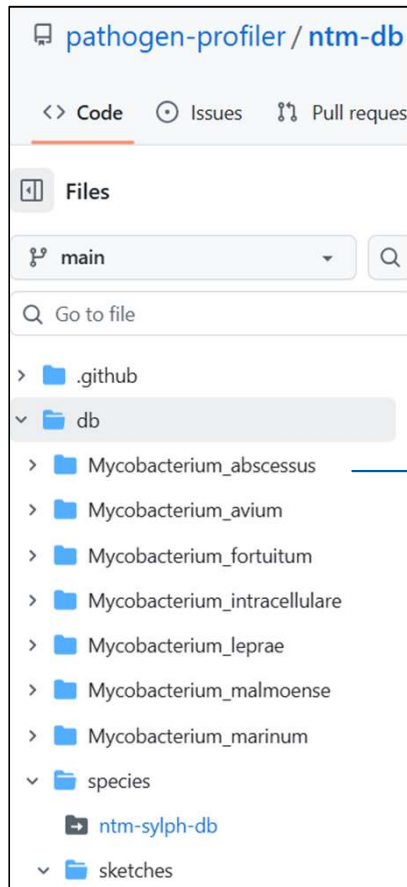
rrs	n.1375A>C	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	A1408C
rrs	n.1375A>G	drug_resistance	amikacin	10.1038/s41467-021-25484-9	yes	A1408G
rrs	n.1375A>T	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	A1408T
rrs	n.1376C>A	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	C1409A
rrs	n.1376C>G	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	C1409G
rrs	n.1376C>T	drug_resistance	amikacin	10.1038/s41467-021-25484-9	yes	C1409T
erm(41)	functionally_normal	drug_resistance	macrolides	10.1038/s41467-021-25484-9		
erm(41)	p.Trp10Arg	loss_of_function		10.1016/j.jmoldx.2021.07.023	yes	
erm(41)	p.Arg7*	loss_of_function		10.1016/j.jmoldx.2021.07.023	no	
gyrA	p.Asp96Asn	drug_resistance	fluoroquinolones	10.1128/aac.01051-24		

<https://github.com/pathogen-profiler/ntm-db/tree/main/db>

NTM PROFILER



- CLI: <https://github.com/jodyphelan/NTM-Profiler>



Gene	Mutation	type	drug	literature	hain	E.coli-nomenclature
rrl	n.2269A>C	drug_resistance	macrolides	10.1016/j.jmoldx.2021.07.023	no	A2057C
rrl	n.2269A>G	drug_resistance	macrolides	10.1016/j.jmoldx.2021.07.023	no	A2057G
rrl	n.2269A>T	drug_resistance	macrolides	10.1016/j.jmoldx.2021.07.023	no	A2057T
rrl	n.2270A>C	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2058C
rrl	n.2270A>G	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2058G
rrl	n.2270A>T	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2058T
rrl	n.2271A>C	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2059C

⋮

rrs	n.1375A>C	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	A1408C
rrs	n.1375A>G	drug_resistance	amikacin	10.1038/s41467-021-25484-9	yes	A1408G
rrs	n.1375A>T	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	A1408T
rrs	n.1376C>A	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	C1409A
rrs	n.1376C>G	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	C1409G
rrs	n.1376C>T	drug_resistance	amikacin	10.1038/s41467-021-25484-9	yes	C1409T
erm(41)	functionally_normal	drug_resistance	macrolides	10.1038/s41467-021-25484-9		
erm(41)	p.Trp10Arg	loss_of_function		10.1016/j.jmoldx.2021.07.023	yes	
erm(41)	p.Arg7*	loss_of_function		10.1016/j.jmoldx.2021.07.023	no	
gyrA	p.Asp96Asn	drug_resistance	fluoroquinolones	10.1128/aac.01051-24		

<https://github.com/pathogen-profiler/ntm-db/tree/main/db>

- <https://pathogen-profiler.github.io/ntm-db>

NTM-DB Knowledge Base

NTM-DB Knowledge Base

NTM-DB

Mycobacterium abscessus

Mycobacterium avium

Mycobacterium fortuitum

Mycobacterium intracellulare

Mycobacterium leprae

Mycobacterium malmoense

Mycobacterium marinum

Species

Mycobacterium abscessus

General information

Key	value
Species	Mycobacterium abscessus
Reference sequence accession	CU458896
Subspecies detection	Yes
Resistance detection	Yes

Subspecies detection

The following subspecies are detected:

- subsp. abscessus
- subsp. bolletii
- subsp. massiliense

Drug resistance

Drug resistance is detected for:

- macrolides
- amikacin
- fluoroquinolones

Genes of interest

Gene	Drug	Literature
rrl	macrolides	10.1038/s41467-021-25484-9
rrs	amikacin	10.1038/s41467-021-25484-9
MAB_2297	macrolides	10.1038/s41467-021-25484-9
MAB_0019	fluoroquinolones	10.1038/s41467-021-25484-9
MAB_0006	fluoroquinolones	10.1038/s41467-021-25484-9
MAB_0185c	ethambutol	10.1093/jac/dkr578
MAB_3542c	fluoroquinolones	10.1128/aac.01051-24

Custom Search Builder Search:

Gene	Mutation	type	drug	literature	hain	E.coli-nomenclature
erm(41)	functionally_normal	drug_resistance	macrolides	10.1038/s41467-021-25484-9		
erm(41)	p.Trp10Arg	loss_of_function		10.1016/j.jmoldx.2021.07.023	yes	
erm(41)	p.Arg7*	loss_of_function		10.1016/j.jmoldx.2021.07.023	no	
gyrA	p.Asp96Asn	drug_resistance	fluoroquinolones	10.1128/aac.01051-24		
rrl	n.2269A>C	drug_resistance	macrolides	10.1016/j.jmoldx.2021.07.023	no	A2057C
rrl	n.2269A>G	drug_resistance	macrolides	10.1016/j.jmoldx.2021.07.023	no	A2057G
rrl	n.2269A>T	drug_resistance	macrolides	10.1016/j.jmoldx.2021.07.023	no	A2057T
rrl	n.2270A>C	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2058C
rrl	n.2270A>G	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2058G
rrl	n.2270A>T	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2058T

Showing 1 to 10 of 28 entries

- **Access:**
 - CLI: <https://github.com/JordyCoolen/MyCodentifier>
- **Author:** Jordy Coolen and Heleen Severin (Radboud University, Netherlands)
- **Input:** FastQ
- **Output:** Species, (drug resistance: currently being implemented)
- **Principle for NTM ID:** Sequences matched against a custom *hsp65* database and whole-genome database (KMA/centrifuge)
- **Comments**
 - Developed for WGS from early positive MGIT cultures
 - Still under development

