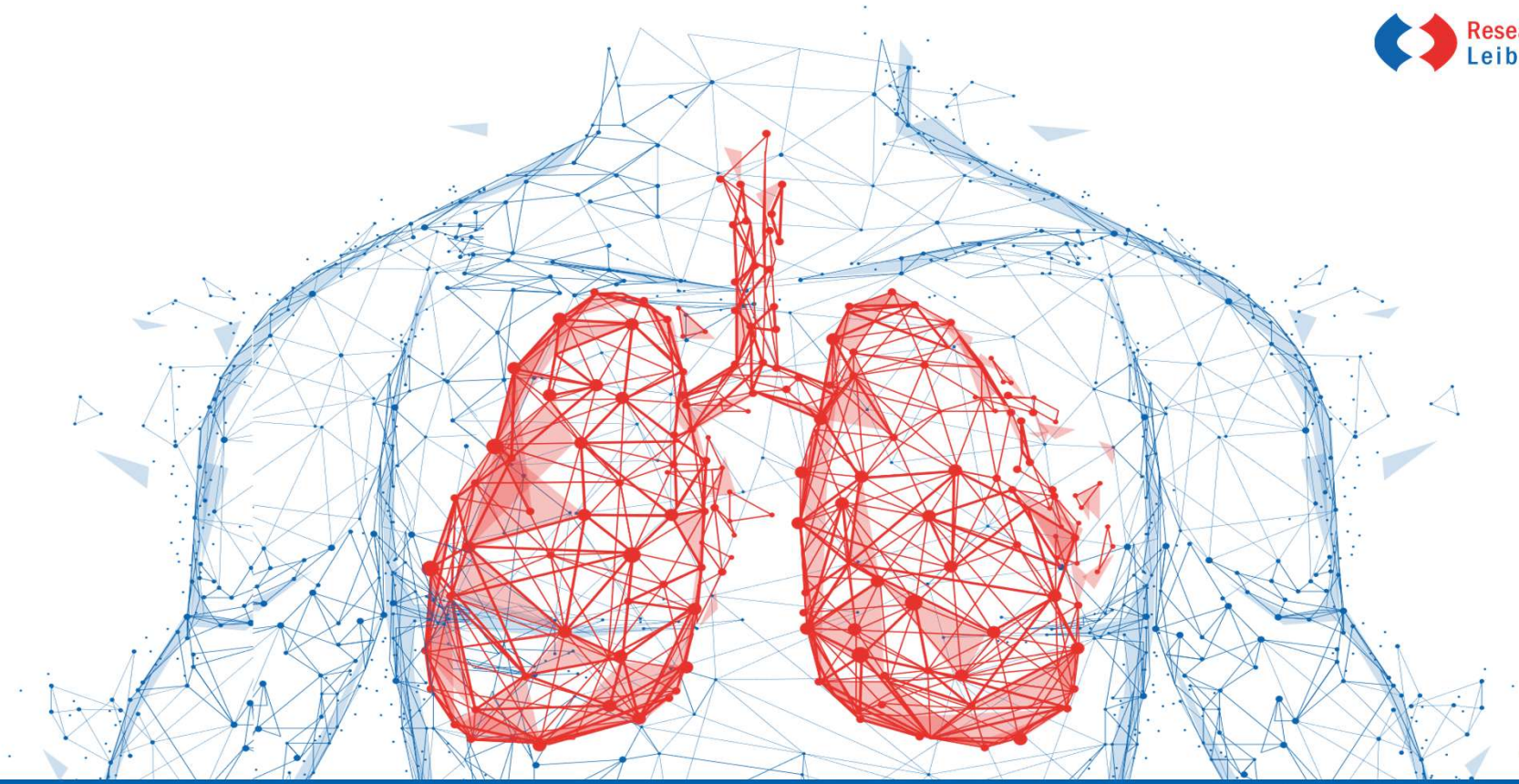
Leibniz Association

Research Center Borstel
Leibniz Lung Center



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

WHOLE GENOME SEQUENCING

Research Center Borstel
Leibniz Lung Center

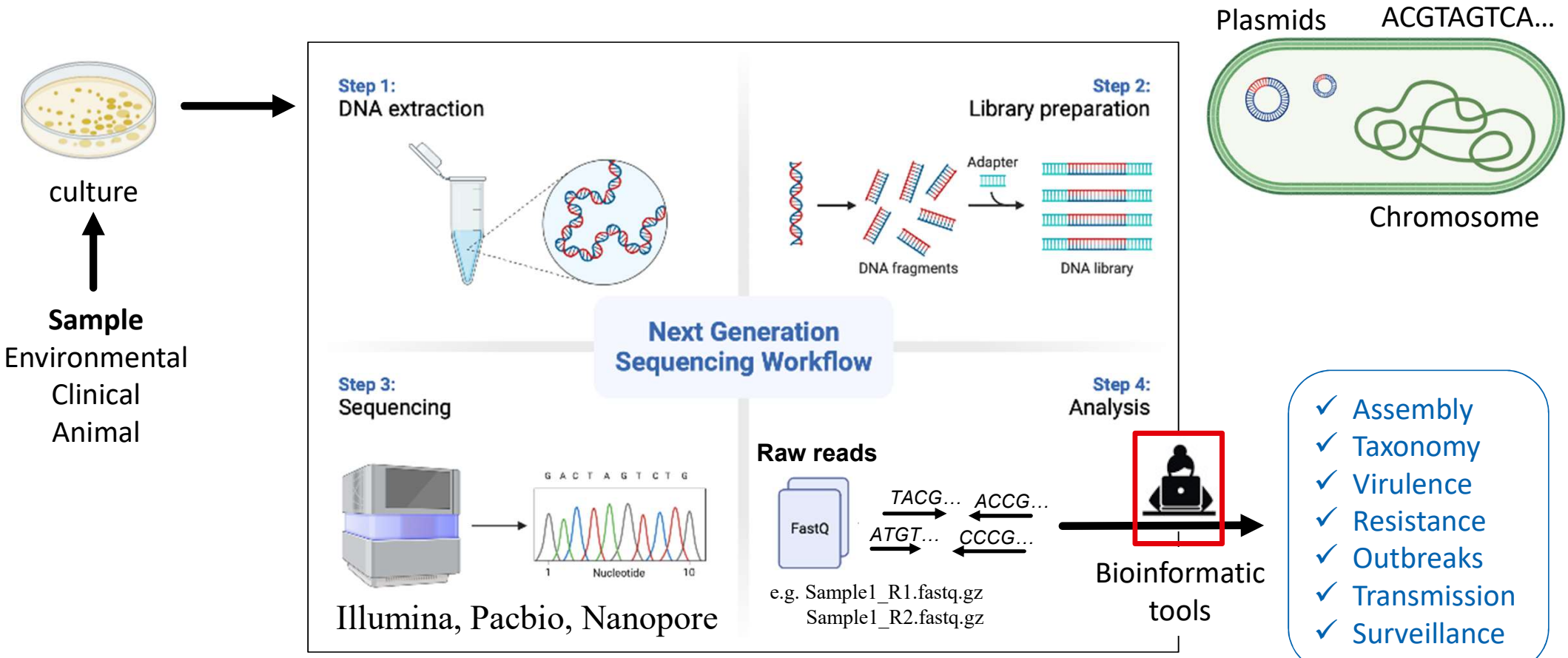
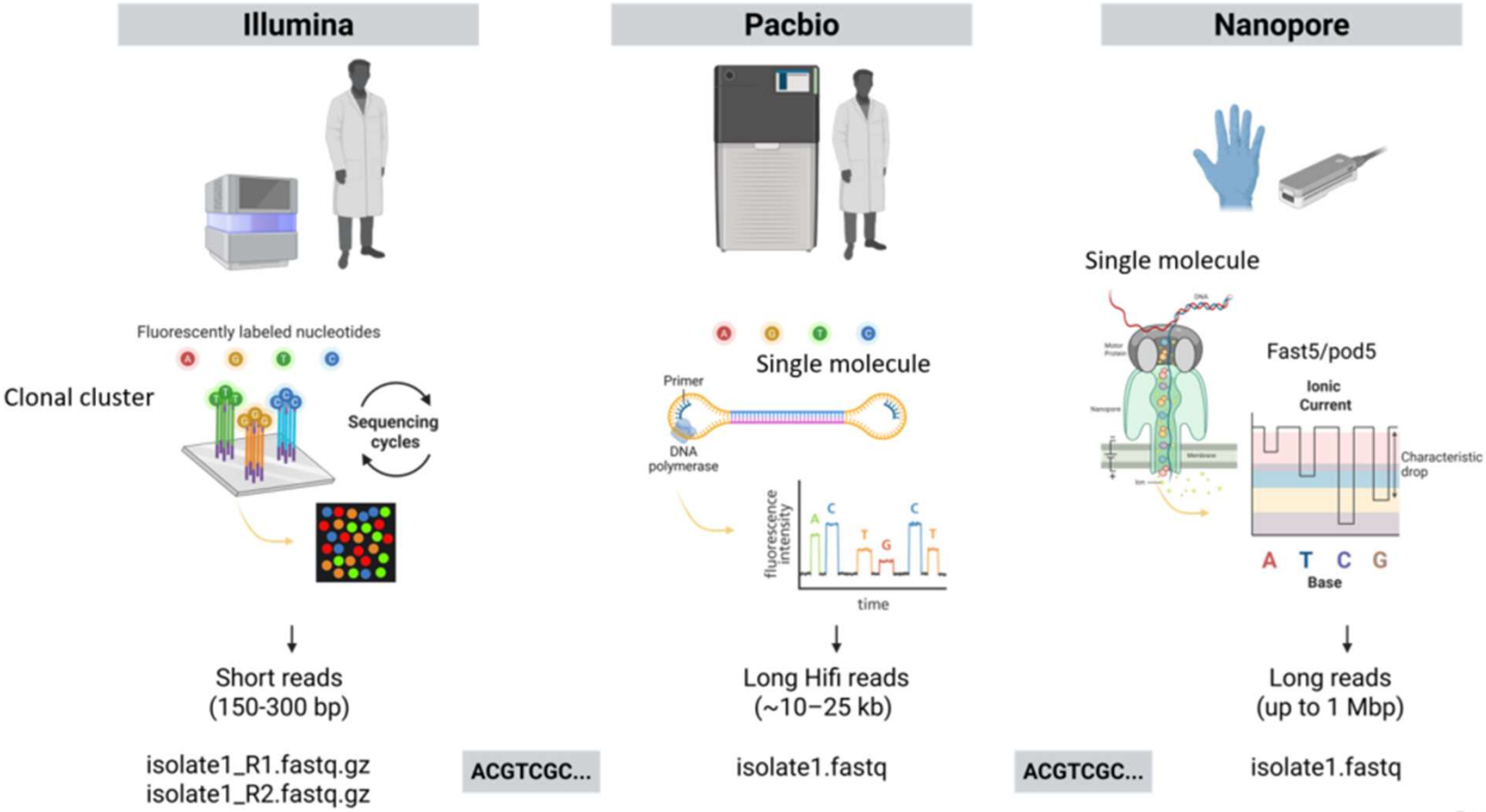


Figure adapted from biorender

WHOLE GENOME SEQUENCING TECHNOLOGIES

Research Center Borstel
Leibniz Lung Center



Created with biorender

SEQUENCING TECHNOLOGIES

Research Center Borstel
Leibniz Lung Center

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	☑ Yes (PCR / bridge amplification)	✗ No	✗ 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	✗ No (needs bisulfite, etc.)	☑ Yes (polymerase kinetics)	☑ 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

SEQUENCE DATA FILE FORMATS

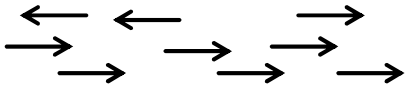
Research Center Borstel
Leibniz Lung Center

