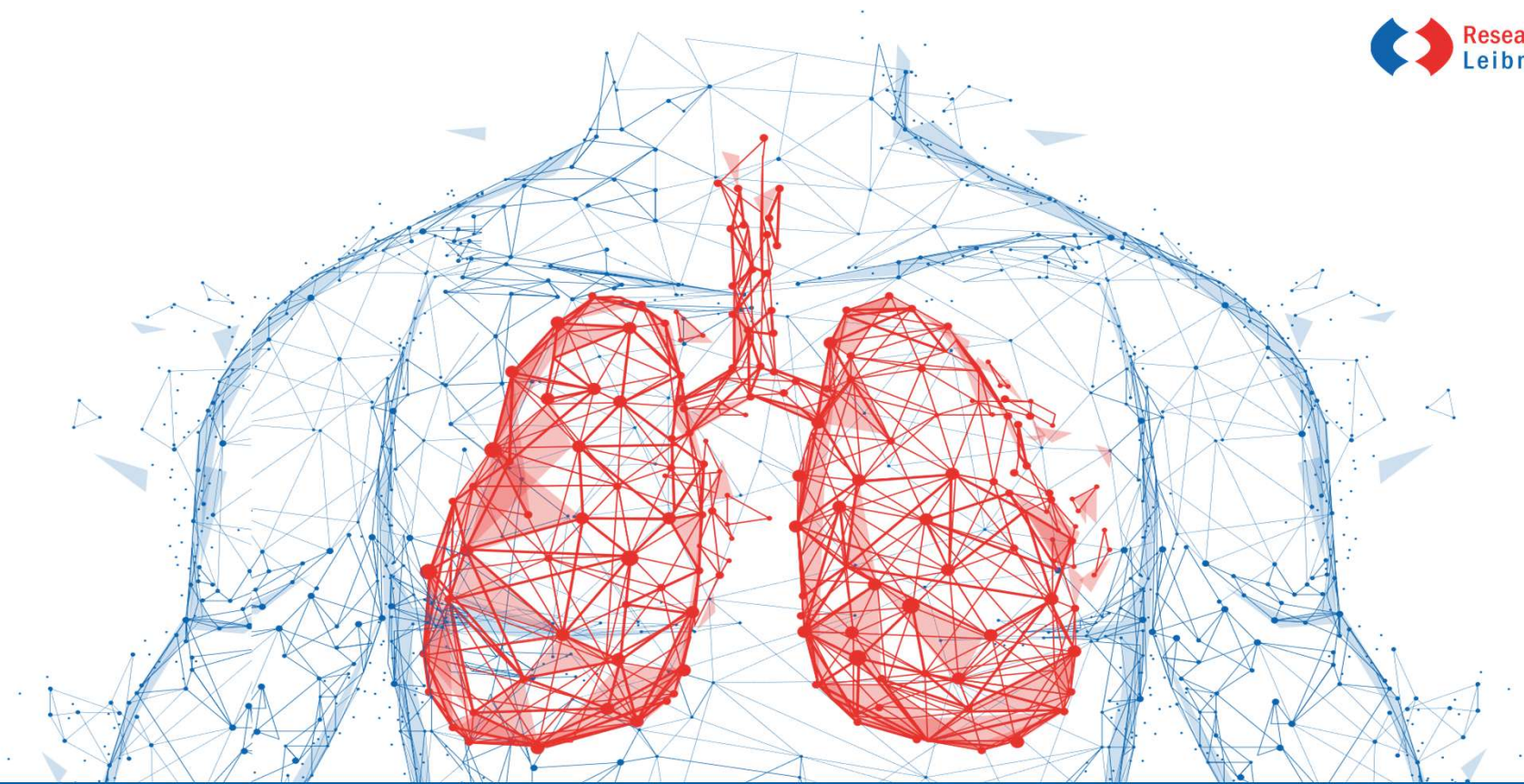