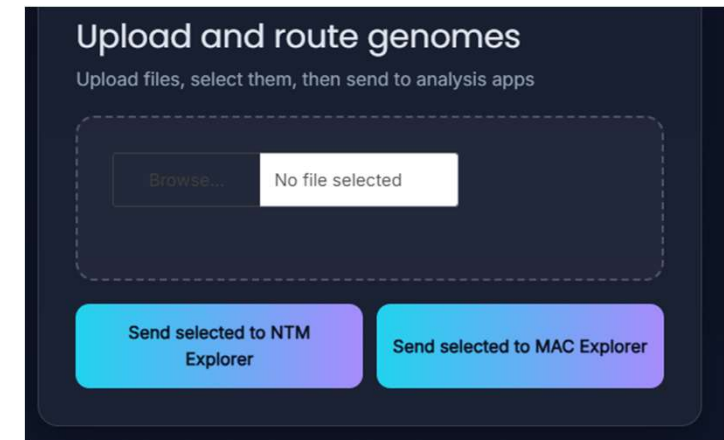


TB/NTM species identification pipeline Version 1.0 (beta)

MYCOBACTERIA EXPLORER



- **Access:**
 - Web-based: Mycobacteriaexplorer.fr
- **Author:** France (paper in revision)
- **Principle for NTM ID:** K-mer ML
- **Input:** FastA
- **Output:** Species ID, subspecies, lineage, resistance
- **Comments**
 - Still under development



Show 15 entries

Search:

	genome	gene	gene_found	identity	coverage	mutation	status	antibiotic	impact
1	4434-15.fasta	rrs	YES	90.9	100	A1408G	ABSENCE	Amikacine	-
2	4434-15.fasta	rrs	YES	90.9	100	C1409T	ABSENCE	Amikacine	-
3	4434-15.fasta	rrs	YES	90.9	100	G1491C	PRESENCE	Amikacine	Résistance possible
4	4434-15.fasta	rrs	YES	90.9	100	G1491T	ABSENCE	Amikacine	-

MAC Explorer v1.0

terium avium complex (MAC) prediction by k-mer ML: upload genome(s), run ML, get consensus names and GTDB correspondence.

Notice Results Subspecies Resistance

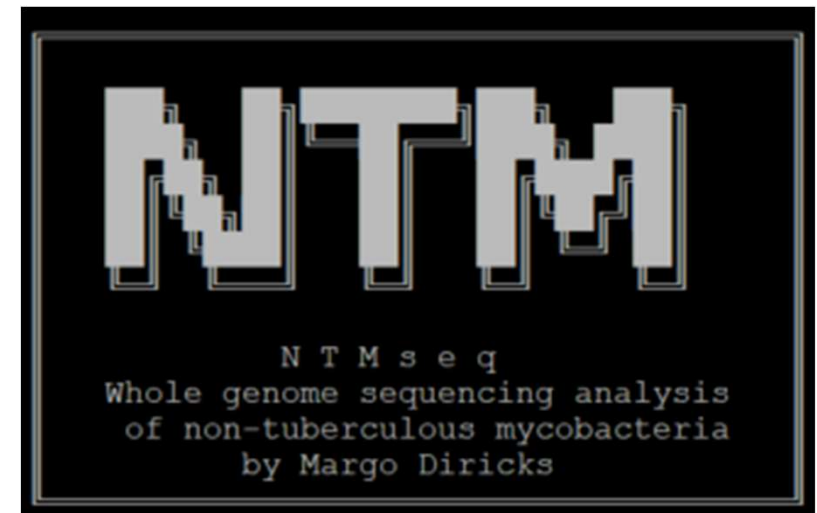
CSV Excel

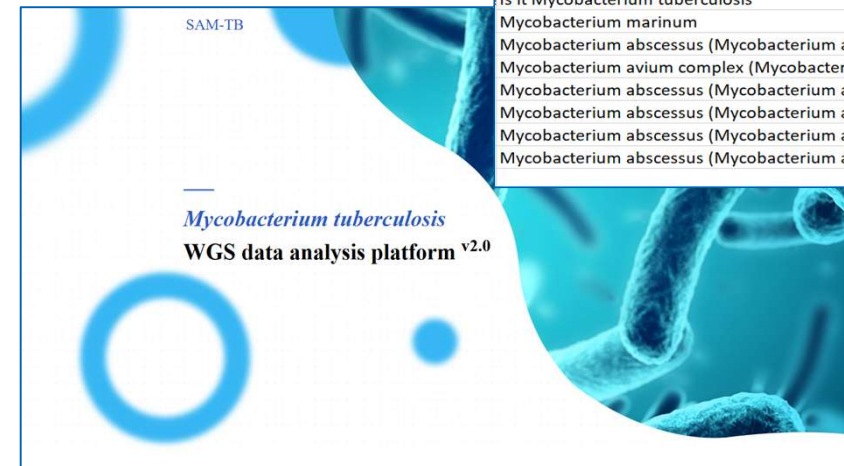
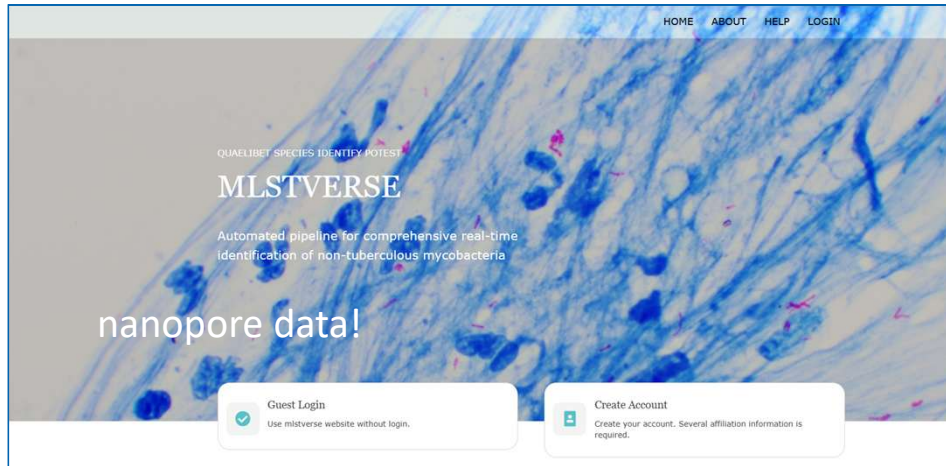
Search:

File	Name	GTDG_correspondence	Probability_1	IS_Analysis	Paratb_Check	Chimerism_Check	Resistance_Analysis	Name_2	Probability_2
4434-15.fasta	M. Intracellulare subsp. chimera lineage 2	M. Intracellulare	99.9%				Find Resistance	Mycobacterium genhospites MAC-96 (GTDG placeholder, no standing in nomenclature)	0.1%
1911-15.fasta	M. Intracellulare subsp. chimera lineage 2	M. Intracellulare	99.8%				Find Resistance	Mycobacterium Intracellulare subsp. Intracellulare	0.3%
1913-15.fasta	M. Intracellulare subsp. chimera lineage 2	M. Intracellulare	99.5%				Find Resistance	Mycobacterium Intracellulare subsp. Intracellulare	0.4%
1914-15.fasta	Mycobacterium Intracellulare subsp. chimera	M. Intracellulare	99.3%			Lineage ID	Find Resistance	Mycobacterium Intracellulare subsp. Intracellulare	0.4%