FastQ

~1 million reads

```
@read1
SEQUENCE
+
QUALITY
@read2
```

Illumina: ~150 bp
Pacbio (~15 kbp)
Nanopore (~5-50 kbp)



```
@HWI-ST911:111:C0N4WACXX:5:1101:2249:2216 1:N:0:TTAGGC CGATC:@@@FF
NATGGCACCATTAAAAAGAATGTTTTATATGGTGTTGAGAAGGACAAAGCTGAAGAAGAAATTTAGTCTGCACTTGATGTTGCAAATGCAAAGAAA
+
#2A2<CCFHIIIIIIIIIIIIGCCHIIIGIIIFFHIIDGHIGIIIIIIICHGIIIGGCECEGICFHCECDEFFFFDEEEEEDDDDDCDDCDDDDBC
@HWI-ST911:111:C0N4WACXX:5:1101:2509:2197 1:N:0:TTAGGC CGATC:1+4=B
NATGAGATAAATCAATTGTCTTTAATGAAGTACAGTCTTTGAATAATGAGTTTTGAACTCTTCTGCAACTTTTTGGAAACTTTAAAGTTTGTAATG
+
#4A2<AADHIIIIIIIIIHHIIIIIIIIIIFGIII@GIIFIIIGIIIIEIDHEHIIIHIIIIIIIIIIICHIIIHHEEDFFFFFEEECEEEADDFC
@HWI-ST911:111:C0N4WACXX:5:1101:3746:2179 1:N:0:TTAGGC CCATC:+11+A
NATGTCATCCATCTTTTCCTATCTAAAAAAGAATCAAAAAAGGGATAGTACAGAGGGAAAGTTCAATCCAGAGGACGATGAAACACTGATTGATGG
+
```

e.g. Isolate1_R1.fastq.gz

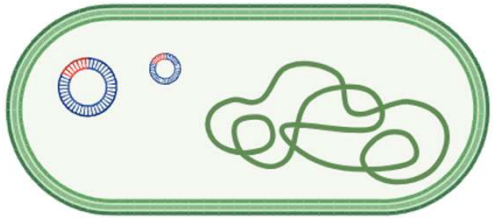
200 MB

Short reads:
SPAdes, skesa,…

De novo assembly

Long reads:
flye, canu,…

Plasmids (e.g. contig 2-3)



Chromosome (e.g. contig 1)

FastA

Pacbio/Nanpore:
1 contig
Illumina:
30-500 contigs

```
>contig1
SEQUENCE
>contig2
SEQUENCE
```

Illumina:
~1.5 Mbp
Pacbio/Nanopore:
5 Mbp

TACGACGTCGTCGTCG Contig 1 | *GTCTA* Contig 2 | *GTC* Contig 3

Draft genome

```
>contig00001 len=1551651 cov=62.3 corr=0 origname=Contig_11_6
CGCCACCGCCGCCCCCTCCGCCGGTGACGACGCATGAGGCGCCGCCGCCGGTGACTACCA
CGACGACACCGCCGCCGCCTCCGCCCCCGACGACCACGACACCGGCCGGGCCACGGCAGA
TCACGTACTCGGTGACCGGTTCCAAGGCTCCGCTCGATCGCATCTCGATCACCTGGACCG
ACGGTTCGGGACGCACCCGGGTGAACCCGAATGTGTACATCCCGTGGTCGATCACGGTTA
CCCCCATTTCGAATTCGGAGATCGGATCGGTGTCGGCGAGTAGCTTCCTGCGGCTGAGTC
AGCTCAACTGCACGATCACCACCAGCGACGGTCAGGTTTTGTCCTCCAACAACAACAATT
CGGCGCAGGCAACCTGCTGATGCCAGCAGGTCGATTGGCCAAGCCACTGGAGGCAATCGC
GGGCCCGCTGCGGCGGTCATCGCCGGAGTCCGTTGACCGGATGCTCATCGGCTTGTGCGC
```

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!!

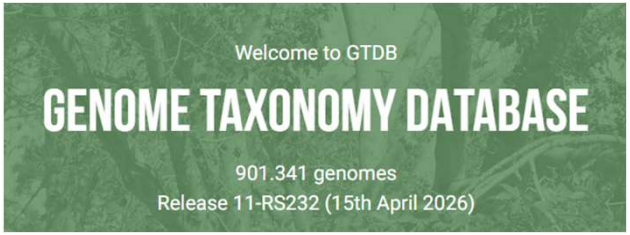
Mycobacteria/NTM-specific tools

NTM-PROFILER

- **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

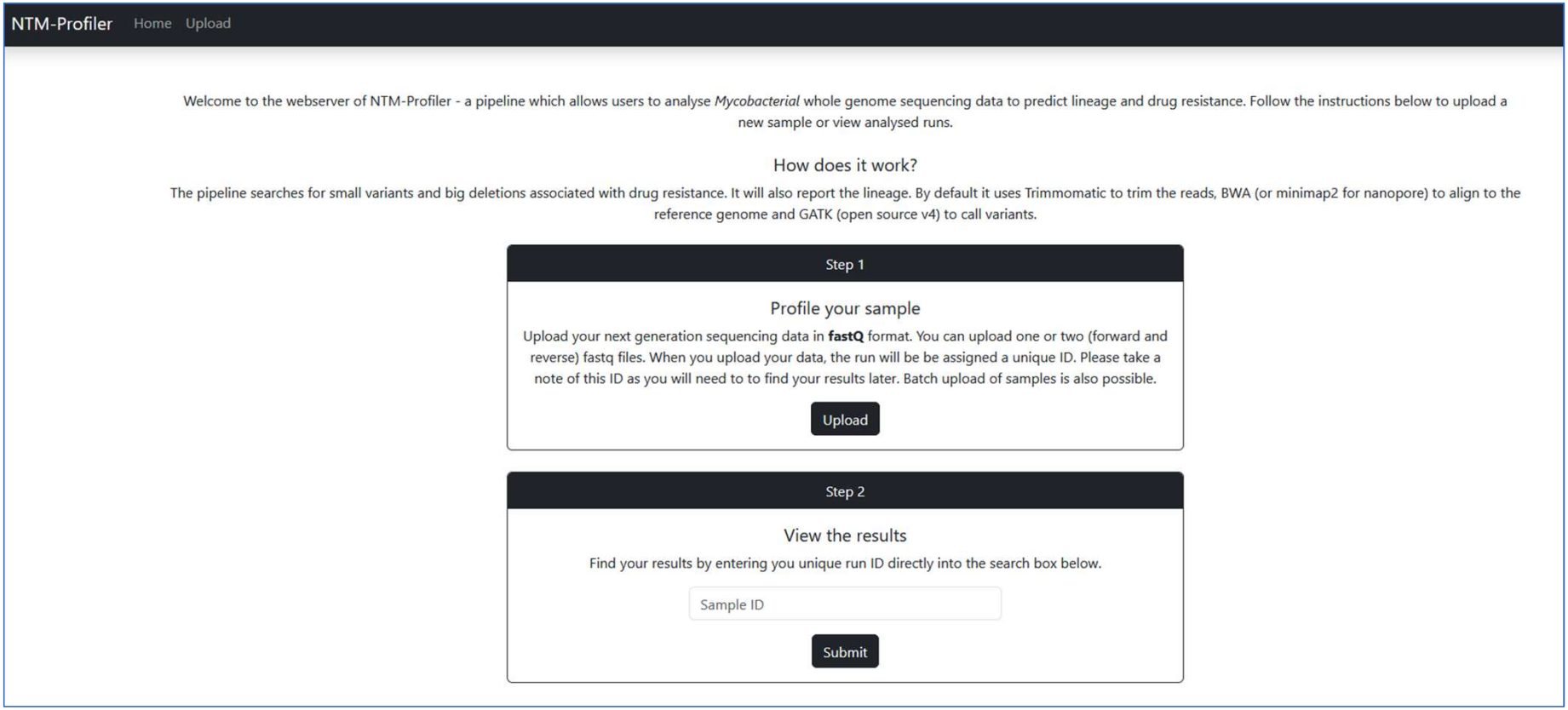
Jody Phelan

https://gtdb.ecogenomic.org/



NTM-PROFILER

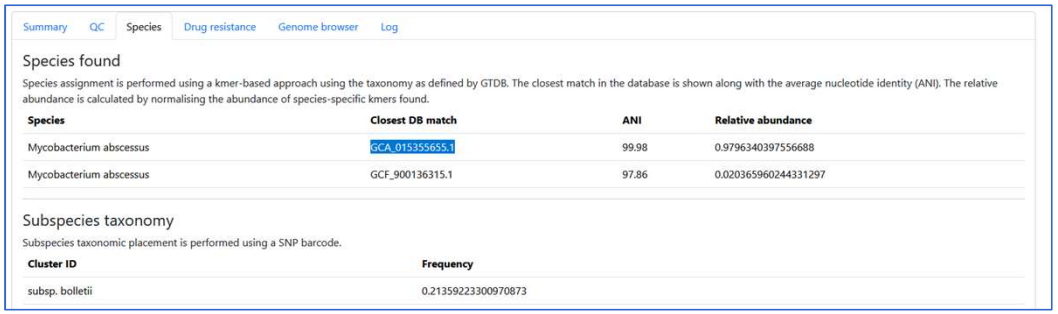
- **Web-based**: https://bioinformatics.lshtm.ac.uk/ntm-profiler/



Jody Phelan

NTM-PROFILER

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



Summary | QC | Species | Drug resistance | Genome browser | Log

Species found

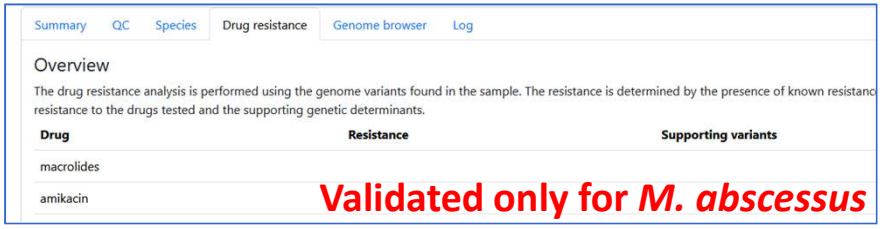
Species assignment is performed using a kmer-based approach using the taxonomy as defined by GTDB. The closest match in the database is shown along with the average nucleotide identity (ANI). The relative abundance is calculated by normalising the abundance of species-specific kmers found.

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

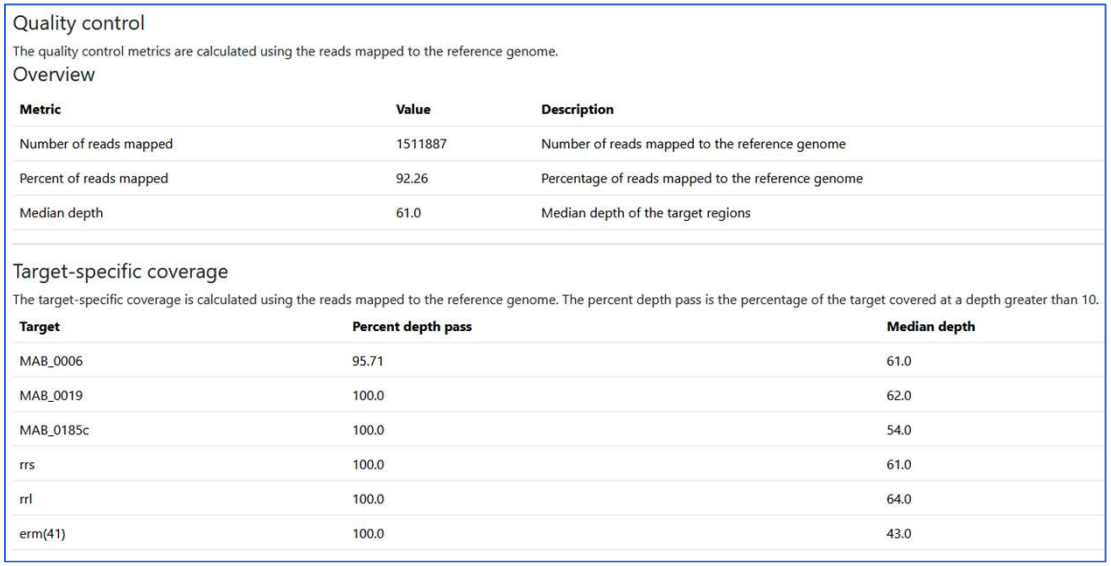
The drug resistance analysis is performed using the genome variants found in the sample. The resistance is determined by the presence of known resistanc
resistance to the drugs tested and the supporting genetic determinants.