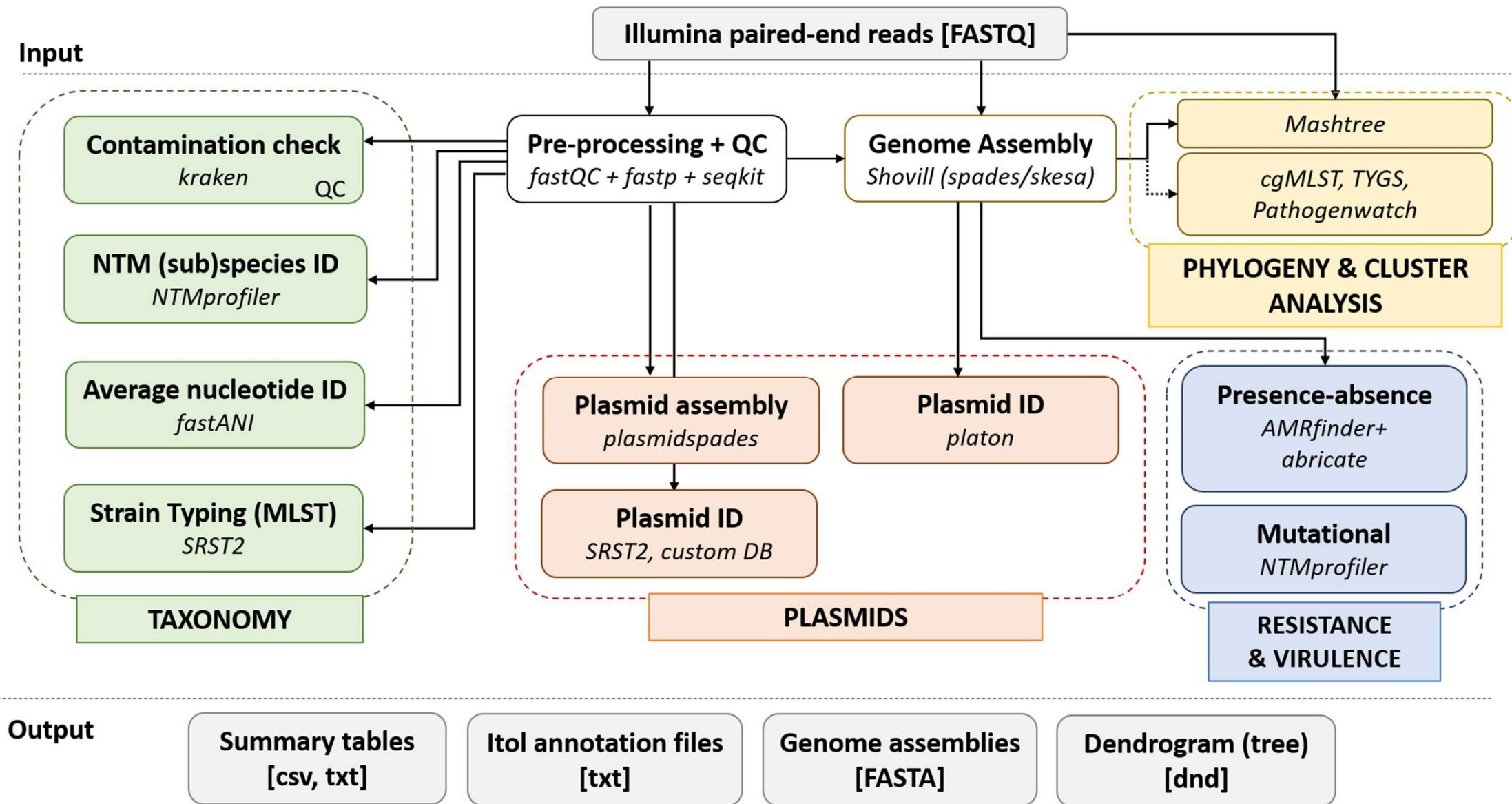