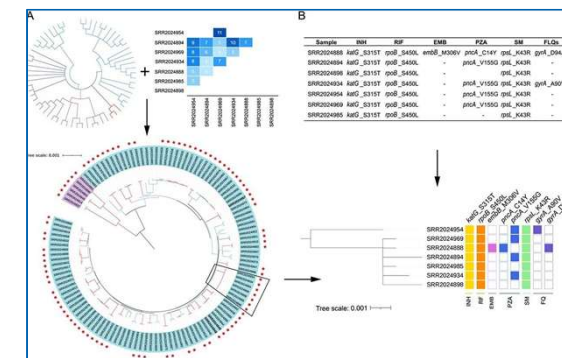
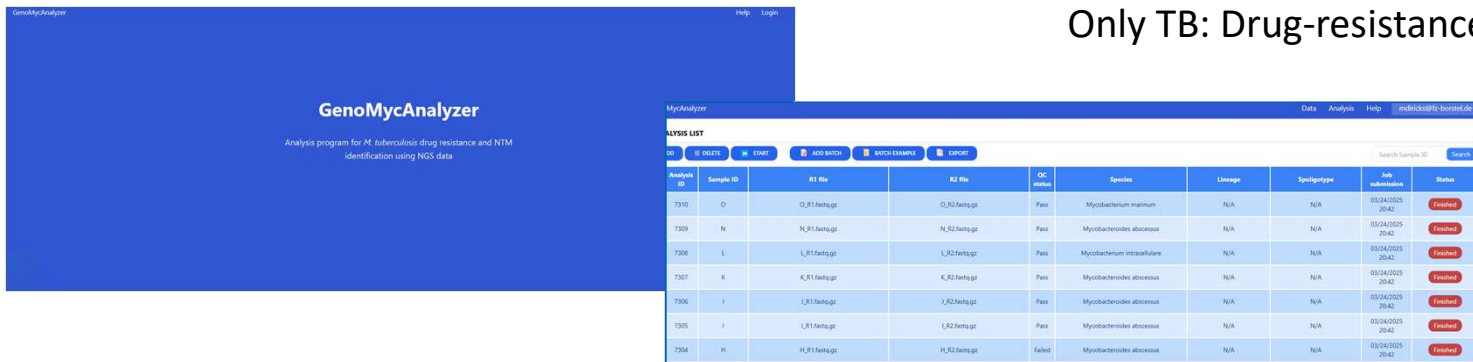
- **Access:**
 - CLI NTMseq: <https://github.com/ngs-fzb/NTMseq>
- **Author:** Margo Diricks (Research Center Borstel)
- **Input:** FastQ
- **Output:** quality control, (sub)species ID, MLST sequence types, assemblies, drug resistance, plasmid prediction, (approximate) phylogenies
- **Principle for NTM ID:** NTM-profiler
- **Comments**
 - Still under development





Japan (Matsumoto, 2019) - 184 genes - nanopore data

China (Yang, 2022) - uses mlstverse, kraken2 – only Illumina
Only TB: Drug-resistance prediction for 17 drugs, phylogeny



South Korea (Kim, 2024) – kraken2) – only Illumina

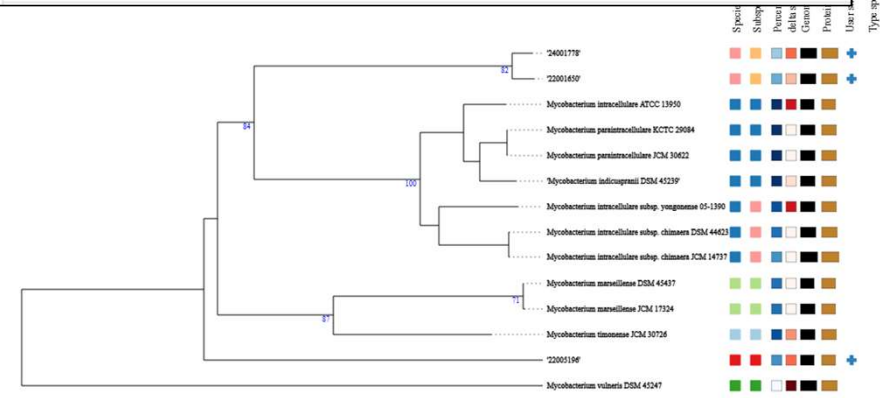
- **Access:**
 - Web-based: <https://tygs.dsmz.de/>
- **Author:** DSMZ (Germany)
- **Input:** FastA
- **Output:** Species ID, phylogeny
- **Principle for NTM ID**
 - Genome-to-Genome Distance Calculator, dDDH
- **Comments**
 - No subspecies information and no resistance prediction
 - Good to identify novel species
 - Data gets removed after period of time!
 - Processing currently takes quite long



Table 2: Identification

Your strain '24010731' belongs to species *Mycobacterium marinum*.

Potential new species detected: your strain '24001778' does not belong to any species found in TYGS database. ?



Methods, Results and References

Materials and Methods

The genome sequence data were uploaded to the Type (Strain) Genome Server (TYGS), a free bioinformatics platform available under <https://tygs.dsmz.de>, for a whole genome-based taxonomic analysis [1]. The analysis also made use of recently introduced methodological updates and features [2]. Information on nomenclature, synonymy and associated taxonomic literature was provided by TYGS's sister database, the List of Prokaryotic names with Standing in Nomenclature (LPSN, available at <https://lpsn.dsmz.de>) [3]. The results were provided by the TYGS on 2025-03-29. The TYGS analysis was subdivided into the following steps:

Determination of closely related type strains

Determination of closest type strain genomes was done in two complementary ways: First, all user genomes were compared against all type strain genomes available in the TYGS database via the MASH algorithm, a fast approximation of intergenomic relatedness [3], and the ten type strains with the smallest MASH distances chosen per user genome. Second, an additional set of ten closely related type strains was determined via the 16S rDNA gene sequences. These were extracted from the user genomes using RNAmmer [4] and each sequence was subsequently BLASTed [5] against the 16S rDNA gene sequence of each of the currently 22001 type strains available in the TYGS database. This was used as a proxy to find the best 50 matching type strains (according to the k-mer) for each user genome and to subsequently calculate precise distances using the Genome BLAST Distance Phylogeny approach (GBDP) under the algorithm 'coverage' and distance formula d_5 [6]. These distances were finally used to determine the 10 closest type strain genomes for each of the user genomes.

Pairwise comparison of genome sequences

For the phylogenetic inference, all pairwise comparisons among the set of genomes were conducted using GBDP and accurate intergenomic distances inferred under the algorithm 'trimming' and distance formula d_5 [6]. 100 distance replicates were calculated each. Digital DDH values and confidence intervals were calculated using the recommended settings of the GGDC 4.0 [2,6].

Phylogenetic inference

The resulting intergenomic distances were used to infer a balanced minimum evolution tree with branch support via FASTME 2.1.6.1 including SPR postprocessing [7]. Branch support was inferred from 100 pseudo-bootstrap replicates each. The trees were rooted at the midpoint [8] and visualized with Phy3D [9].