Drug	Resistance	Supporting variants
macrolides		
amikacin		

Validated only for *M. abscessus*

Other variants

rrl	1464318	n.111C>T	non_coding_transcript_exon_variant	0.9859154929577465
rrl	1464484	n.277C>T	non_coding_transcript_exon_variant	1.0
rrl	1464540	n.333C>T	non_coding_transcript_exon_variant	1.0
rrl	1464557	n.350G>A	non_coding_transcript_exon_variant	1.0
rrl	1464841	n.634C>T	non_coding_transcript_exon_variant	1.0
rrl	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.Trp10Arg	loss_of_function_variant	0.9824561403508771
erm(41)	2346113	c.159T>C	synonymous_variant	1.0
erm(41)	2346192	p.Ile80Val	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

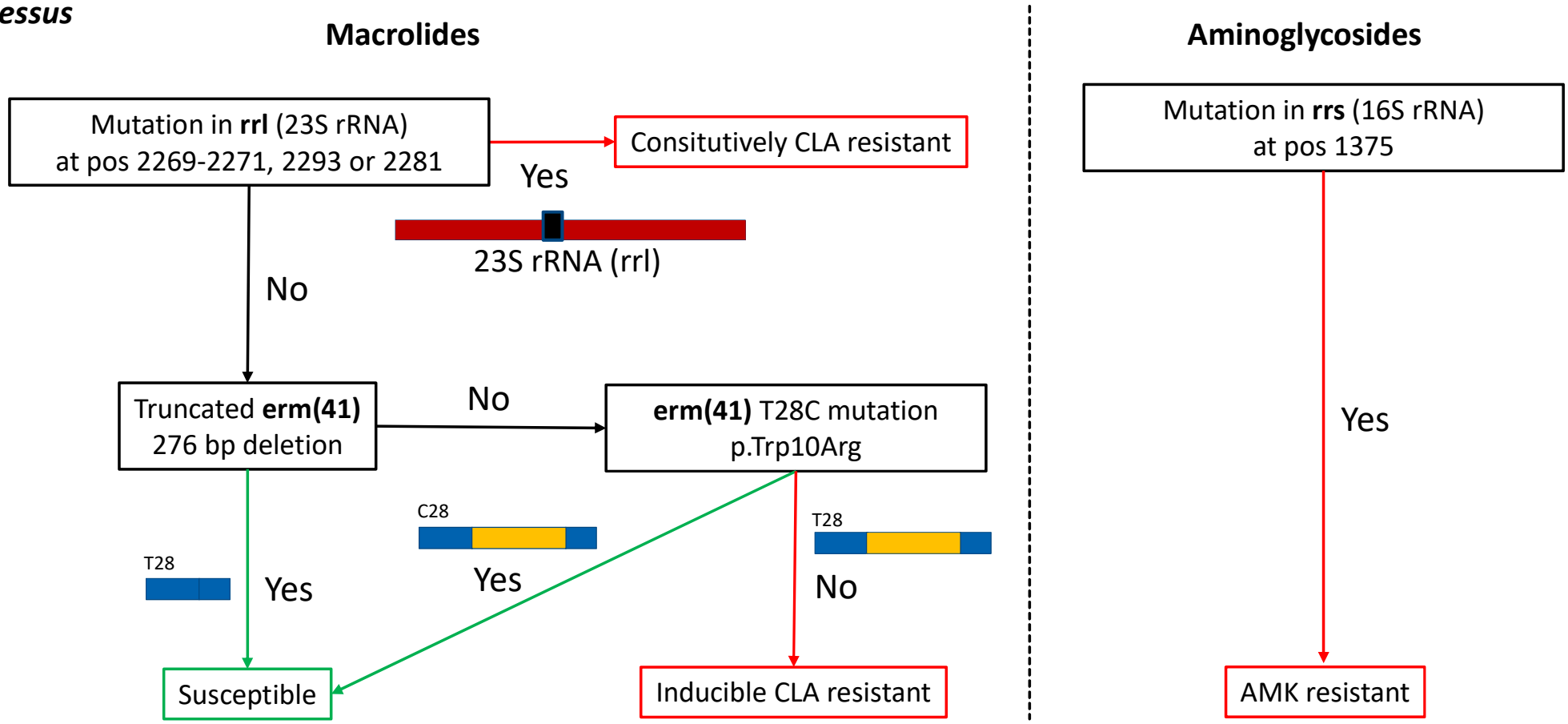
Target-specific coverage

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
rrl	100.0	64.0
erm(41)	100.0	43.0

DECISION ALGORITHM RESISTANCE PREDICTION (*M. ABSCESSUS*)

M. abscessus



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.

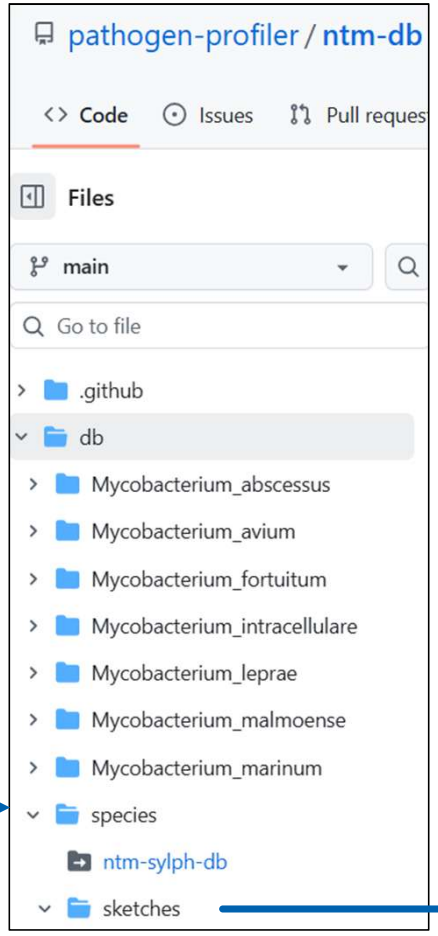
Please beware that this tools is in alpha testing and should not yet be considered for production use. If you would like to get involved and help out with testing or development please drop me a line through the issues tab.

Installation is available through conda:

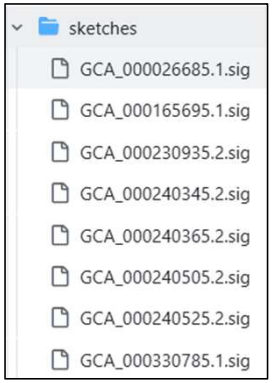
```
conda install bioconda::ntm-profiler
```

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

```
ntm-profiler update_db
```



Set of hashes = fingerprint for each genome)



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.

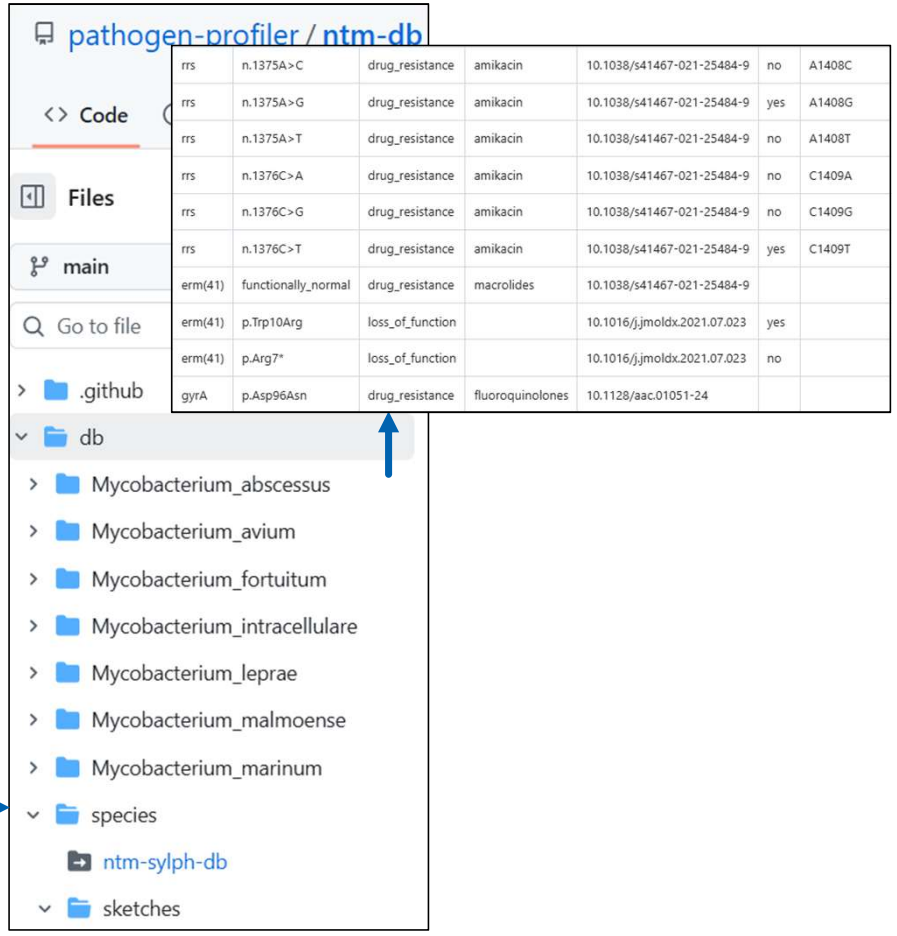
Please beware that this tools is in alpha testing and should not yet be considered for production use. If you would like to get involved and help out with testing or development please drop me a line through the issues tab.