NTMseq

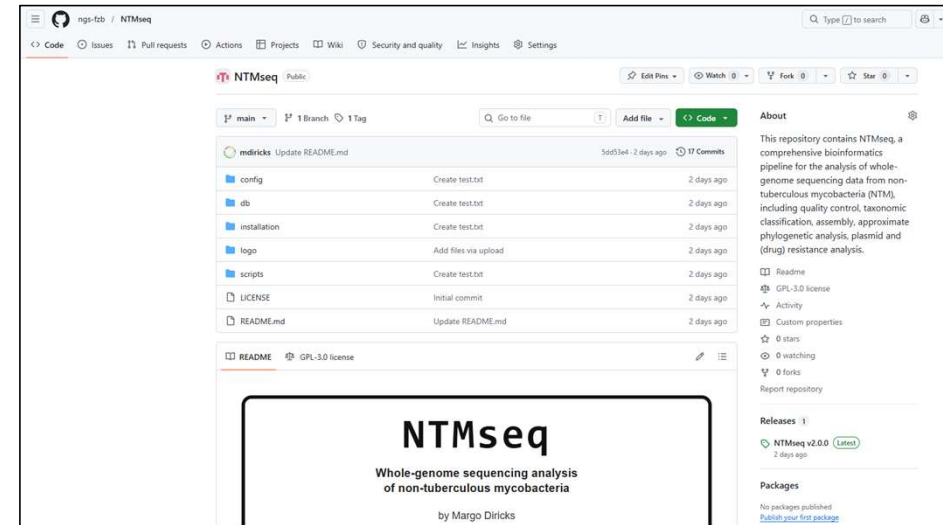
NTMseq



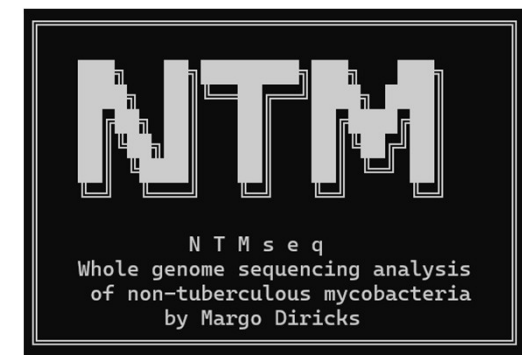
INSTALLATION AND RUNNING NTMSEQ



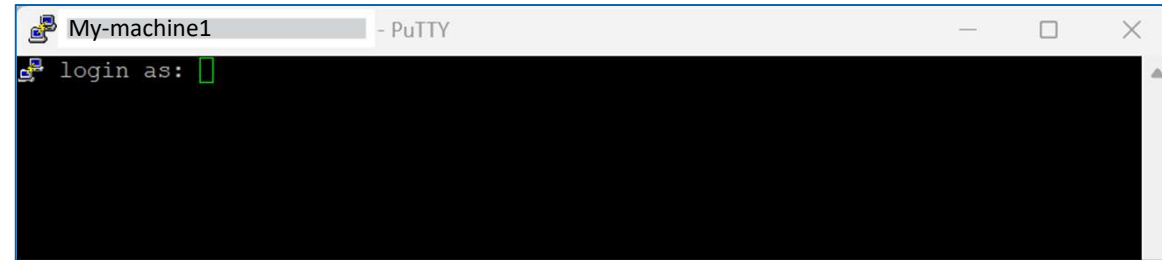
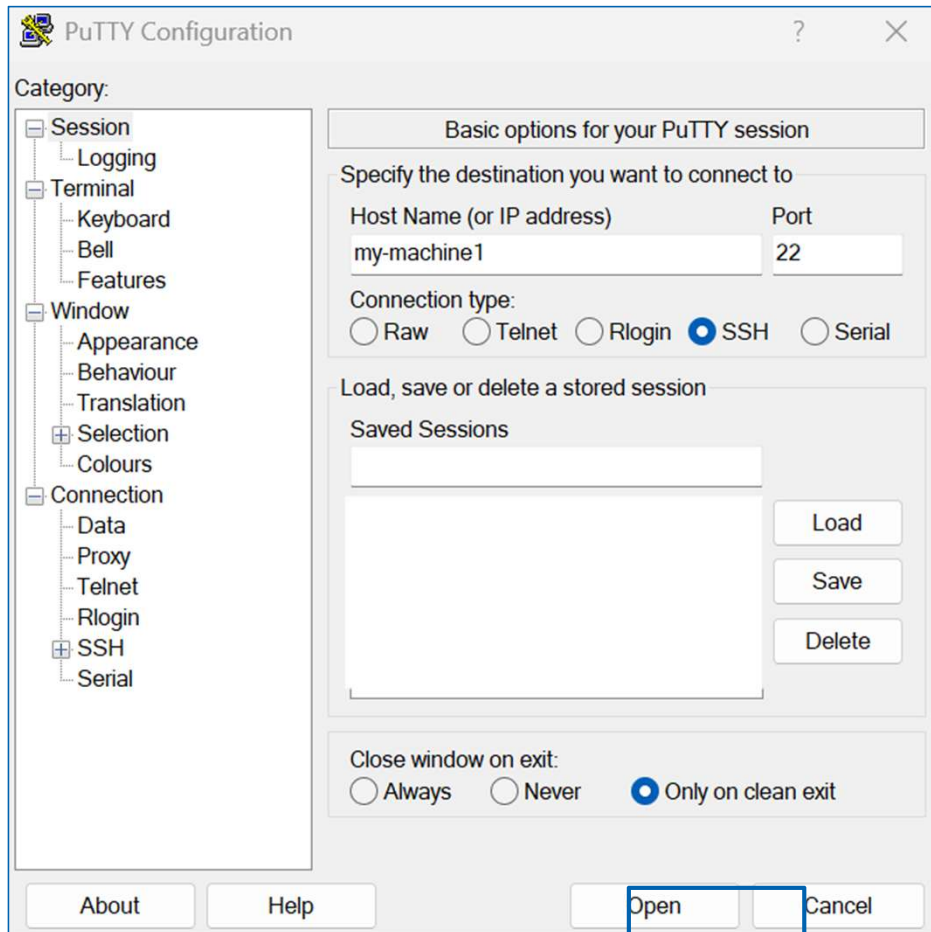
- <https://github.com/ngs-fzb/NTMseq>
- Optimized for short-read sequencing data (Illumina)
- Requirements
 - Linux-based system (e.g. VM, wsl,...)
 - Basic command-line experience
 - Conda (e.g. miniconda) or mamba for faster installation
- Steps
 - Connect to a linux-based system (e.g. to a VM via putty)
 - Download repository (wget)
 - Install required tools: `$ bash Installation_NTMseq.sh`
 - Install databases: instructions at github
 - Configure your analyses: in NTMseq.config
 - Run analyses: `$ bash starter_NTMseq.sh ../config/NTMseq.config`



NTMseq is for Research Use Only!