Type-based species and subspecies clustering

The type-based species clustering using a 70% dDDH radius around each of the 11 type strains was done as previously described [1]. The resulting groups are shown in Table 1 and 4. Subspecies clustering was done using a 79% dDDH threshold as previously introduced [10].

Table 3: Pairwise comparisons of user genomes vs. type strain genomes

Copy CSV PDF Excel Print Search:

Query strain	Subject strain	dDDH (d_0 , in %)	C.I. (d_0 , in %)	dDDH (d_4 , in %) ★	C.I. (d_4 , in %)	dDDH (d_5 , in %)	C.I. (d_5 , in %)	G+C content difference (in %)
'22001650'	'24001778'	95.2	[93.0 - 96.8]	95.4	[93.9 - 96.6]	96.9	[95.5 - 97.9]	0.06
'24001778'	<i>Mycobacterium intracellulare</i> subsp. <i>chimaera</i> JCM 14737	78.2	[74.3 - 81.8]	51.0	[48.4 - 53.7]	74.5	[71.0 - 77.7]	0.25

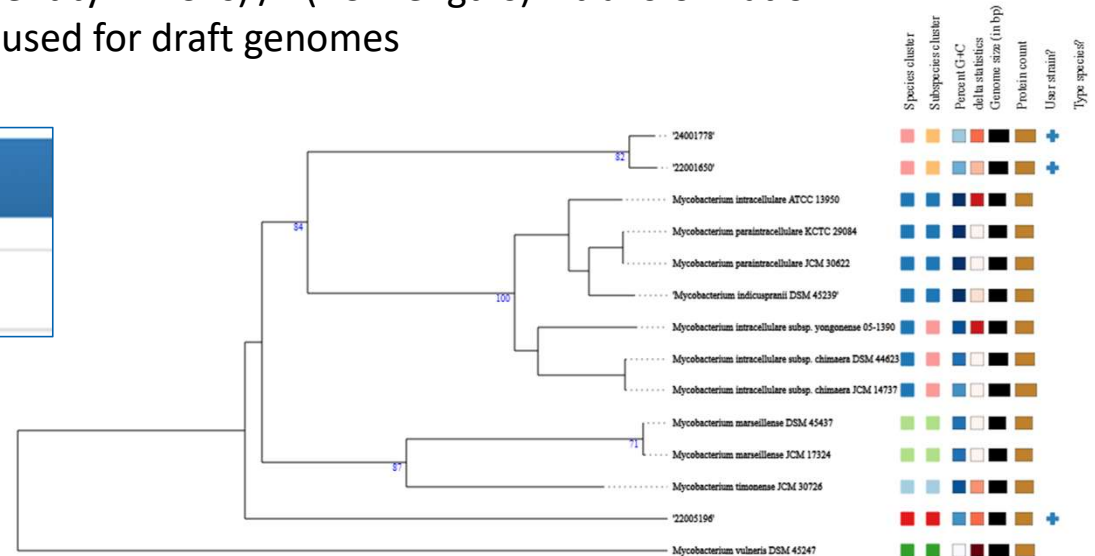
$dDDH\ d_4 \approx \Sigma(\text{identity in HSPs}) / \Sigma(\text{HSP lengths}) + \text{transformation}$
 → Can also be used for draft genomes

Table 2: Identification

Your strain '24010731' belongs to species *Mycobacterium marinum*.

Table 2: Identification

- Potential new species detected: your strain '22001650' does not belong to any species found in TYGS database. ?
- Potential new species detected: your strain '24001778' does not belong to any species found in TYGS database. ?
- Potential new species detected: your strain '22005196' does not belong to any species found in TYGS database. ?



General tools for typing and comparative genomics

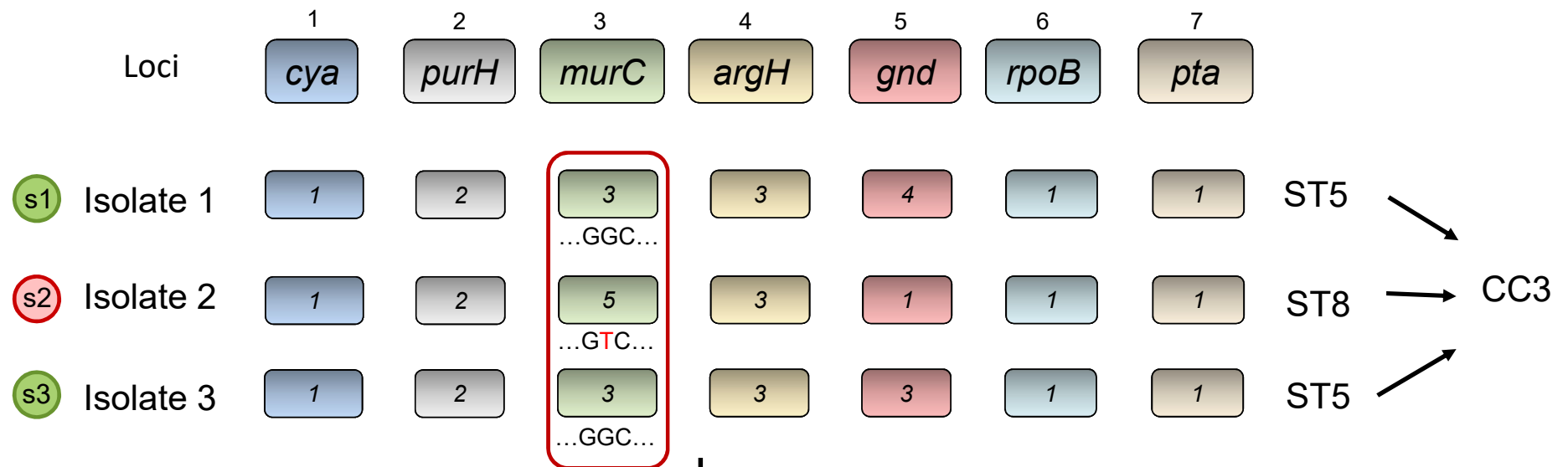
- Are my isolates closely related?
 - Is there any (indirect) person-to-person transmission or an outbreak in my hospital?
 - What is the source of the outbreak/transmission (e.g. compare environmental with clinical isolates)?
 - Did my patient get a relapse (same strain) or reinfection (different strain)?
 - Do my isolates belong to one of the dominant circulating clones?
 - How genetically diverse is a specific NTM species in my country/city or globally (population structure)?



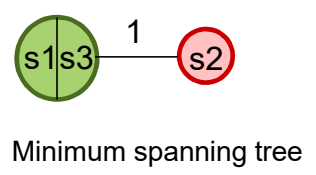
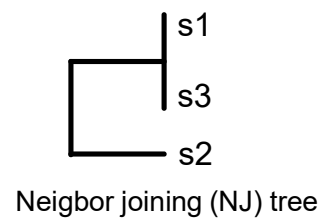
Can be answered with

- Multi-locus sequence typing (MLST) – e.g. using pubMLST, Ridom Typer, SRST2, mlst,...
- core genome multi-locus sequence typing (cgMLST) – e.g. using pubMLST, Ridom Typer, chewbbaca,...
- core genome single nucleotide polymorphism analysis (cgSNP) – e.g. using snippy, MTBseq,...