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
```

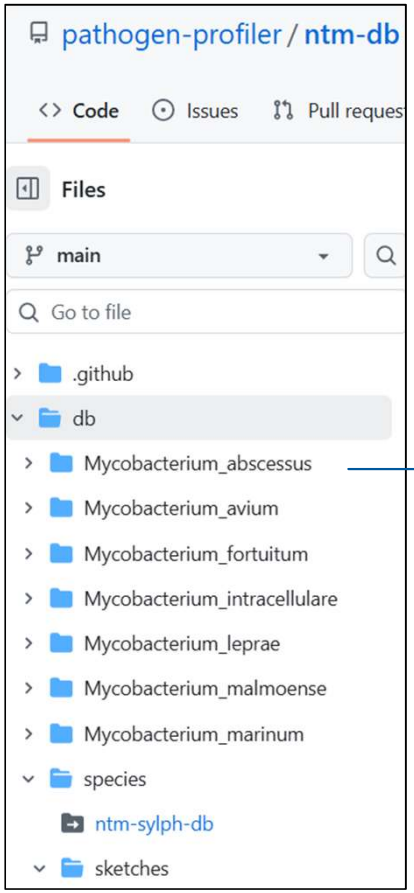


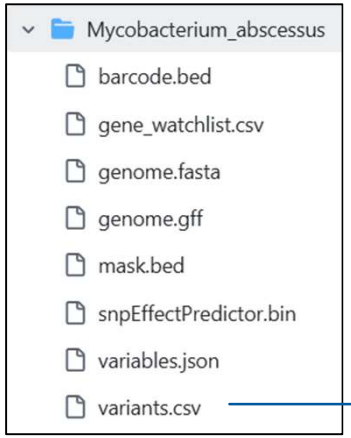
rrs	n.1375A>C	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	A1408C
rrs	n.1375A>G	drug_resistance	amikacin	10.1038/s41467-021-25484-9	yes	A1408G
rrs	n.1375A>T	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	A1408T
rrs	n.1376C>A	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	C1409A
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



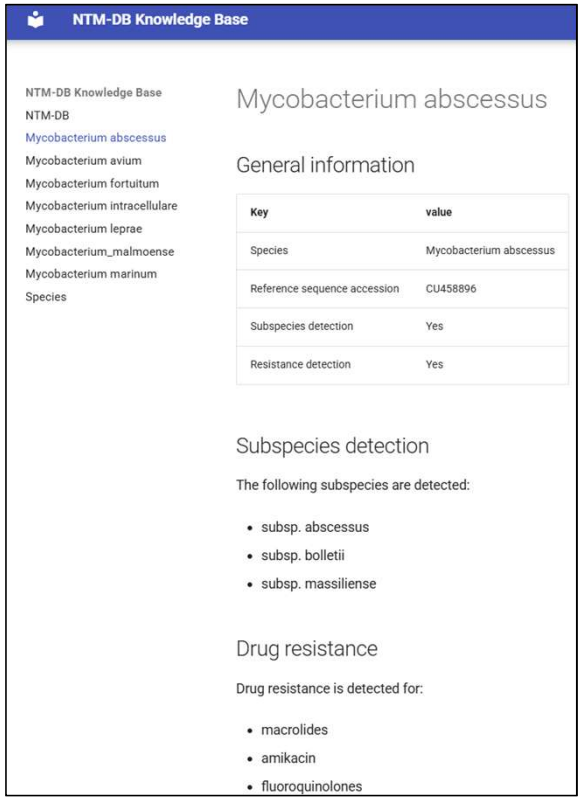


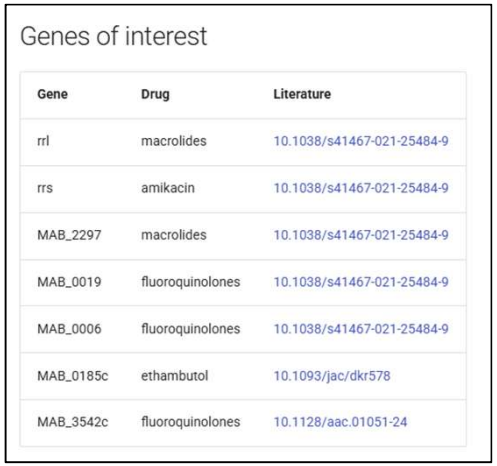
Gene	Mutation	type	drug	literature	hain	E.coli-nomenclature
rrl	n.2269A>C	drug_resistance	macrolides	10.1016/j.jmoldx.2021.07.023	no	A2057C
rrl	n.2269A>G	drug_resistance	macrolides	10.1016/j.jmoldx.2021.07.023	no	A2057G
rrl	n.2269A>T	drug_resistance	macrolides	10.1016/j.jmoldx.2021.07.023	no	A2057T
rrl	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

NTM PROFILER

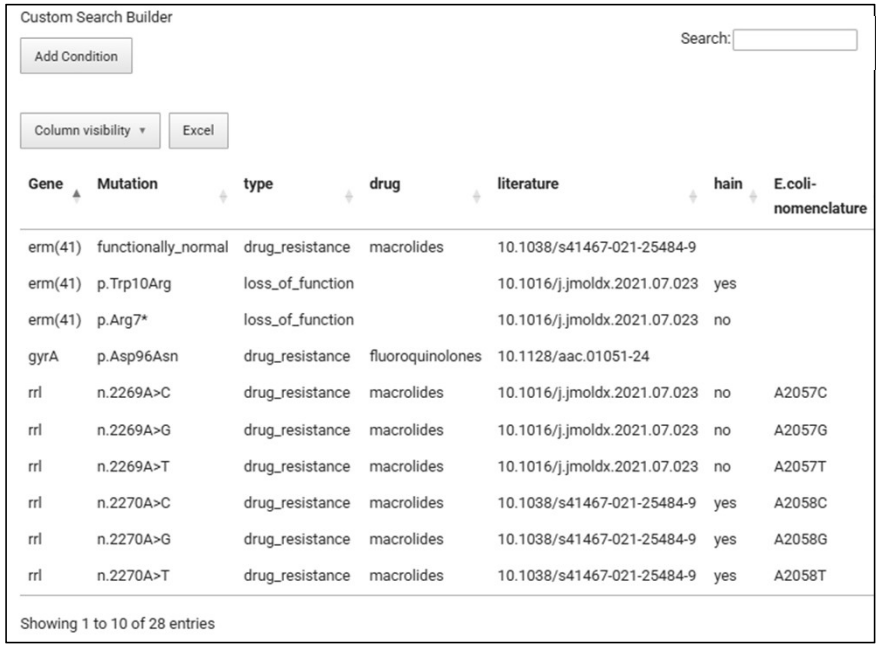
- **https://pathogen-profiler.github.io/ntm-db**





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

Add Condition

Search:

Column visibility ▾ Excel

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

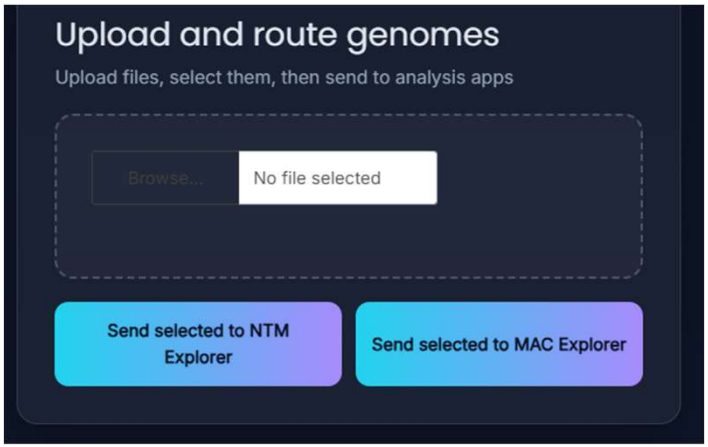
MYCODENTIFIER

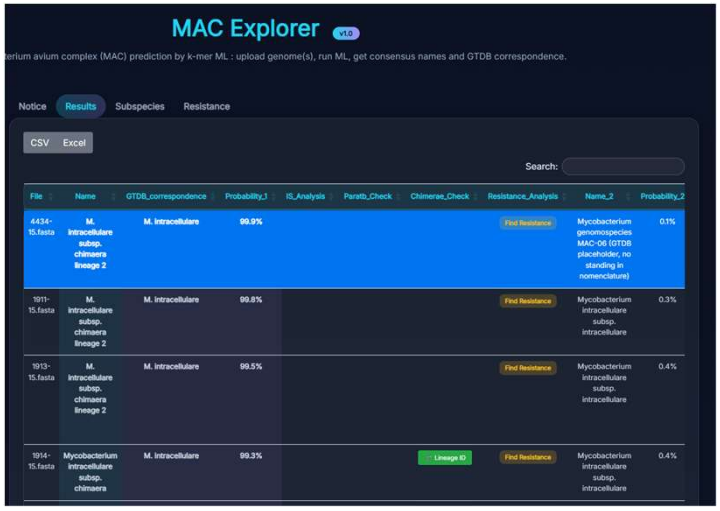
- **Access**:
  - CLI: https://github.com/JordyCoolen/MyCodentifier
- **Author:** Jordy Coolen and Heleen Severin (Radboud University, Netherlands)
- **Input:** FastQ
- **Output**: Species, (drug resistance: currently being implemented)
- **Principle for NTM ID:** Sequences matched against a custom *hsp65* 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



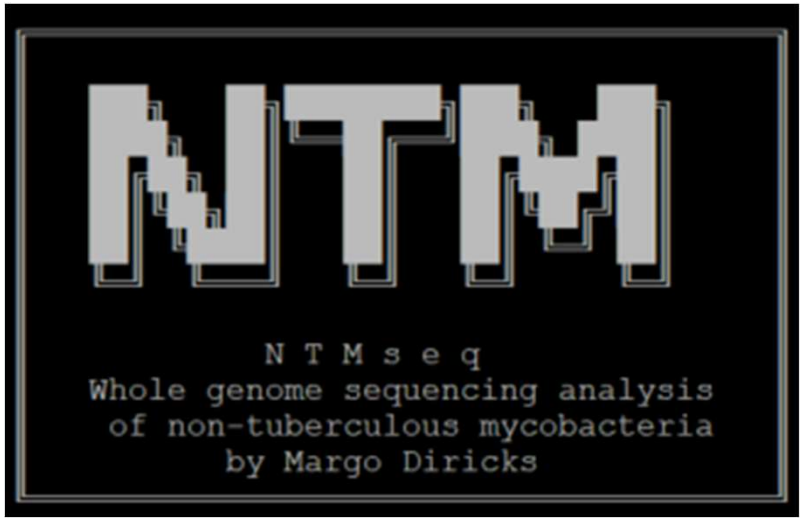


Show 15 entries Search:

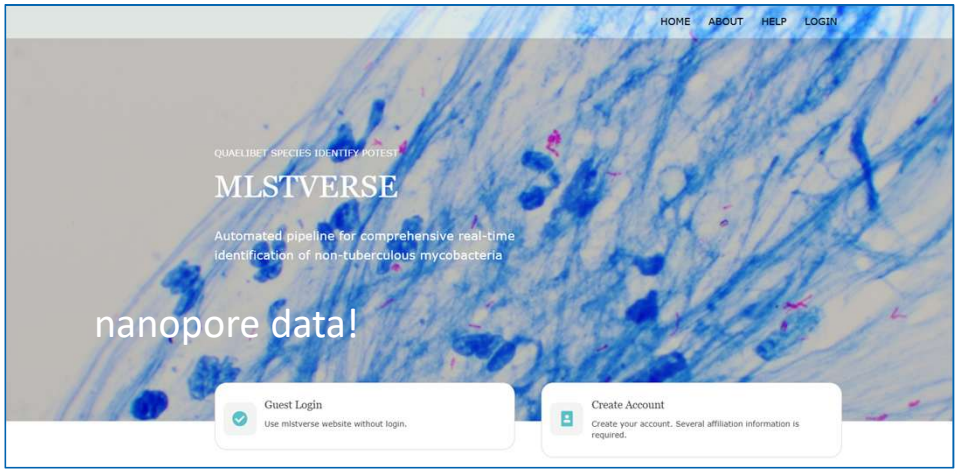
	genome	gene	gene_found	identity	coverage	mutation	status	antibiotic	impact
1	4434-15.fasta	rrs	YES	90.9	100	A1408G	ABSENCE	Amikacine	-
2	4434-15.fasta	rrs	YES	90.9	100	C1409T	ABSENCE	Amikacine	-
3	4434-15.fasta	rrs	YES	90.9	100	G1491C	PRESENCE	Amikacine	Résistance possible
4	4434-15.fasta	rrs	YES	90.9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