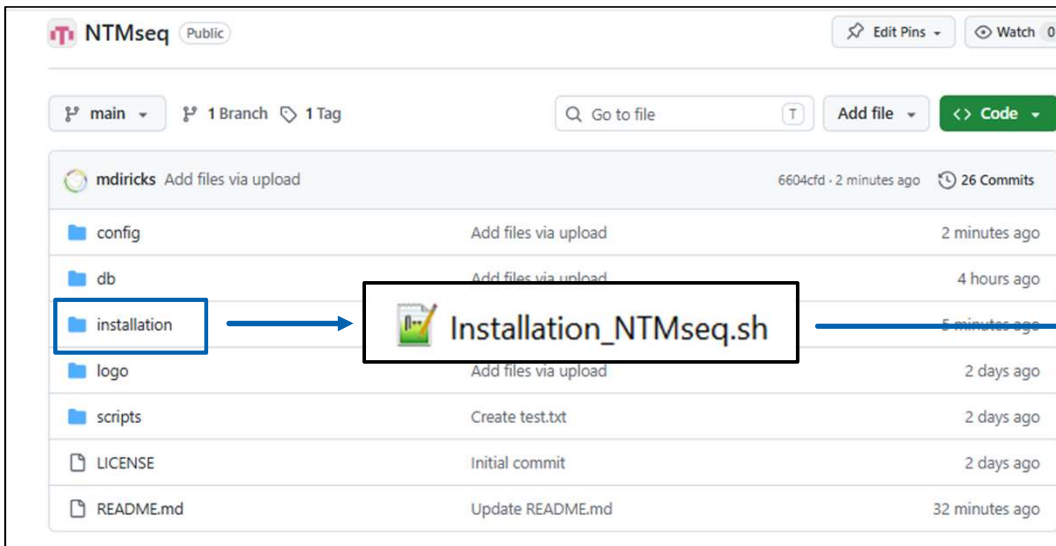
CONNECTION TO A LINUX BASED SYSTEM



16 threads, 47 Gb RAM

```
(base) mdiricks@my-hostname:~$ mkdir NTMseq
(base) mdiricks@my-hostname:~$ cd NTMseq
(base) mdiricks@my-hostname:~/NTMseq$ wget https://github.com/ngs-fzb/NTMseq/archive/refs/heads/main.tar.gz && tar -xzf main.tar.gz
(base) mdiricks@my-hostname:~/NTMseq$ cd ./NTMseq-main/installation
(base) mdiricks@my-hostname:~/NTMseq/NTMseq-main/installation$ bash Installation_NTMseq.sh
```


INSTALLATION



```
install_env "NTMseq_multiqc" \
  bioconda::fastqc=0.12.1 \
  bioconda::multiqc=1.34

install_env "NTMseq_seqkit" \
  bioconda::seqkit=2.13.0 \
  conda-forge::pigz=2.8

install_env "NTMseq_fastp" \
  bioconda::fastp=1.3.2

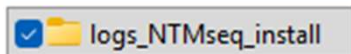
install_env "NTMseq_fastani" \
  bioconda::fastani=1.34

install_env "NTMseq_kraken2" \
  bioconda::kraken2=2.17.1 \
  bioconda::krona=2.8.1
```

```
if command -v mamba >/dev/null 2>&1; then
  CONDA_CREATE_CMD="mamba"
  info "Using mamba for environment creation."
else
  CONDA_CREATE_CMD="conda"
  warn "mamba was not found. Falling back to conda."
fi
```

If mamba is installed, it will be used; otherwise, conda will be used.

After installation, check



☐ Name	Änderungsdatum	Typ	Größe
checks_20260424_115842.log	24.04.2026 12:40	Textdokument	5 KB
installation_20260424_115842.log	24.04.2026 12:40	Textdokument	167 KB

NTMseq2 post-installation checks
Started: 2026-04-24 12:39:09
Generated by: Installation_NTMseq.sh (v1.1.0)

```
=====
CHECK: multiqc version
TIME : 2026-04-24 12:39:09
CMD : conda run -n NTMseq_multiqc multiqc --version
-----
multiqc, version 1.34
```

STATUS: OK

```
=====
CHECK: seqkit version
TIME : 2026-04-24 12:39:22
CMD : conda run -n NTMseq_seqkit seqkit version
-----
seqkit v2.13.0
```

STATUS: OK

INSTALLATION OF NTMSEQ



```

  NTMseq
  Whole genome sequencing analysis
  of non-tuberculous mycobacteria
  by Margo Diricks

[2026-05-03 19:29:33] INFO: Using mamba for environment creation.

-----
[2026-05-03 19:29:33] NTMseq2 installation started
-----

Script      : Installation_NTMseq.sh
Version     : 1.1.0
Started     : 2026-05-03 19:29:33
Log dir     : /home/mdiricks/NTMseq/NTMseq-main/installation/logs_NTMseq_install
Install log : /home/mdiricks/NTMseq/NTMseq-main/installation/logs_NTMseq_install/installation_20260503_192933.log
Check log   : /home/mdiricks/NTMseq/NTMseq-main/installation/logs_NTMseq_install/checks_20260503_192933.log

[2026-05-03 19:29:33] INFO: This script installs the required tools for NTMseq2 using mamba.
[2026-05-03 19:29:33] INFO: Each tool is installed in a separate environment to reduce dependency clashes.
[2026-05-03 19:29:33] WARNING: Pinned versions are used for reproducibility and may not be the newest available releases.

-----
[2026-05-03 19:29:33] Installing NTMseq_multiqc

```

- Problems to install some of the tools → have a look at the tools original website
- Ask ChatGPT or Claude
 - Often specifying specific compatible versions of tools and dependencies and/or specifying order of channels can fix the problem
 - Example: `conda create -n NTMseq_NTMprofiler bioconda::ntm-profiler=0.8.1 -y` was not working on my WSL
 - `conda config --add channels defaults \`
`conda config --add channels bioconda \`
`conda config --add channels conda-forge \`
`conda config --set channel_priority strict \`
`conda create -n NTMseq_NTMprofiler -c conda-forge -c bioconda -c defaults ntm-profiler=0.8.1`
`python=3.12 htlib=1.21 -y`

INSTALL DATABASES



- <https://github.com/ngs-fzb/NTMseq>

3. Download required databases

⚠ Note: Some databases are large (>50 GB) and must be downloaded in a location with enough storage space.