MULTI-LOCUS SEQUENCE TYPING



CC3

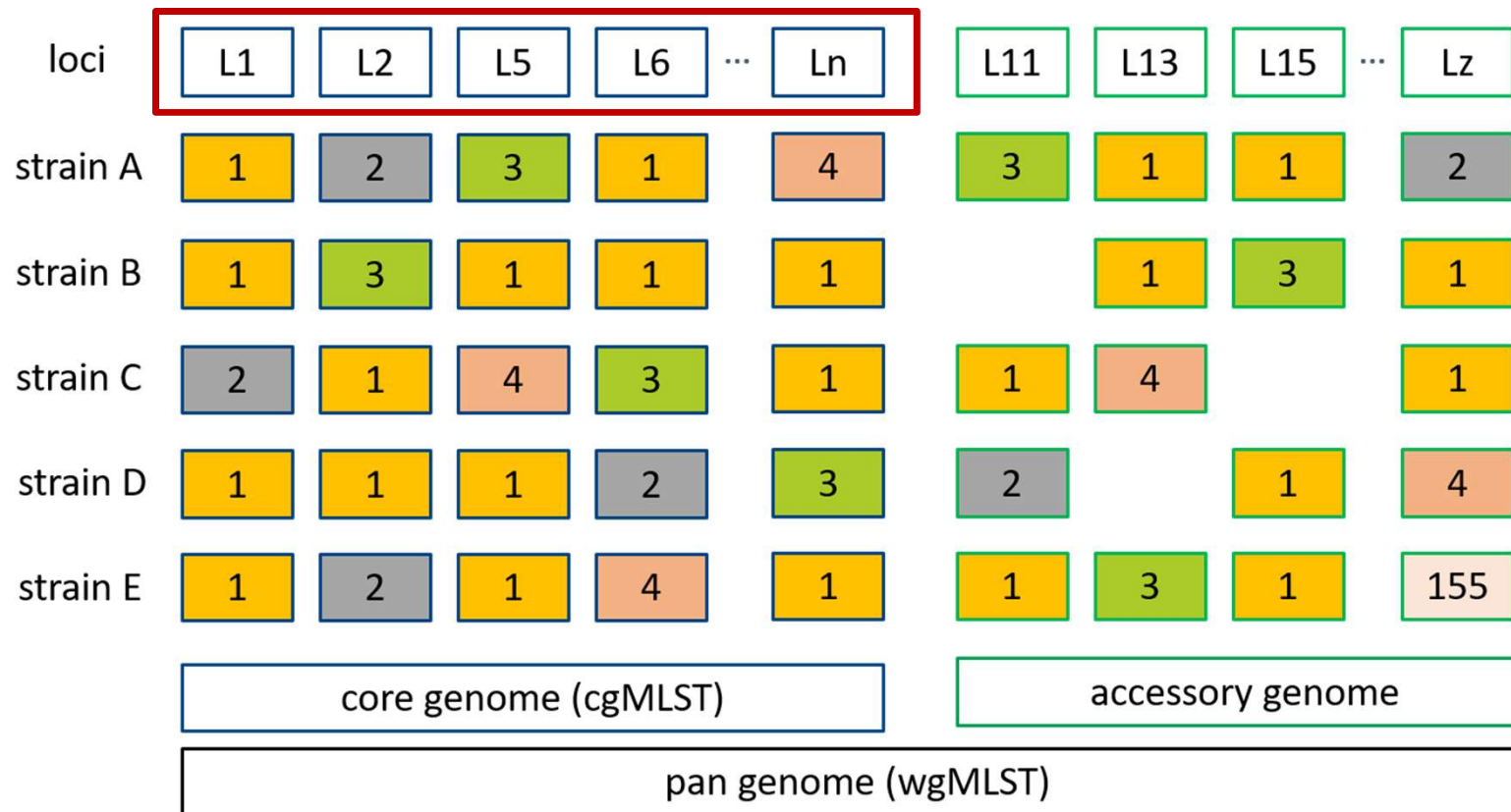


Only available for *M. abscessus*

CORE GENOME MULTI-LOCUS SEQUENCE TYPING



cgMLST scheme





Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

EBioMedicine
Published by THE LANCET

Research Paper

Harmonized Genome Wide Typing of Tubercle Bacilli Using a Web-Based Gene-By-Gene Nomenclature System



Thomas A. Kohl ^a, Dag Harmsen ^c, Jörg Rothgänger ^d, Timothy Walker ^e, Roland Diel ^f, Stefan Niemann ^{a,b,*}

Delineating *Mycobacterium abscessus* population structure and transmission employing high-resolution core genome multilocus sequence typing

Margo Diricks [✉], Matthias Merker, Nils Wetzstein, Thomas A. Kohl, Stefan Niemann & Florian P. Maurer

Nature Communications **13**, Article number: 4936 (2022) | [Cite this article](#)

2855 Accesses | 5 Citations | 22 Altmetric | [Metrics](#)

cgMLST schemes:

M. tuberculosis
2,891 genes (72%)

M. abscessus
2,904 genes (59%)

- Steps of new scheme creation

Scheme creation

- Select core genes using a diverse isolate collection

Scheme validation

- Test scheme in a larger dataset
 - Test robustness
- Compare with other genotyping methods

Threshold calibration

- Analyse epidemiologically related samples (e.g. type strains, sequential samples, outbreaks)
- Determine allele distance threshold

Delineating *Mycobacterium abscessus* population structure and transmission employing high-resolution core genome multilocus sequence typing

Margo Diricks , Matthias Merker, Nils Wetzstein, Thomas A. Kohl, Stefan Niemann & Florian P. Maurer

Nature Communications 13, Article number: 4936 (2022) | [Cite this article](#)

2855 Accesses | 5 Citations | 22 Altmetric | [Metrics](#)

- Examples of bioinformatic tools for (cg)MLST analysis
 - pubMLST: web-based GUI (<https://pubmlst.org/>) → only for *M. abscessus*
 - Ridom Typer: desktop GUI (<https://www.ridom.de/ridom-typer/>)
 - Pathogenwatch: web-based GUI (<https://pathogen.watch/>) → only for *M. abscessus* (based on pubMLST)
 - ChewBBaca: CLI (<https://chewbbaca.readthedocs.io/en/latest/>) → no scheme available for NTM

- **Access**
 - Web-based GUI: <https://pubmlst.org/>
- **Author:** Keith Jolley et al. (University of Oxford)
- **Input:** FastA
- **Output:** (cg)MLST ST, isolate/genome collection, trees,...
- **Comments**
 - Free but requires login to access latest data
 - Currently only (cg)MLST schemes for *M. abscessus*