NTMSEQ

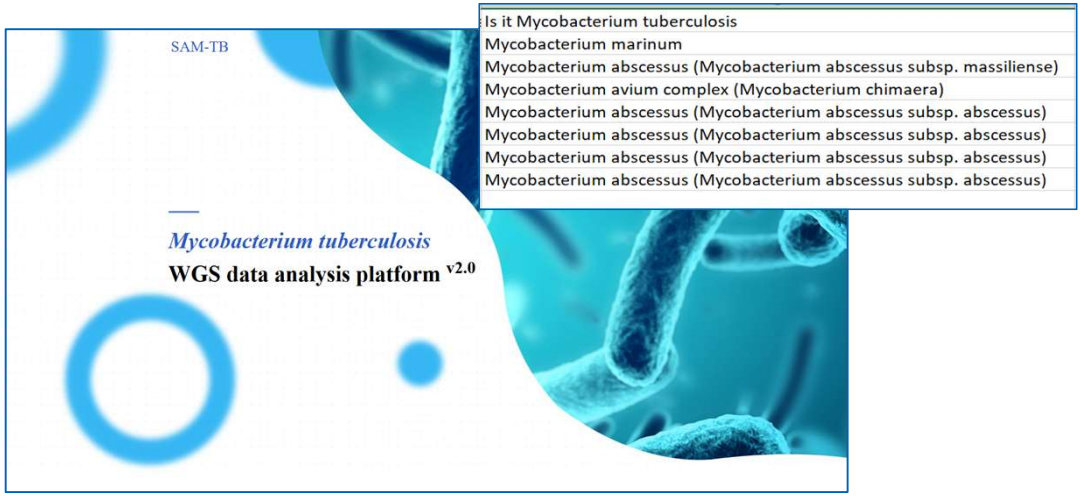
- **Access**:
  - CLI NTMseq: https://github.com/ngs-fzb/NTMseq
- **Author:** Margo Diricks (Research Center Borstel)
- **Input:** FastQ
- **Output**: quality control, (sub)species ID, MLST sequence types, assemblies, drug resistance, plasmid prediction, (approximate) phylogenies
- **Principle for NTM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



OTHER

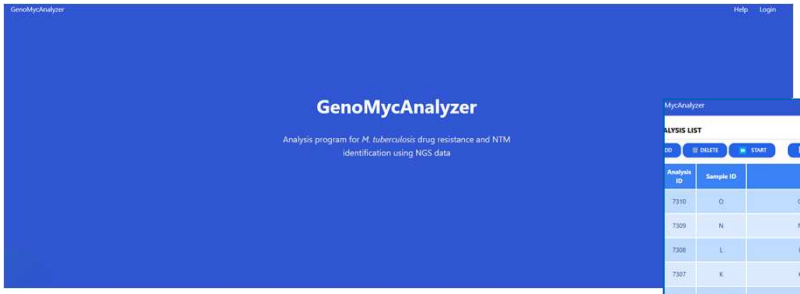


Japan (Matsumoto, 2019) - 184 genes - nanopore data

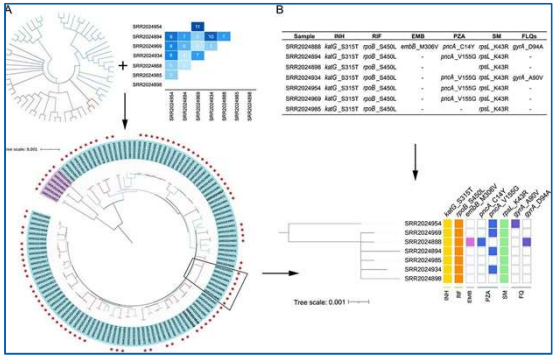


China (Yang, 2022) - uses mlstverse, kraken2 – only Illumina

Only TB: Drug-resistance prediction for 17 drugs, phylogeny

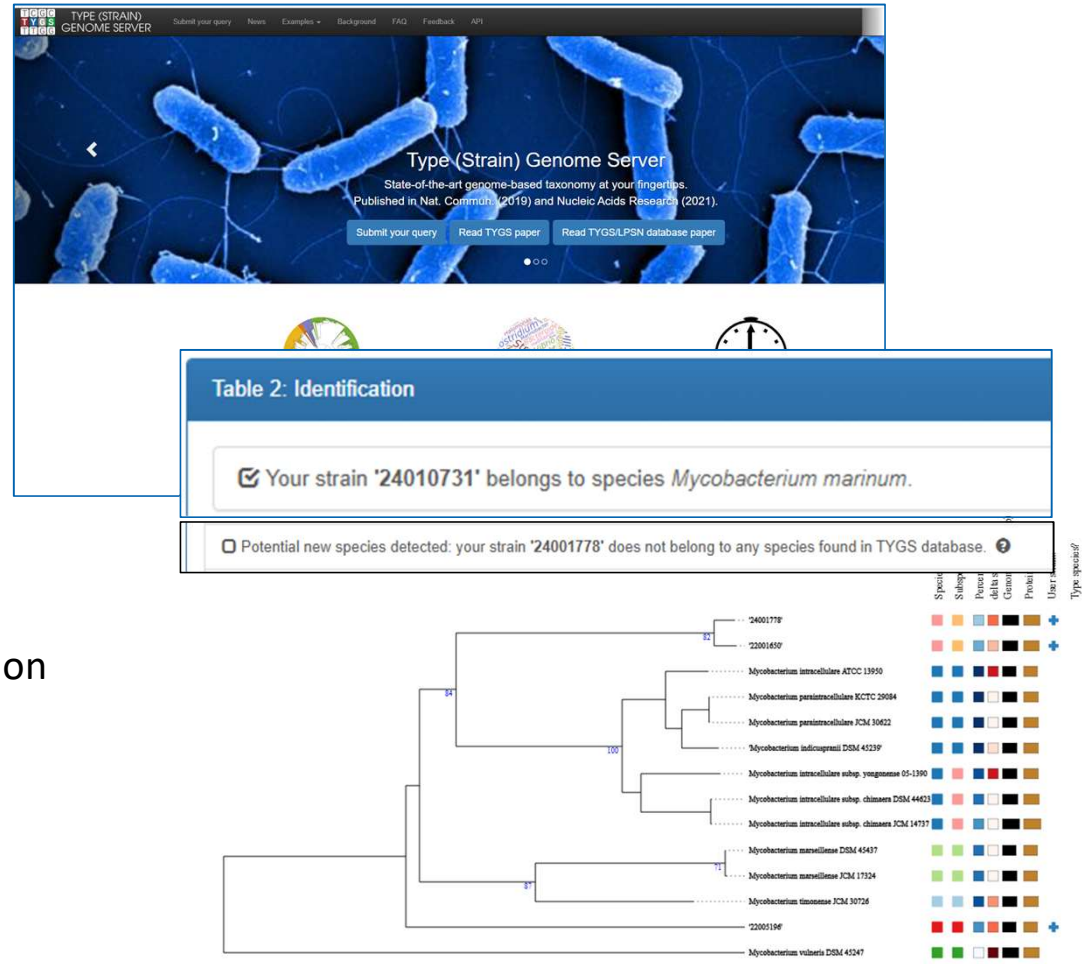


South Korea (Kim, 2024) – kraken2) – only Illumina

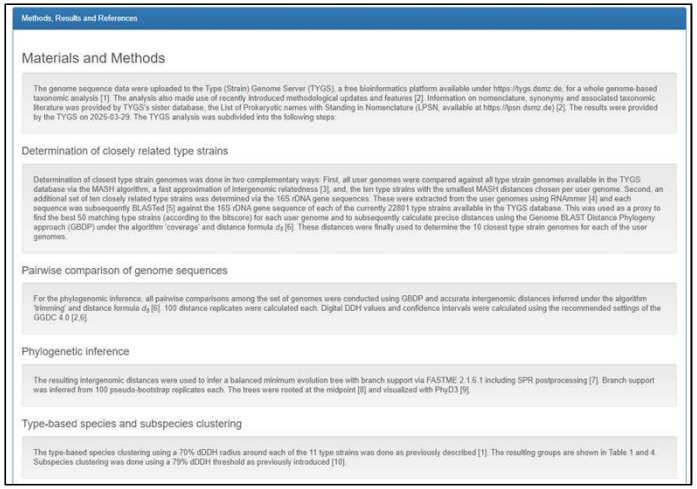


TYGS

- **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



TYGS


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) [2]. 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 22861 type strains available in the TYGS database. This was used as a proxy to find the best 50 matching type strains (according to the bitscore) 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 phylogenomic 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 PhyD3 [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].

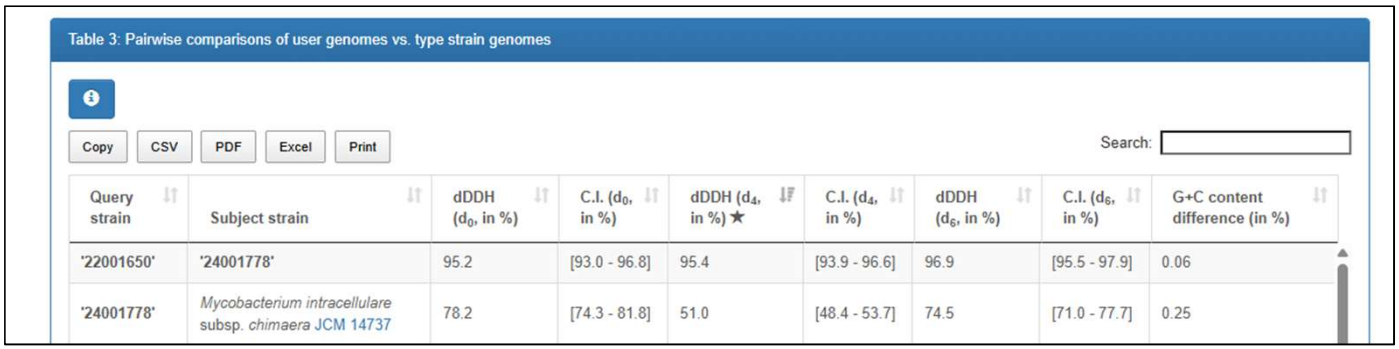


Table 3: Pairwise comparisons of user genomes vs. type strain genomes

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_6, in %)	C.I. (d_6, 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'	*Mycobacterium intracellulare* subsp. *chimaera* JCM 14737	78.2	[74.3 - 81.8]	51.0	[48.4 - 53.7]	74.5	[71.0 - 77.7]	0.25

dDDH d4 ≈ Σ(identity in HSPs) / Σ(HSP lengths) + transformation
→ Can also be used for draft genomes

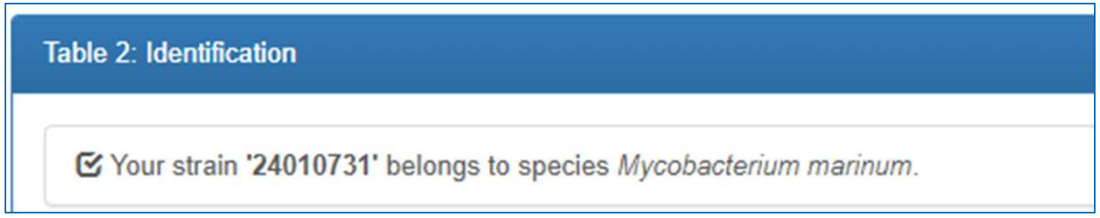


Table 2: Identification

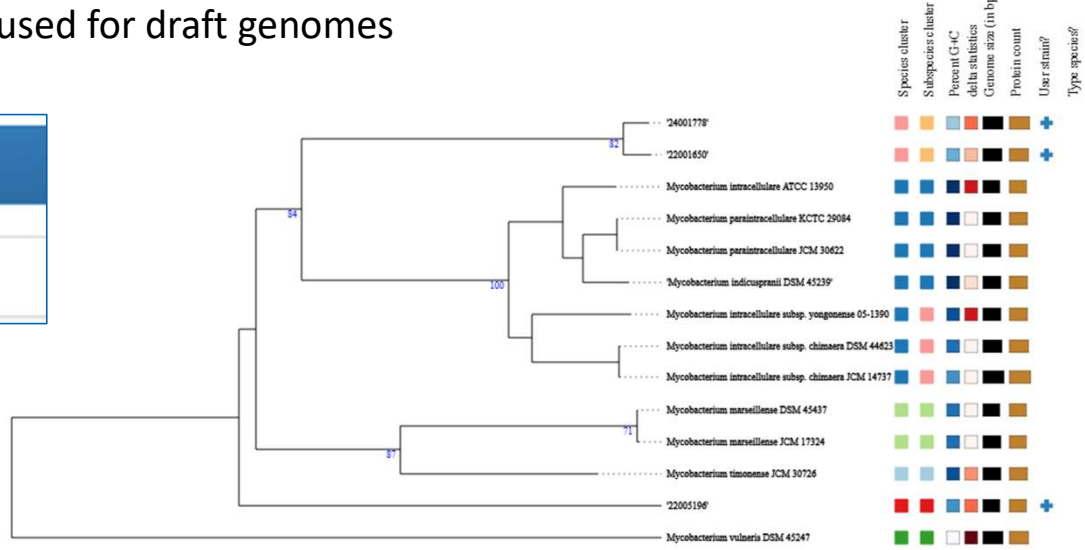
Your strain '24010731' belongs to species *Mycobacterium marinum*.


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

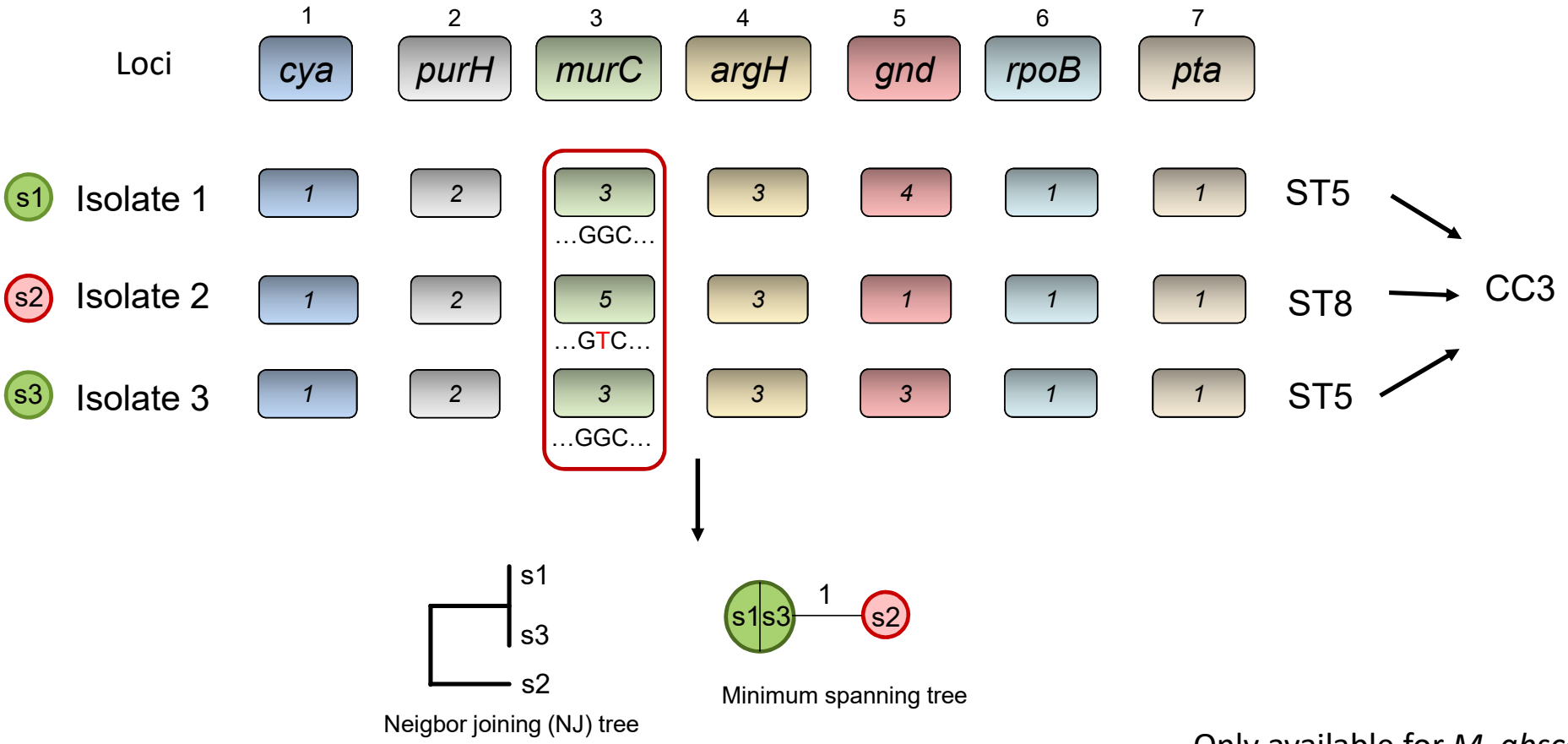