Taxonomy

1. Kraken 2: [<https://benlangmead.github.io/aws-indexes/k2>] --> choose your preferred database, depending on the space you have available

TLDR: wget https://genome-index.s3.amazonaws.com/kraken/k2_pluspf_16_GB_20260226.tar.gz && tar -xvzf k2_pluspf_16_GB_20260226.tar.gz

3. Krona: [<https://github.com/marbl/Krona/wiki/KronaTools>]

TLDR: conda run -n NTMseq_kraken2 ktUpdateTaxonomy.sh /path/to/krona-db

Typing

1. MLST via SRST2: [<https://github.com/katholt/srst2>]

This database is updated automatically when running the pipeline. However, due to a recent change in data access policy at pubMLST [<https://pubmlst.org/change-data-access-policy>], only ST types and profiles submitted up to December 2024 will be reported. For the newest ST types, please submit your assemblies directly to the pubMLST website, after login. Alternatively, download the latest database via instructions mentioned at <https://github.com/MDU-PHL/mlstdb>.

Currently, MLST typing is only available for *M. abscessus*

Resistance prediction

1. NTM-Profiler: [<https://github.com/jodyphelan/NTM-Profiler>]

TLDR: cd into folder where you want to install database; conda run -n NTMseq_NTMprofiler ntm-profiler update_db

3. AMRFinder: [<https://github.com/ncbi/amr/wiki/Upgrading#database-updates>]

TLDR: conda run -n NTMseq_NTMprofiler amrfinder_update -d </database_directory>

Plasmid prediction

1. Custom plasmid database: [<https://ccb-microbe.cs.uni-saarland.de/plsdb/plasmids/>]

You can download the full PLSDB plasmids database or a subset thereof (e.g. only those of mycobacteriaceae) via API. You will need to create multi-fasta file containing all plasmids.

Alternatively, use the pre-downloaded multi-fasta file of 208 mycobacteriaceae plasmids that can be found in this repository (db folder): 2023_11_03_v2_PLSDB_mycobacteriaceae_208plasmids.fasta

Optional:

2. Platon: [https://zenodo.org/search?page=1&size=20&q=conceptrecid:3349651&all_versions&sort=-version] or [<https://github.com/oschwengers/platon#database>]

SCRIPTS



NTMseq Public

main 1 Branch 1 Tag

Go to file T Add file <> Code

mdiricks Add files via upload 6604cfd · 2 minutes ago 26 Commits

config	Add files via upload	2 minutes ago
db	Add files via upload	4 hours ago
installation	Add files via upload	5 minutes ago
logo	Add files via upload	2 days ago
scripts	Create test.txt	2 days ago
LICENSE	Initial commit	2 days ago
README.md	Update README.md	32 minutes ago

NTMseq / scripts /

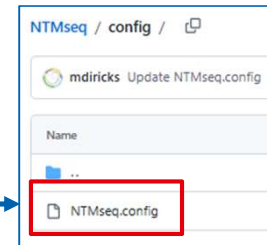
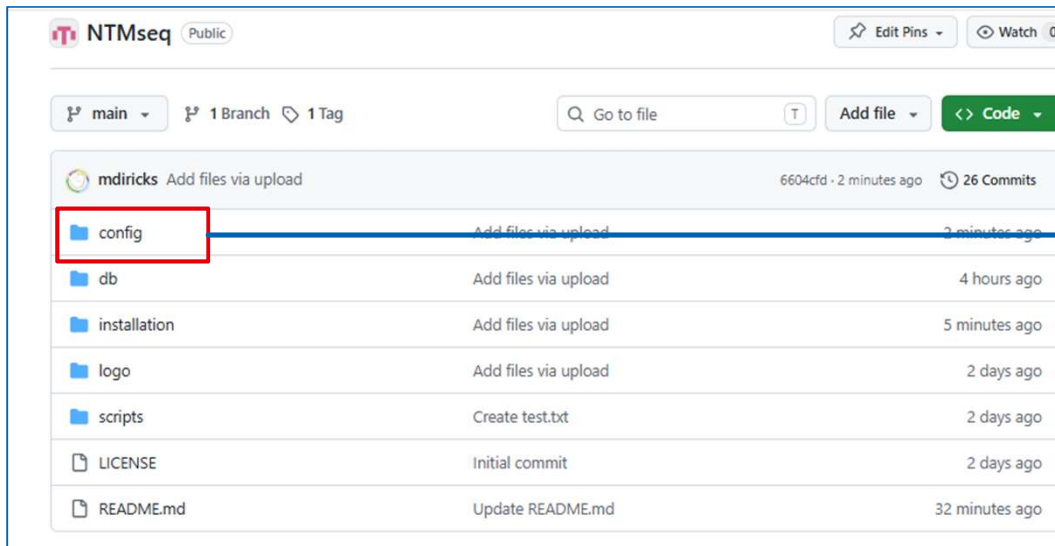
mdiricks Add files via upload