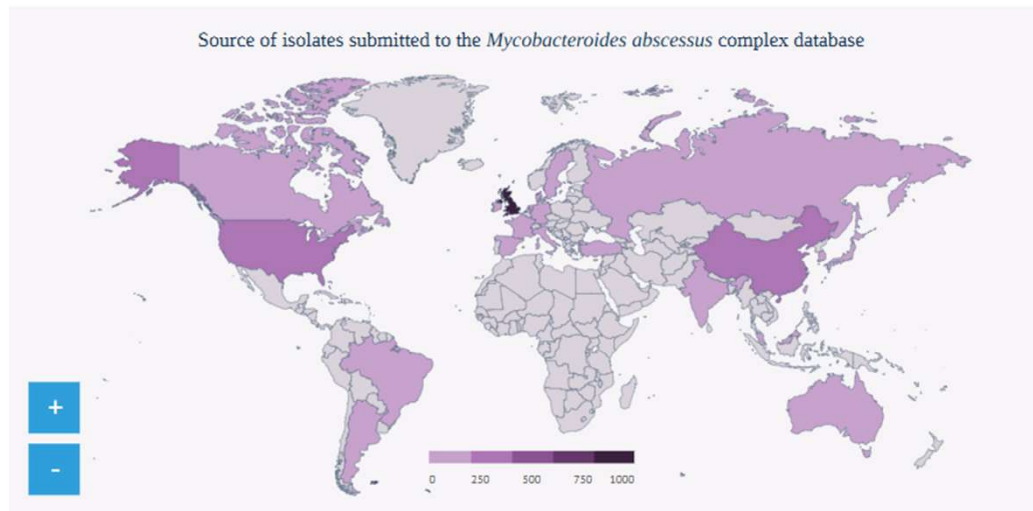
PubMLST Public databases for molecular typing and microbial genome diversity

A collection of open-access, curated databases that integrate population sequence data with provenance and phenotype information for over 140 different microbial species and genera.

47,470,754 ALLELES	2,044,340 ISOLATES	1,460,955 GENOMES
-----------------------	-----------------------	----------------------

Mycobacteroides abscessus complex

<https://pubmlst.org/>



[SUBMIT](#)

This MLST scheme was developed by Song Yee Kim and colleagues at the Yonsei University College of Medicine, Seoul, Korea. It is described in Kim *et al.* 2013 *Diagn Microbiol Infect Dis* **77**:143-9 *et.*

Database curated by Margo Diricks.

The separate scheme for *M. massiliense* has been removed since this was leading to confusion with the *M. abscessus* scheme as they used the same loci with different allele numbers. Allele and profile definitions for the legacy *M. massiliense* scheme are available for download [here](#).

The preferred citation for this website is:

Jolley *et al.* *Wellcome Open Res* 2018, **3**:124 [version 1; referees: 2 approved] *et.*

Typing

The typing database contains nomenclature - allele definitions that provide an identifier for every unique allele sequence, and MLST profiles that index each unique combination of alleles with a sequence type (ST).

Allele sequences: 191,889

Last updated: 2026-02-10

Isolate collection

The isolate database consists of isolate records containing provenance and phenotype information linked to molecular typing information. These records may also include genome assemblies.

Isolates: 2,123

Last updated: 2026-02-28

Genome collection

A subset of records within the isolate database may contain genomes assemblies. You can access these from the isolate database by filtering on sequence bin size in a query.

Genomes: 2,079

Last updated: 2025-05-27

Typing

The typing database contains nomenclature - allele definitions that provide an identifier for every unique allele sequence, and MLST profiles that index each unique combination of alleles with a sequence type (ST).

Allele sequences: 191,889

Last updated: 2026-02-10

Isolate collection

The isolate database consists of isolate records containing provenance and phenotype information linked to molecular typing information. These records may also include genome assemblies.

Isolates: 2,123

Last updated: 2026-02-28

Genome collection

A subset of records within the isolate database may contain genomes assemblies. You can access these from the isolate database by filtering on sequence bin size in a query.

Genomes: 2,079

Last updated: 2025-05-27

Nomenclature:

Query a sequence

Single sequence

Query a single sequence or whole genome assembly to identify allelic matches.

Batch sequences

Query multiple independent sequences in FASTA matches.

Uploaded file: H.fasta

7 exact matches found.

Locus	Allele	Length	Contig	Start position	End position
argH	1	510	contig20001_lem=1531651_cov=61.3_cor=0_origname=Contig_11_61_3418_pilon_svwshovills-skeas/1.1.0_date=202111201	490151	490660
cya	2	546	contig20001_lem=169208_cov=6.6_cor=0_origname=Contig_11_61_3752_pilon_svwshovills-skeas/1.1.0_date=202111201	490649	491188
gnd	3	506	contig20001_lem=1531651_cov=61.3_cor=0_origname=Contig_11_61_3418_pilon_svwshovills-skeas/1.1.0_date=202111201	505839	506364
murC	2	445	contig20001_lem=1531651_cov=61.3_cor=0_origname=Contig_11_61_3418_pilon_svwshovills-skeas/1.1.0_date=202111201	45742	46186
pta	2	520	contig20008_lem=189718_cov=68.3_cor=0_origname=Contig_19_68_252_pilon_svwshovills-skeas/1.1.0_date=202111201	128141	128660
purH	25	497	contig20004_lem=318344_cov=63.5_cor=0_origname=Contig_18_63_4806_pilon_svwshovills-skeas/1.1.0_date=202111201	224905	225401
rpoB	4	503	contig20009_lem=153652_cov=63.9_cor=0_origname=Contig_2_63_8717_pilon_svwshovills-skeas/1.1.0_date=202111201	139878	140380

Only exact matches are shown above. If a locus does not have an exact match, try querying specifically against that locus to find the closest match.

MLST

Matching profile

ST: 222

Find alleles

By specific criteria

Find alleles by matching criteria (all loci together)

By locus

Select, analyse and download specific

Search for allelic profiles

By specific criteria

Search, browse or enter list of profiles

By allelic profile

This can include partial matches to find

PubMLST Public databases for molecular typing and microbial genome diversity

HOME ORGANISMS SPECIES ID ABOUT US UPDATES

Home » Organisms » *Mycobacteriales: abscessus complex* » *Mycobacteriales: abscessus complex* typing » Scheme information

Scheme information - MLST

Curator: This scheme is curated by: Mergo Diricks, Fz Borstel

Fields: ST, [Hemolysin](#), [clonal_complex](#)

Loci: This scheme consists of alleles from 7 loci: argH, cya, gnd, murC, pta, purH, rpoB

Profiles: This scheme has 320 profiles defined.