COMPARATIVE ANALYSIS

- 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 got 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

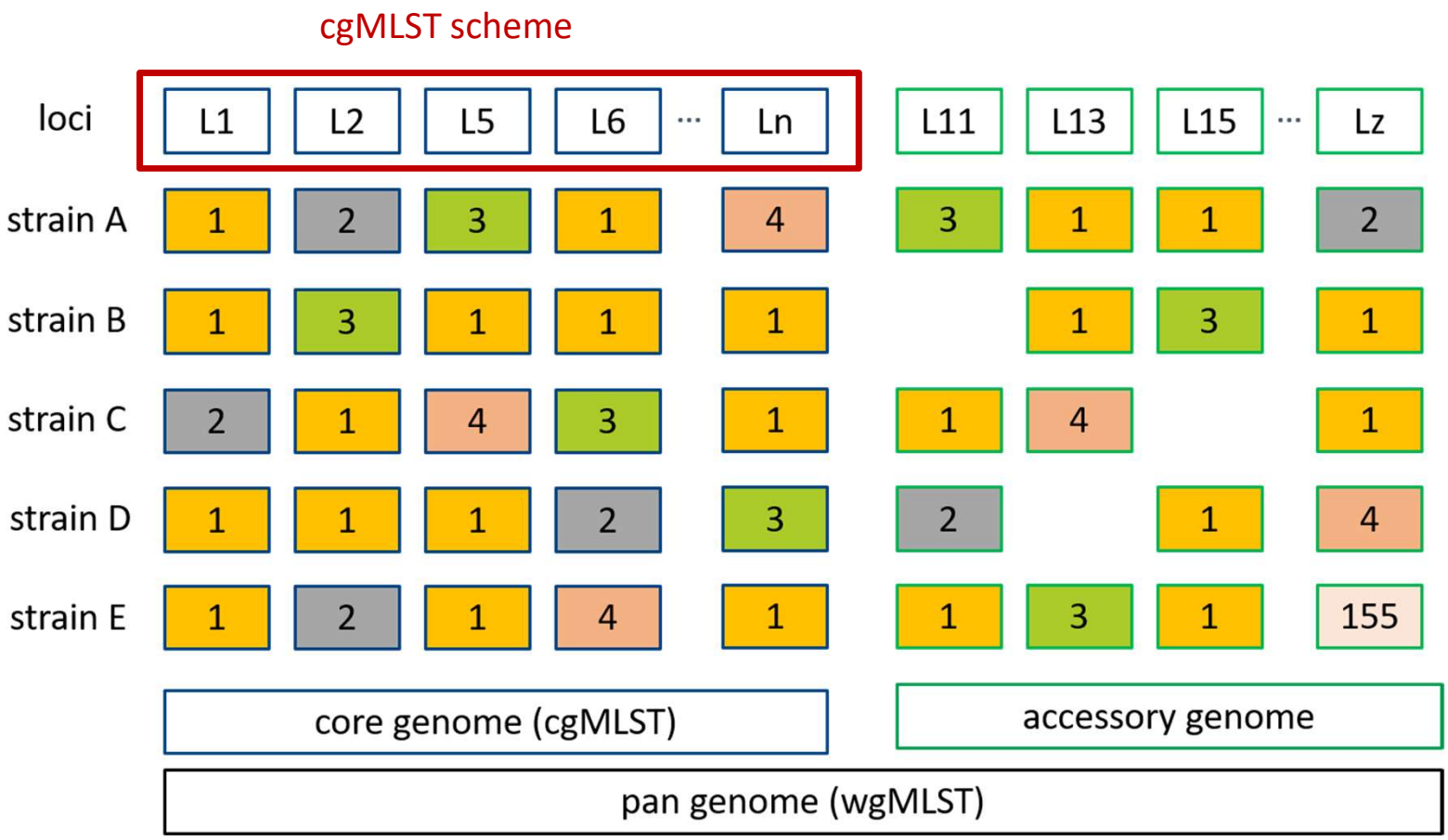
- Multi-locus sequence typing (MLST) – e.g. using pubMLST, Ridom Typer, SRST2, mlst,…
- core genome multi-locus sequence typing (cgMLST) – e.g. using pubMLST, Ridom Typer, chewbbaca,…
- core genome single nucleotide polymorphism analysis (cgSNP) – e.g. using snippy, MTBseq,…

MULTI-LOCUS SEQUENCE TYPING



Only available for *M. abscessus*

CORE GENOME MULTI-LOCUS SEQUENCE TYPING



CORE GENOME MULTI-LOCUS SEQUENCE TYPING

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,*]

cgMLST schemes:

M. tuberculosis
2,891 genes (72%)

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

M. abscessus
2,904 genes (59%)

CORE GENOME MULTI-LOCUS SEQUENCE TYPING

- 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

(CORE GENOME) MULTI-LOCUS SEQUENCE TYPING

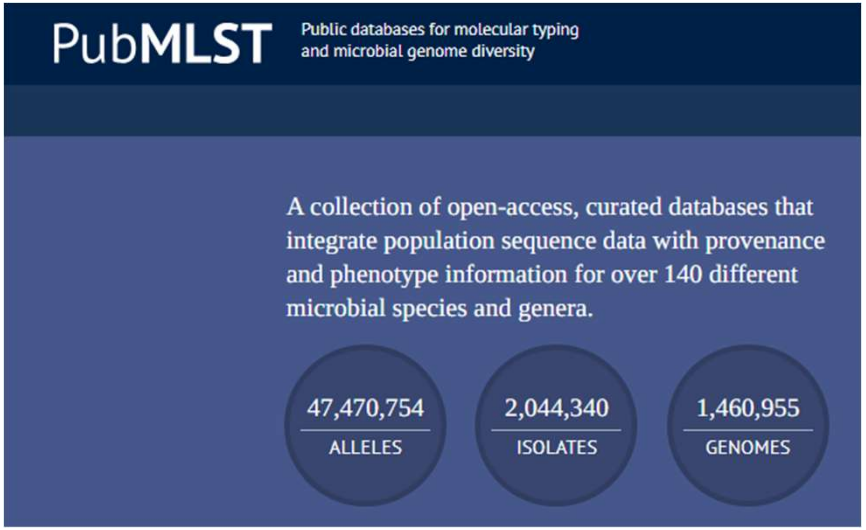
Research Center Borstel
Leibniz Lung Center

- 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

PUBMLST

Research Center Borstel
Leibniz Lung Center

- **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*

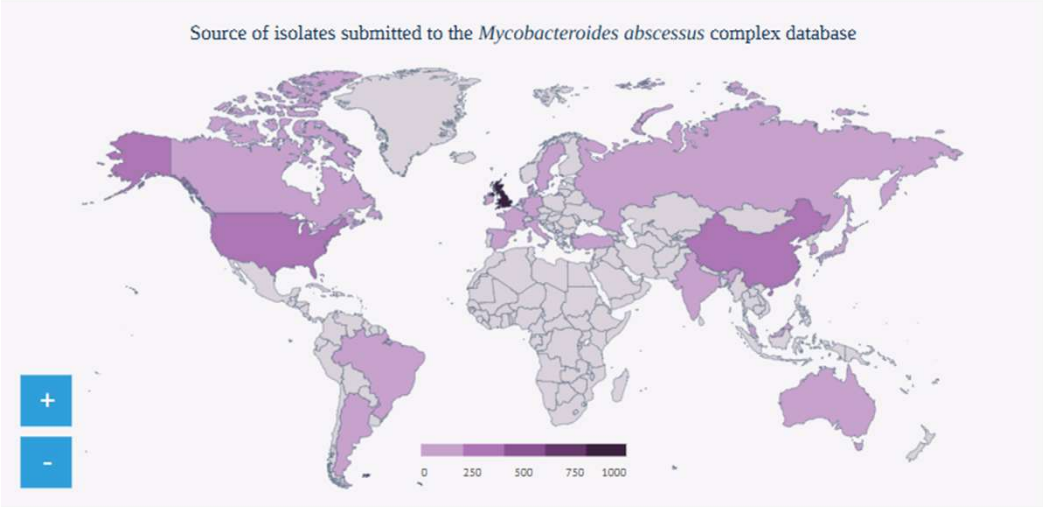


PUBMLST

Research Center Borstel
Leibniz Lung Center

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.

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].

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

PUBMLST

Research Center Borstel
Leibniz Lung Center

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

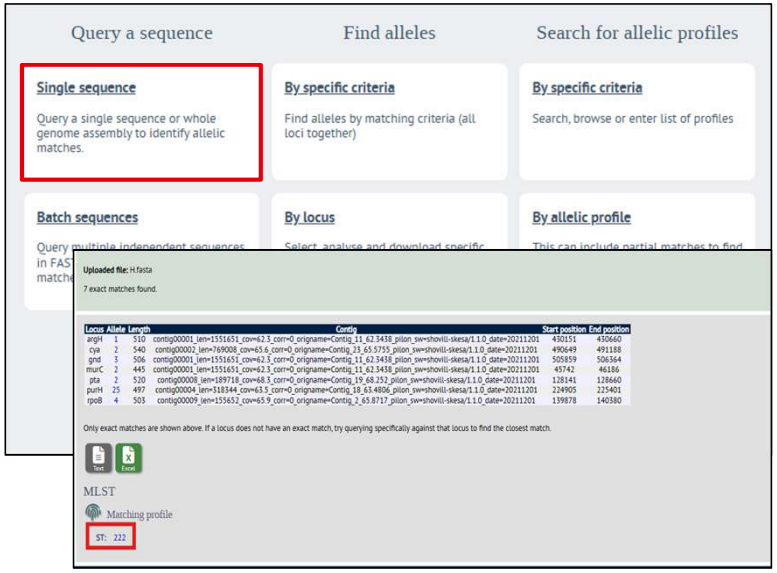
Genome collection

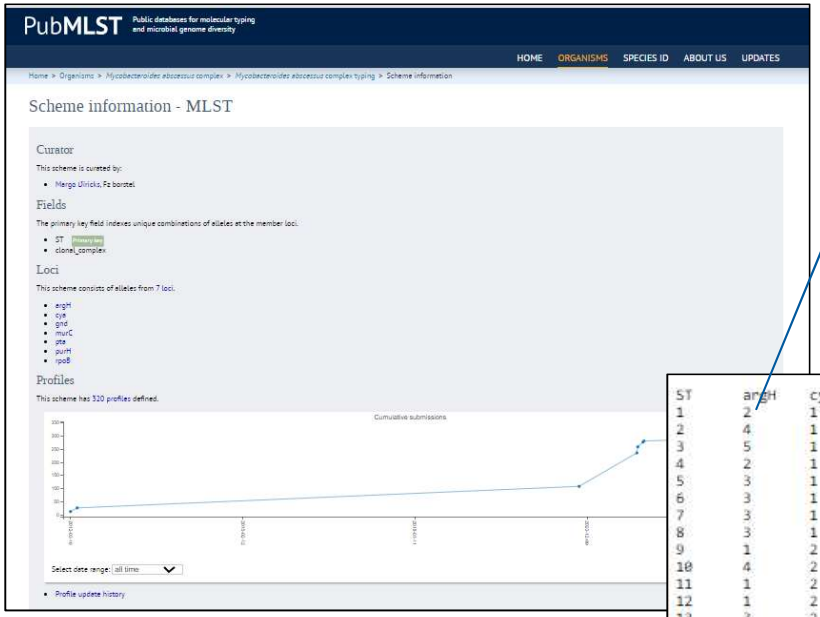
A subset of records within the isolate database may contain genomes assemblies. You can access these from the isolate database by filtering on sequence bin size in a query.

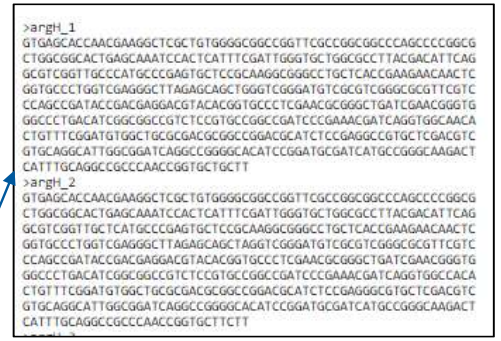
Genomes: 2,079

Last updated: 2025-05-27

Nomenclature:







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	

PUBMLST

Research Center Borstel
Leibniz Lung Center

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