Name	Last commit message
..	
Extract_fasta_from_multifasta.sh	Add files via upload
MashDistance_to_groups.R	Add files via upload
PLSDB_to_SRST2_db.sh	Add files via upload
Query_plasmids_plsdbapi.py	Add files via upload
customDB_to_SRST2_db.sh	Add files via upload
guess-encoding.py	Add files via upload
kraken_parse_results.v2.0.0.pl	Add files via upload
starter_Assemble_usingShovill.sh	Add files via upload
starter_MLST_SRST2.sh	Add files via upload
starter_Mashtree.sh	Add files via upload
starter_NTMseq.sh	Add files via upload
starter_PLSDB_SRST2.sh	Add files via upload
starter_Platon.sh	Add files via upload
starter_abricate.sh	Add files via upload
starter_amrfinder.sh	Add files via upload
starter_customDB_SRST2.sh	Add files via upload
starter_fastANI.sh	Add files via upload
starter_fastANI_AllAgainstAll.sh	Add files via upload
starter_fastp.sh	Add files via upload
starter_kraken2.sh	Add files via upload
starter_multiqc.sh	Add files via upload
starter_plasmidspades.sh	Add files via upload
starter_subsampling.sh	Add files via upload

Run analyses: `$ bash starter_NTMseq.sh ../config/NTMseq.config`

orchestrates the execution of multiple subscripts

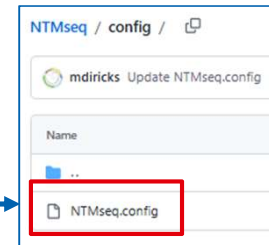
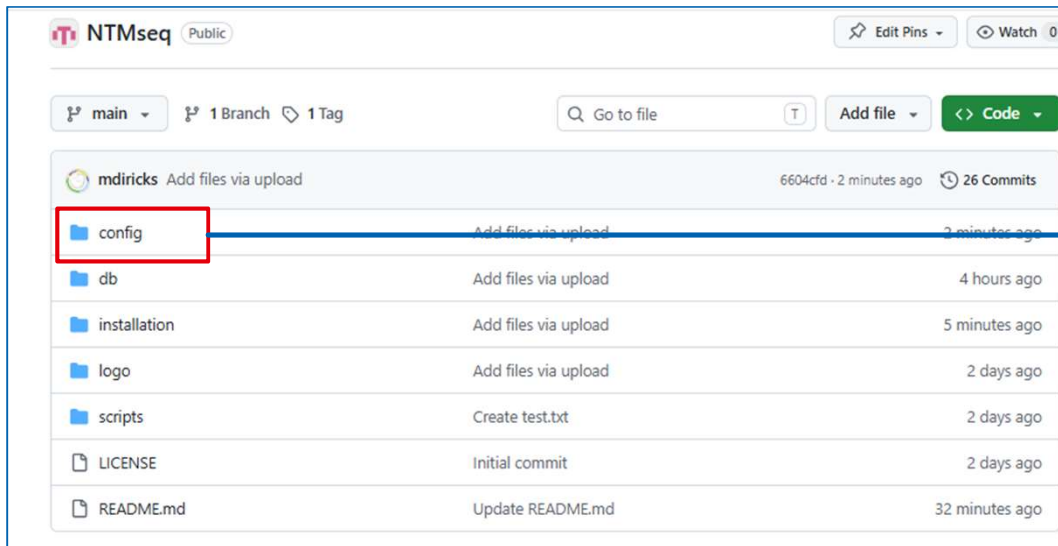
CONFIG



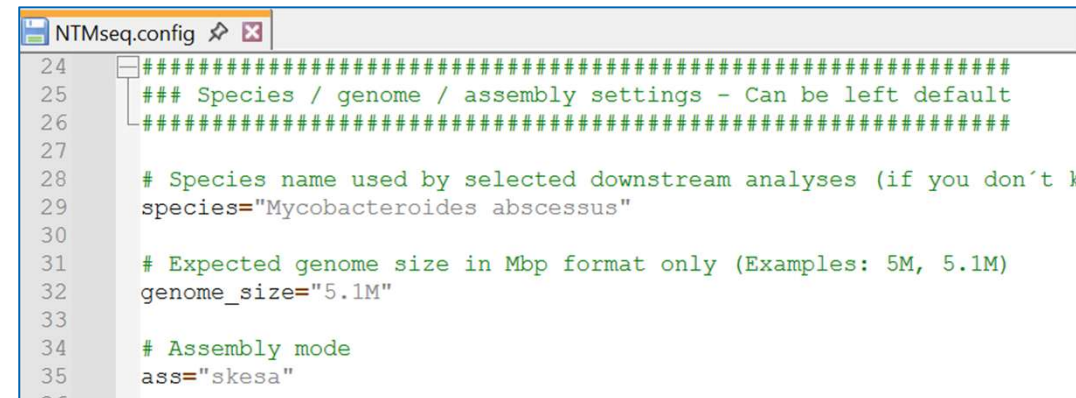
```
NTMseq.config
1  #!/usr/bin/env bash
2  #version=v1.0.0
3
4  #####
5  ### NTMseq configuration
6  #####
7  # This config file is mandatory.
8  #NTMseq is for Research use only!
9
10 #####
11 ### Input / Output - REQUIRED
12 #####
13
14 # Path to the folder containing the NTMseq scripts
15 PATH_scripts="/path/to/NTMseq/scripts"
16
17 # Path to folder containing input FastQ files
18 PATH_fastQ="/path/to/FastQ"
19 #Note: Output folders and result files are named using only
20
21 # Path to folder where results will be written
22 PATH_output="/path/to/NTMseq/Results"
23
```