Select date range: all time

Profile update history

```
>argH_1
GTGAGCACCAACGAAGGCTCGCTGTGGGGCGGCCGGTTCGCCGGCGGCCAGCCCGGGG
CTGGCGGACCTGAGCAATCCACTCATTTTCGATTGGGTGCTGGCGCCCTACGACATTCAG
GGTGGTGGTGGCCATGCCGAGTGTCTCCGAAGCGGGCCGTCTCACGAAGAACAATC
GGTGGCCCTGGTCGAGGGCTTAGAGCAGCTGGTTCGGGATGTCGGCTGGCGCGTTCGT
CCAGCCGATACCGACGAGGAGCTACACGGTGCCTCGAACCGGGCTGATCGAACGGGTG
GGCCCTGACATCGCGGGCGCTCTCGCTGGCCGATCCGAAACGATCAGGTGGCAACA
CTGTTTCGGATGTGGTGCAGCAGCGGGCCGACGATCTCCGAGGCCTGCTCGAGCTC
GTGAGGCATTGGCGGATCAGGCCGGGGCACATCCGGATCGGATCATGCGGGCAAGACT
CATTTCGAGGCCTGCCCAACCGGTGCTCTT

>argH_2
GTGAGCACCAACGAAGGCTCGCTGTGGGGCGGCCGGTTCGCCGGCGGCCAGCCCGGGG
CTGGCGGACCTGAGCAATCCACTCATTTTCGATTGGGTGCTGGCGCCCTACGACATTCAG
GGTGGTGGTGGCCATGCCGAGTGTCTCCGAAGCGGGCCGTCTCACGAAGAACAATC
GGTGGCCCTGGTCGAGGGCTTAGAGCAGCTAGGTCGGGATGTCGGCTGGCGCGTTCGT
CCAGCCGATACCGACGAGGAGCTACACGGTGCCTCGAACCGGGCTGATCGAACGGGTG
GGCCCTGACATCGCGGGCGCTCTCGCTGGCCGATCCGAAACGATCAGGTGGCCACA
CTGTTTCGGATGTGGTGCAGCAGCGGGCCGACGATCTCCGAGGCCTGCTCGAGCTC
GTGAGGCATTGGCGGATCAGGCCGGGGCACATCCGGATCGGATCATGCGGGCAAGACT
CATTTCGAGGCCTGCCCAACCGGTGCTCTT
```

ST	argH	cya	gnd	murC	pta	purH	rpoB	clonal_complex
1	2	1	1	2	2	3	1	
2	4	1	1	2	7	4	4	
3	5	1	1	4	5	6	1	
4	2	1	7	2	9	2	4	
5	3	1	4	3	1	2	1	1
6	3	1	4	8	1	2	1	1
7	3	1	4	3	1	2	3	1
8	3	1	4	5	1	2	1	1
9	1	2	1	2	4	4	1	2
10	4	2	1	7	10	2	1	
11	1	2	3	1	3	1	2	
12	1	2	1	2	2	5	1	
13	3	2	4	3	1	2	1	1
14	1	2	5	1	2	1	2	

Typing

The typing database contains nomenclature - allele definitions that provide an identifier for every unique allele sequence, and MLST profiles that index each unique combination of alleles with a sequence type (ST).

Allele sequences: 191,889

Last updated: 2026-02-10

Isolate collection

The isolate database consists of isolate records containing provenance and phenotype information linked to molecular typing information. These records may also include genome assemblies.

Isolates: 2,123

Last updated: 2026-02-28

Genome collection

A subset of records within the isolate database may contain genomes assemblies. You can access these from the isolate database by filtering on sequence bin size in a query.

Genomes: 2,079

Last updated: 2025-05-27

Search or browse database

Enter search criteria or leave blank to browse all records. Modify form parameters to filter or enter a list of values.

Isolate provenance/primary metadata fields

id = Enter value... + ⓘ

Sequence bin

total length (Mbp) >= 1 + ⓘ

Display/sort options

Order by: id

Display: 25 records per page

Action

RESET SEARCH

Isolate count

2,079

▲0 [month]

Genome count

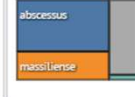
2,079

▲0 [month]

Continent



Subspecies



Year



Analysis tools

Breakdown: Fields Two Field Combinations Publications Sequence bin

Analysis: Codons Gene Presence GeneScanner Genome Comparator BLAST rMLST species Id PCR

Export: Dataset Contigs Sequences

Third party: GrapeTree ITOL PhyloViz ReporTree

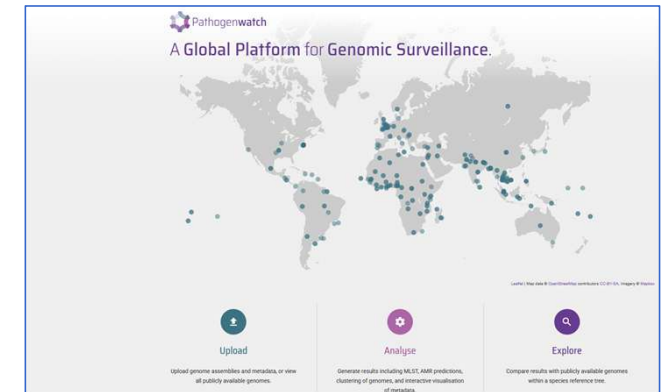
id	isolate	aliases	country	subspecies	year	source	MLST									
							argH	cya	gnd	murC	pta	purH	rpoB	ST	clonal complex	
1	ATCC 19977	CU458896; GCA_000069185.1; NC_010397	Unknown	abscessus		unknown	3	1	4	3	1	2	1	5	1	
2	ERR114969		Unknown	massiliense		unknown	30	31	17	27	21	31	18	33	3	
3	ERR114985		Unknown	massiliense		unknown	30	31	17	27	21	31	18	33	3	
4	ERR233333		UK		1998	unknown	16	6	18	14	14	16	7	41		
5	ERR369183		Unknown	abscessus		unknown	1	2	1	2	4	4	1	9	2	
6	bolletii 50594	GCA_000445035.1; NC_021278.1; NC_021279.1; NC_021282.1	Unknown	bolletii		unknown	31	31	17	28	16	32	19	120		
7	bolletii CCUG 48898	AHAR01; GCA_000239055.2	Unknown	bolletii		unknown	30	31	19	28	16	15	20	63	7	
8	1518	GCA_000523895.1; JAOI01	USA	abscessus		unknown	3	1	4	3	1	2	1	5	1	

BRUKER MBIOSEQ RIDOM TYPER (SEQSPHERE+)

- **Access**
 - Desktop GUI: <https://www.ridom.de/ridom-typer/index.shtml>
 - Nomenclature: <https://cgmlst.org/ncs>
- **Author:** Ridom (now Bruker)
- **Input:** FastQ, FastA, BAM
- **Output:** Database, species ID, (resistance), Quality control, (cg)MLST ST, assemblies (also ONT), trees,...
- **Comments**
 - Commercial
 - Early warning alerts
 - Continuous processing with pipeline
 - Publicly available scheme only available for *M. abscessus*, but creation of your own (ad hoc) cgMLST schemes possible




- **Access**
 - **Web-based:** <https://pathogen.watch/>
- **Author:** Wellcome Sanger institute (UK)
- **Input:** FastQ and FastA
- **Output:** Species ID, (cg)MLST ST
- **Principle for NTM ID:** Search against NCBI RefSeq/curated library with mash
- **Comments**
 - Global Platform for Genomic Surveillance
 - No subspecies information and no resistance prediction
 - Very fast for fastA (few sec), but slower for fastQ (>20 min per sample)



☰ List Map Stats Viewing 16 of 16 genomes

<input type="checkbox"/> Name	Organism
<input type="checkbox"/> MABA-NC_010397.1-Ripoll2009	<i>Mycobacteroides abscessus</i>
<input type="checkbox"/> Madipatum_GCF0016445751_YC-RL4_TypeStrain	<i>Mycobacterium</i> sp. YC-RL4
<input type="checkbox"/> Marupensis_GCF0020865151_DSM44942_TypeS	<i>Mycolicibacter arupensis</i>
<input type="checkbox"/> Mattenuatum_GCF9005660851_MK41_TypeStrain	<i>Mycobacterium attenuatum</i>