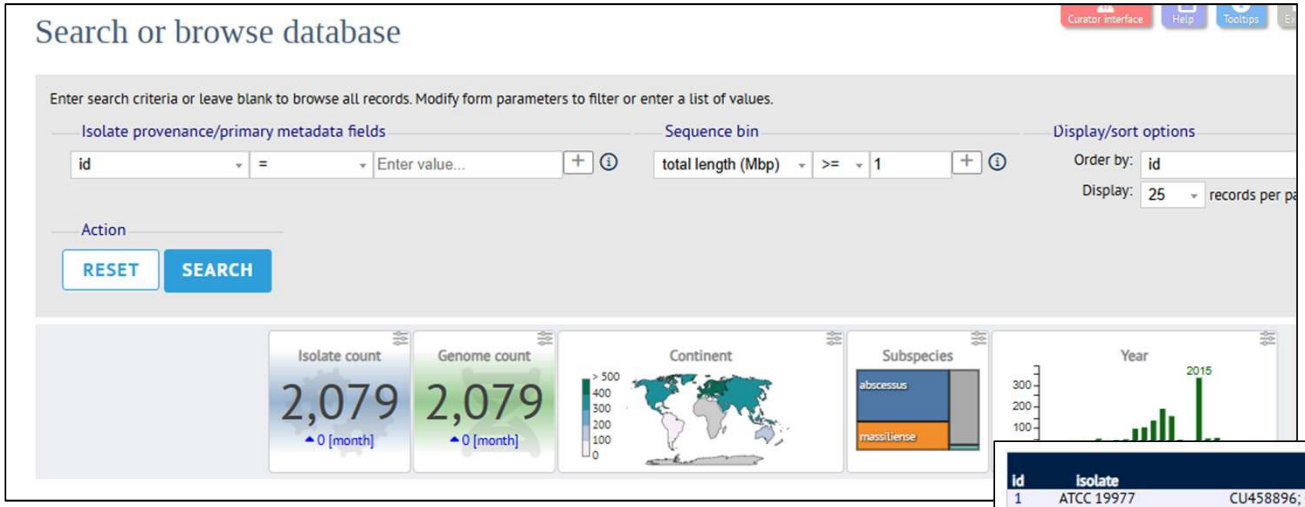
Last updated: 2026-02-28

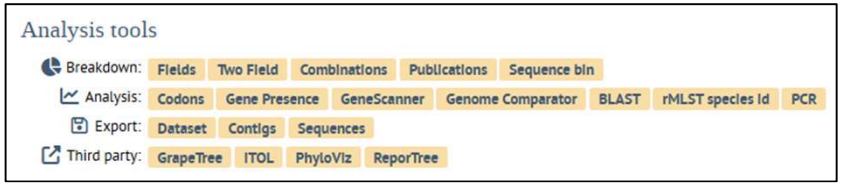
Genome collection

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

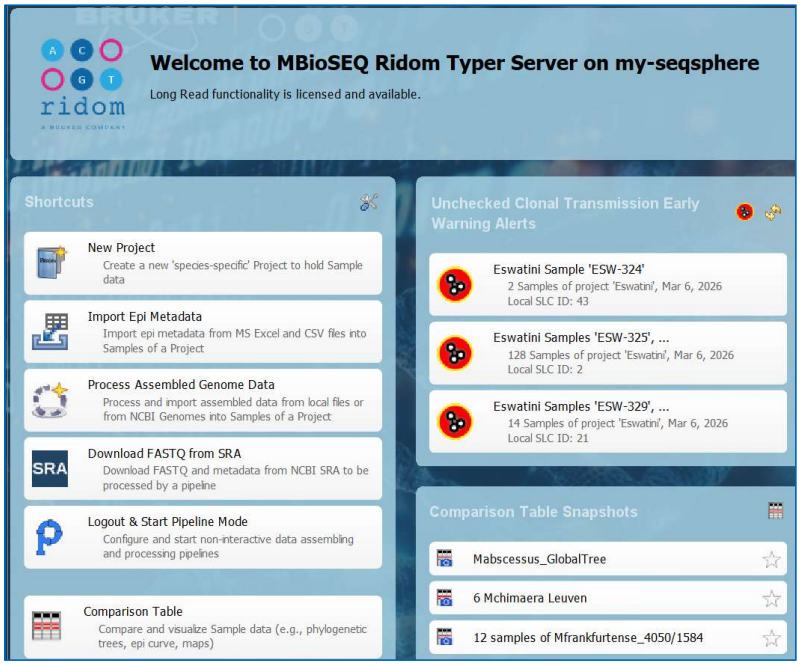




	Isolate fields						MLST								
id	isolate	aliases	country	subspecies	year	source	argH	cya	gnd	murC	pta	purH	rpoB	ST	clonal complex
1	ATCC 19977	CU458896; GCA_000069185.1; NC_010397	Unknown	abscessus		unknown	3	1	4	3	1	2	1	5	1
2	ERR114969		Unknown	massiliense		unknown	30	31	17	27	21	31	18	33	3
3	ERR114985		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

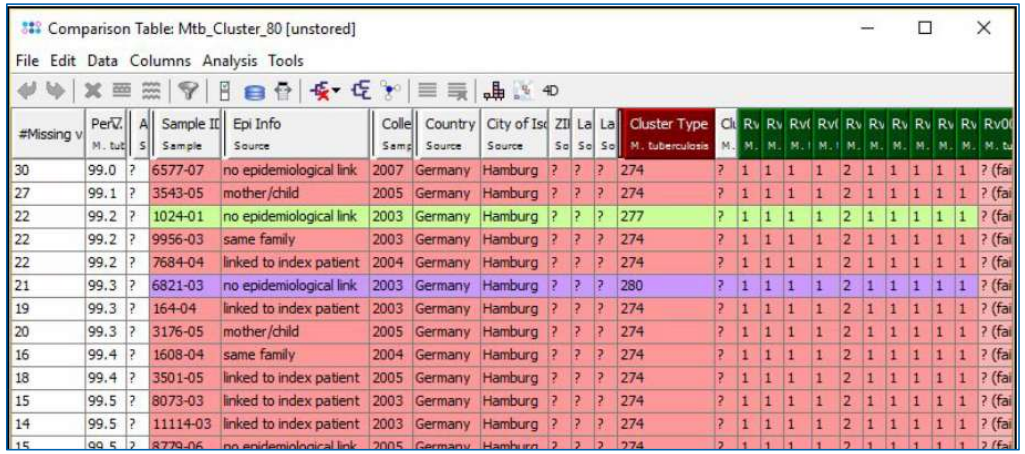
BRUKER MBIOSEQ RIDOM TYPER (SEQSPHERE+)

- **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
  - Continous processing with pipeline
  - Publicly available scheme only available for M. abscessus, but creation of your own (ad hoc) cgMLST schemes possible

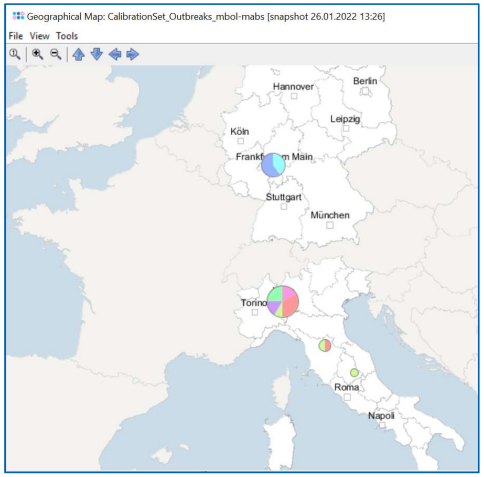


BRUKER MBIOSEQ RIDOM TYPER (SEQSPHERE+)

Research Center Borstel
Leibniz Lung Center



#Missing v	Per√.	A	Sample ID	Epi Info	Colle	Country	City of Isc	Zil	La	La	Cluster Type	Cl	Rv	Rv	Rv(	Rv(	Rv	Rv	Rv	Rv	Rv	Rv	Rv0
	M. tut	S	Sample	Source	Samp	Source	Source	So	So	So	M. tuberculosis	M.	M.	M.	M.	M.	M.	M.	M.	M.	M.	M.	M.	M. tu
30	99.0	?	6577-07	no epidemiological link	2007	Germany	Hamburg	?	?	?	274	?	1	1	1	1	2	1	1	1	1	1	? (fai
27	99.1	?	3543-05	mother/child	2005	Germany	Hamburg	?	?	?	274	?	1	1	1	1	2	1	1	1	1	1	? (fai
22	99.2	?	1024-01	no epidemiological link	2003	Germany	Hamburg	?	?	?	277	?	1	1	1	1	2	1	1	1	1	1	? (fai
22	99.2	?	9956-03	same family	2003	Germany	Hamburg	?	?	?	274	?	1	1	1	1	2	1	1	1	1	1	? (fai
22	99.2	?	7684-04	linked to index patient	2004	Germany	Hamburg	?	?	?	274	?	1	1	1	1	2	1	1	1	1	1	? (fai
21	99.3	?	6821-03	no epidemiological link	2003	Germany	Hamburg	?	?	?	280	?	1	1	1	1	2	1	1	1	1	1	? (fai
19	99.3	?	164-04	linked to index patient	2003	Germany	Hamburg	?	?	?	274	?	1	1	1	1	2	1	1	1	1	1	? (fai
20	99.3	?	3176-05	mother/child	2005	Germany	Hamburg	?	?	?	274	?	1	1	1	1	2	1	1	1	1	1	? (fai
16	99.4	?	1608-04	same family	2004	Germany	Hamburg	?	?	?	274	?	1	1	1	1	2	1	1	1	1	1	? (fai