Run analyses: `$ bash starter_NTMseq.sh ../config/NTMseq.config`

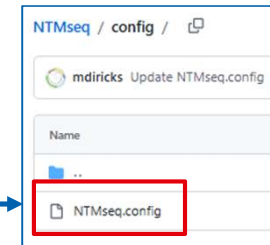
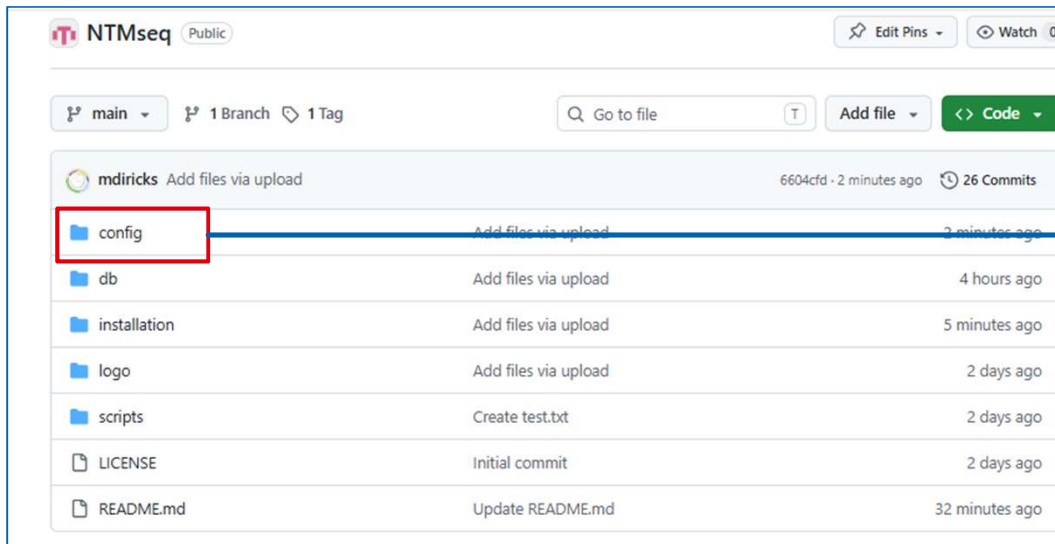
CONFIG



Run analyses: `$ bash starter_NTMseq.sh ../config/NTMseq.config`



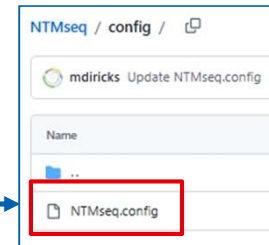
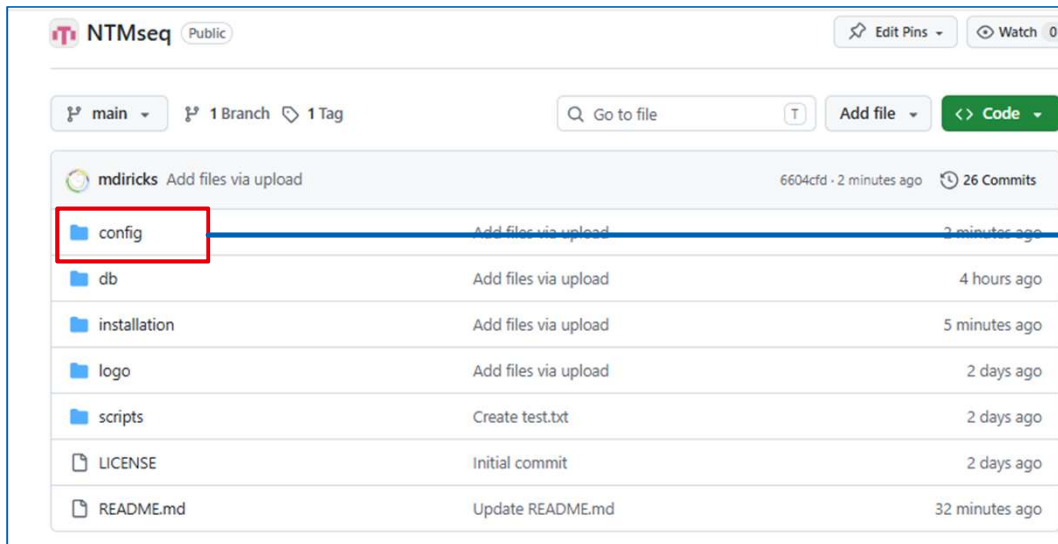
CONFIG