H 
Mycobacteroides abscessus

MLST - Multilocus sequence typing
Source: *Mycobacterium abscessus* - PubMLST

Sequence type
222

[View all ST 222](#)

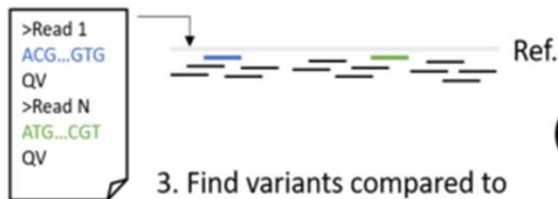
Profile						
argH	cya	gnd	murC	pta	purH	rpoB
1	2	3	2	2	25	4

A. cgSNP analysis

1. Selection of reference genome



2. Map reads to reference genome



3. Find variants compared to reference for each isolate

	IG Pos. 1	Gene N Pos. X	Gene N Pos. Y
Isolate 1	C	A	G
Isolate 2	C	G	T

Distance = 2 SNPs

4. Compare variants between isolates to calculate distance and determine thresholds

vs

B. cgMLST analysis

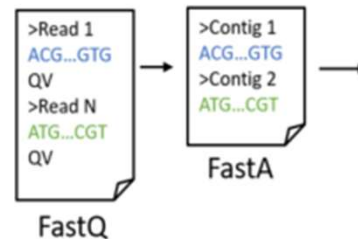
1. cgMLST scheme creation: define core genes and nomenclature

	Core genes			Accessory genes		Allele numbers
	Gene 1	Gene 3	Gene N	Gene 2	Gene Z	
Isolate X	1	1	1	1	1	}
Isolate Y	2	1	1		2	
Isolate Z	3	2	1	2		

Nomenclature:

gene 1 - allele 1 = ATGA...TGA, allele 2 = TGTA...TAA
gene 3 - allele 1 = ATGC...TAA, allele 2 = ATGT...TAA

2. Assemble reads into draft genomes



3. Determine cgMLST profile for each isolate

	Gene 1	Gene 3	Gene N
Isolate 1	1	4	1
Isolate 2	1	4	2

Distance = 1 allele

4. Compare profiles between isolates to calculate distance, validate scheme and determine thresholds

Single nucleotide polymorphisms

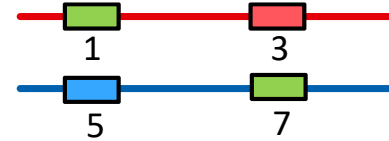
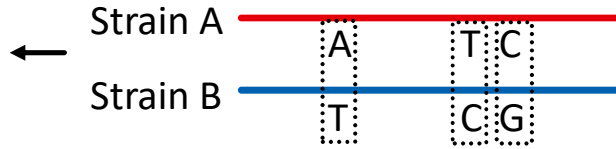
cgSNP

vs

cgMLST

Multi-locus sequence typing

Distance
=
3 SNPs



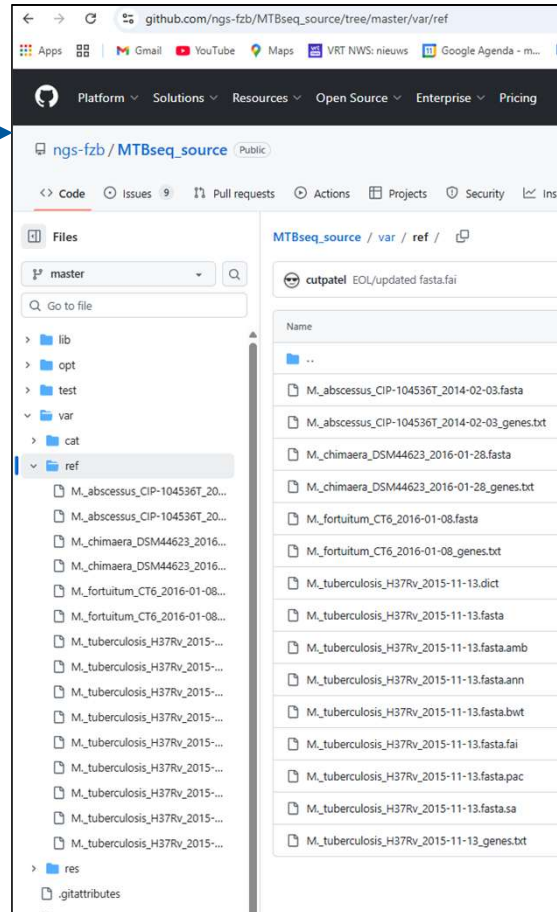
Strain A
Strain B

Distance
=
2 alleles

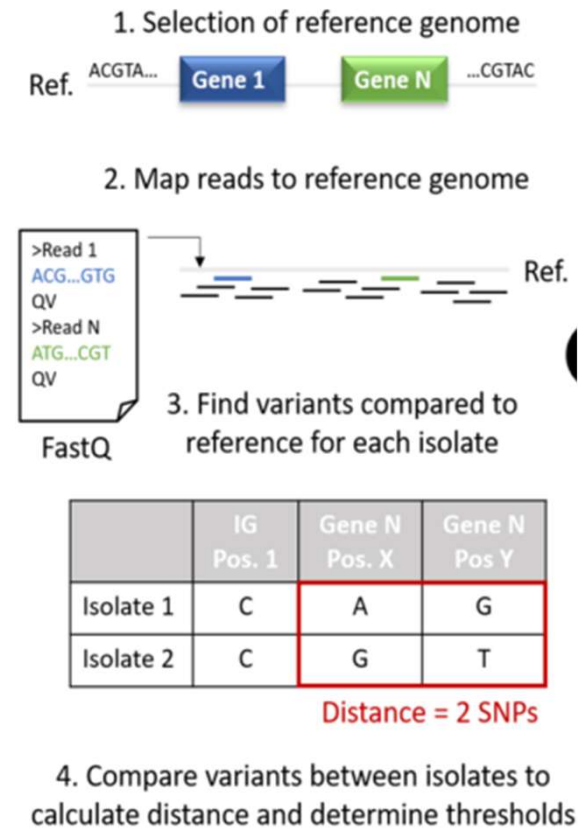
	cgSNP	cgMLST
Level	Single nucleotide substitutions	Genes
Computation power	high	low
Standardized	No	Yes
Resolution	Highest	High
Well suited for	In-depth comparison Retrospective analysis Cluster resolving	Comparison of diverse strains Prospective surveillance Cluster detection

Tools

- MTBseq
- Snippy
- ...
- Special things to consider
 - Removal of recombination?
 - Gubbins/ClonalFrameML



A. cgSNP analysis



- Many different tools and techniques
 - Both web-based and CLI-based
- All with their own (dis)advantages
 - Data gets removed after period of time
 - No subspecies prediction
 - No nice summaries
 - Slow
- All of them still under development/being validated and RUO!
- Try different tools