18	99.4	?	3501-05	linked to index patient	2005	Germany	Hamburg	?	?	?	274	?	1	1	1	1	2	1	1	1	1	1	? (fai
15	99.5	?	8073-03	linked to index patient	2003	Germany	Hamburg	?	?	?	274	?	1	1	1	1	2	1	1	1	1	1	? (fai
14	99.5	?	11114-03	linked to index patient	2003	Germany	Hamburg	?	?	?	274	?	1	1	1	1	2	1	1	1	1	1	? (fai
15	99.5	?	8779-06	no epidemiological link	2005	Germany	Hamburg	?	?	?	274	?	1	1	1	1	2	1	1	1	1	1	? (fai

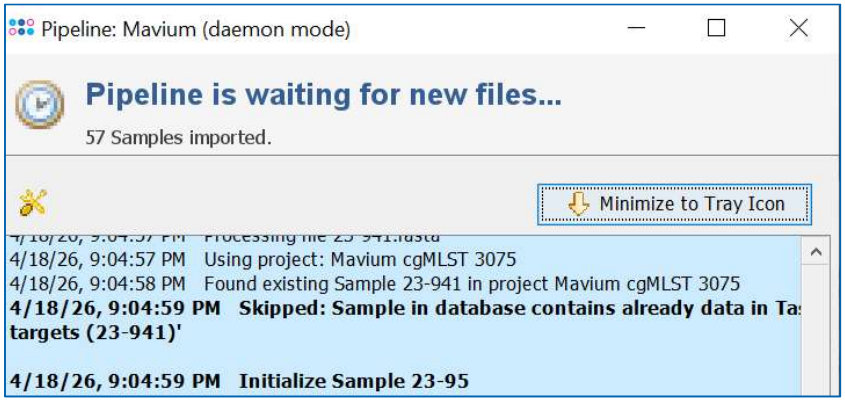


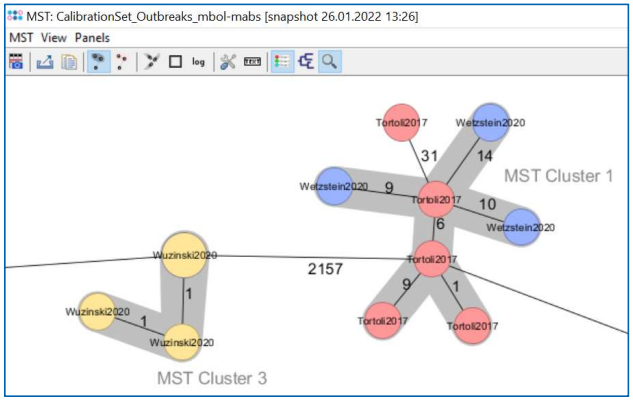
Nomenclature:
https://cgmlst.org/ncs

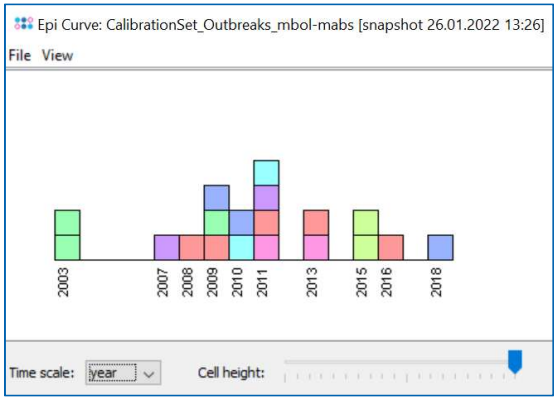
Mycobacteroides abscessus cgMLST

Show Targets

Name	M. abscessus cgMLST
Version	1.1
Seed Genome	(NC_010397.1, 02-AUG-2016)
Genus	Mycobacteroides
Species	abscessus
Locus Count	2,904
Complex Type Distance	24
Complex Type Count	1,058
Curators	M. Diricks, Research Center Borstel, Germany

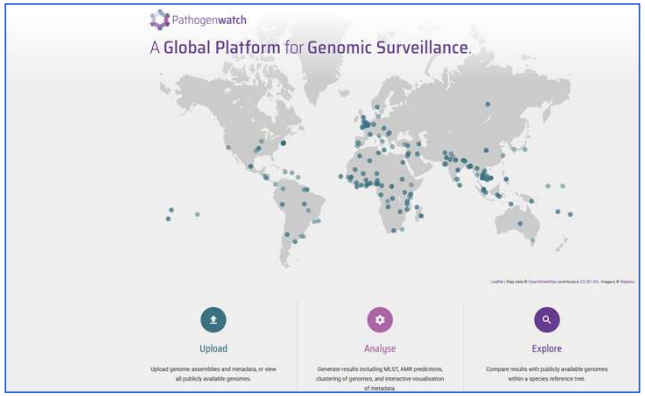


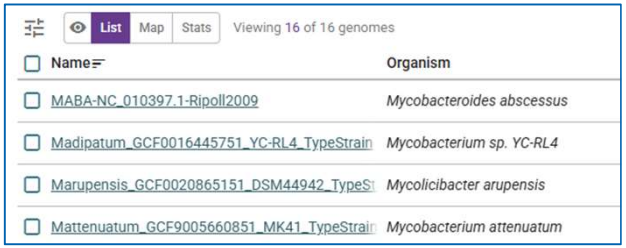


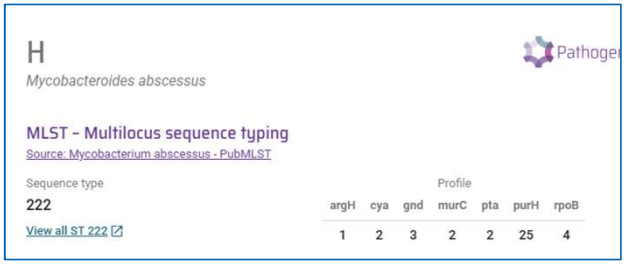


PATHOGENWATCH

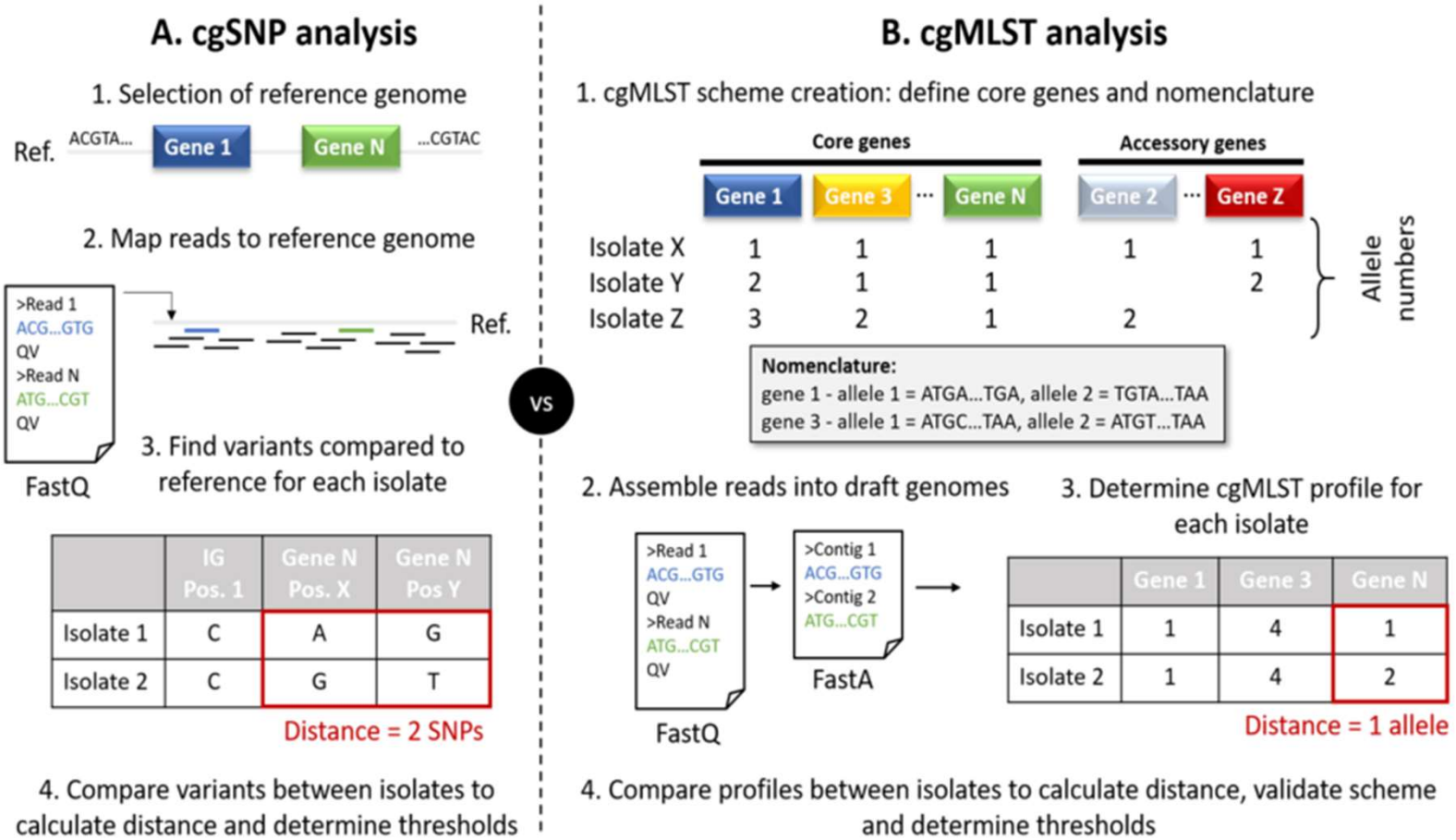
- **Access**
  - **Web-based**: https://pathogen.watch/
- **Author:** Wellcome Sanger institute (UK)
- **Input:** FastQ and FastA
- **Output**: Species ID, (cg)MLST ST
- **Principle for NTM ID:** Search against NCBI RefSeq/curated library with mash
- **Comments**
  - Global Platform for Genomic Surveillance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





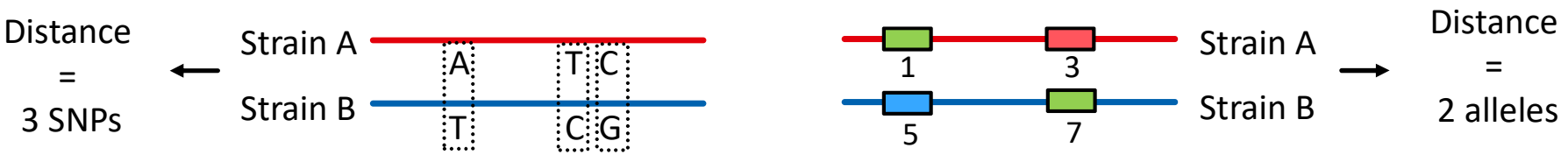


COMPARATIVE GENOMICS



SURVEILLANCE AND OUTBREAK ANALYSIS

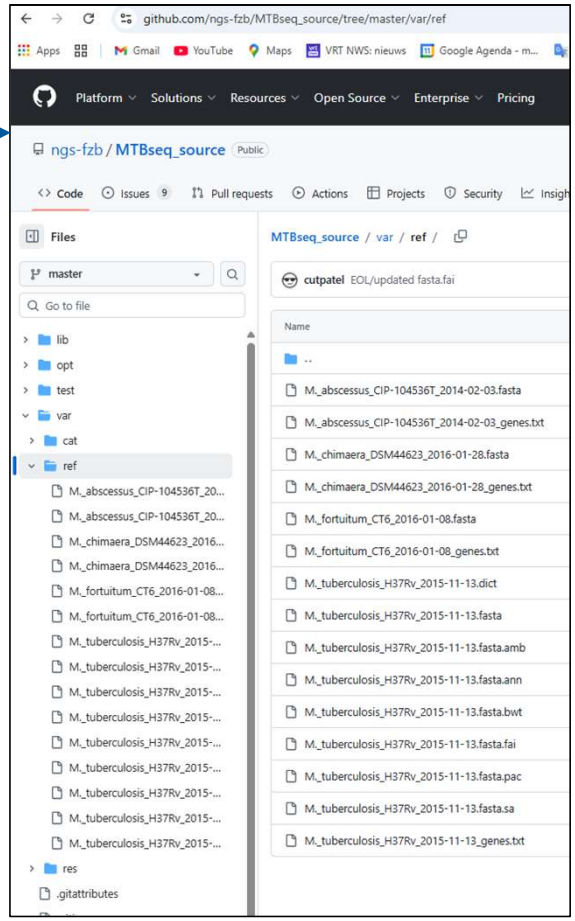
Single nucleotide polymorphisms **cgSNP** vs **cgMLST** Multi-locus sequence typing

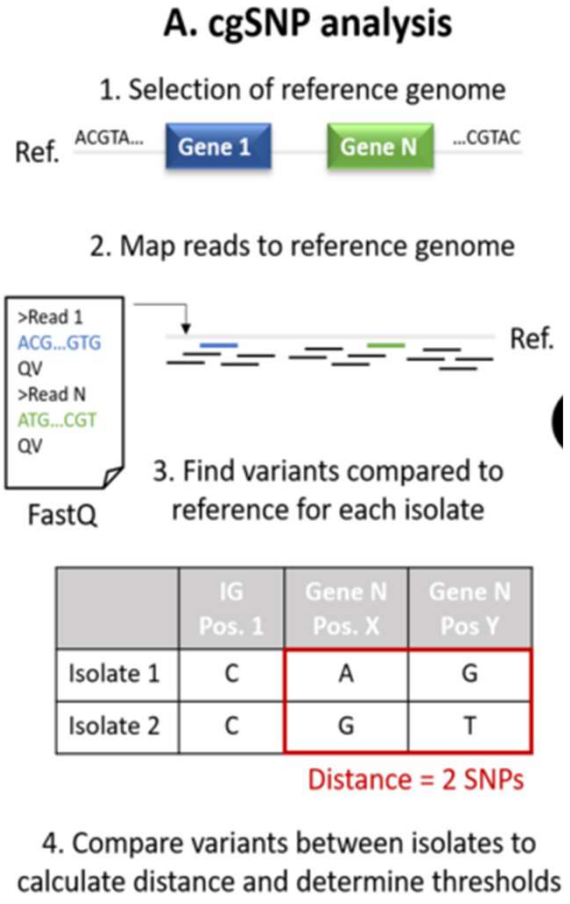


	cgSNP	cgMLST
Level	Single nucleotide substitutions	Genes
Computation power	high	low
Standardized	No	Yes
Resolution	Highest	High
Well suited for	In-depth comparison Retrospective analysis Cluster resolving	Comparison of diverse strains Prospective surveillance Cluster detection

CGSNP ANALYSIS

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





CONCLUSIONS

- 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