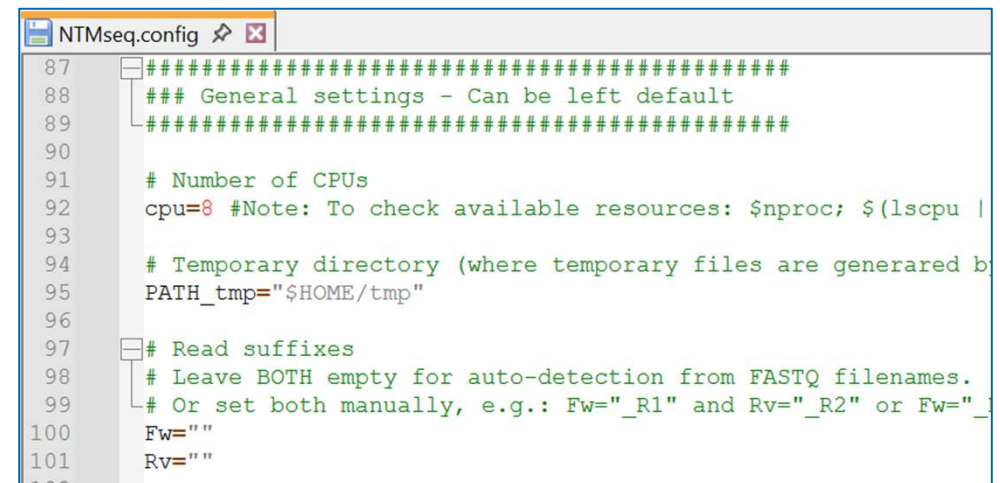
Run analyses: `$ bash starter_NTMseq.sh ../config/NTMseq.config`

```
#####  
### Analysis switches  
#####  
# Allowed values: Yes or No  
  
#Quality control FastQ files  
Do_multiqc="Yes"  
  
#Subsampling of high-coverage reads  
Do_subsampling="Yes"  
cov=100 # Coverage to which samples will be downsampled  
  
#FastQ preprocessing (e.g. adapter trimming, removal of low quality bases)  
Do_fastp="Yes"  
  
#Contamination detection (Requires path to databases to be specified at bottom of this config!!)  
Do_kraken2="Yes"  
  
#NTM species and resistance detection  
Do_NTMprofiler="Yes"  
#Database required, see https://github.com/ngs-fzb/NTMseq and set path to database below!!  
  
#M. abscessus MLST typing  
Do_MLST_fastQ="Yes"  
  
#Assembly with shovill (includes downsampling to cov 100x and adapter trimming)  
Do_assembly="Yes"  
  
#Check average nucleotide identity between your sequence and a reference or between all your sequences  
Do_fastANI="No" #Against one reference, which needs to be specified in the "databases" section of this config  
Do_fastANI_AllagainstAll="Yes"
```

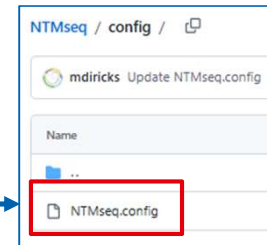
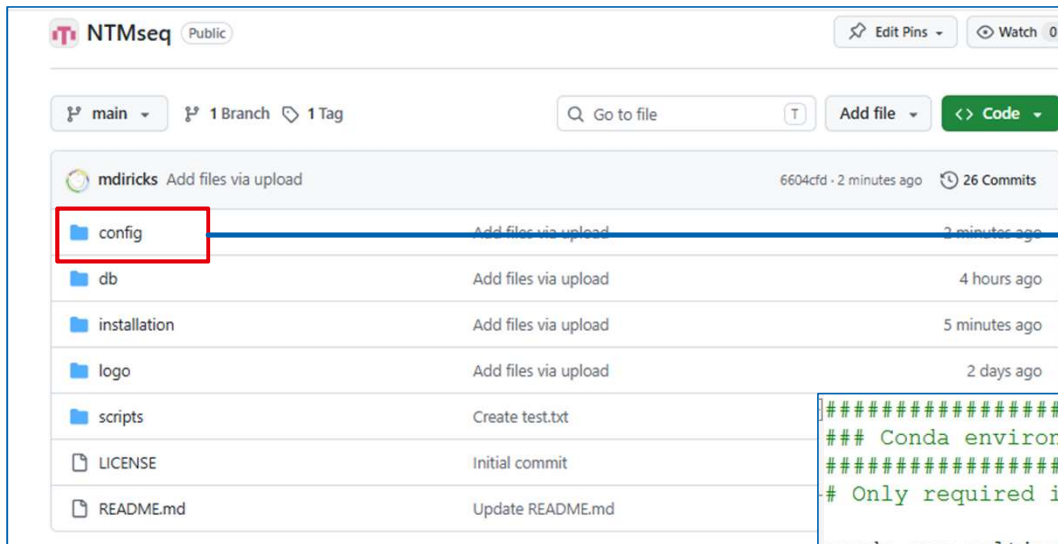

CONFIG



Run analyses: `$ bash starter_NTMseq.sh ../config/NTMseq.config`



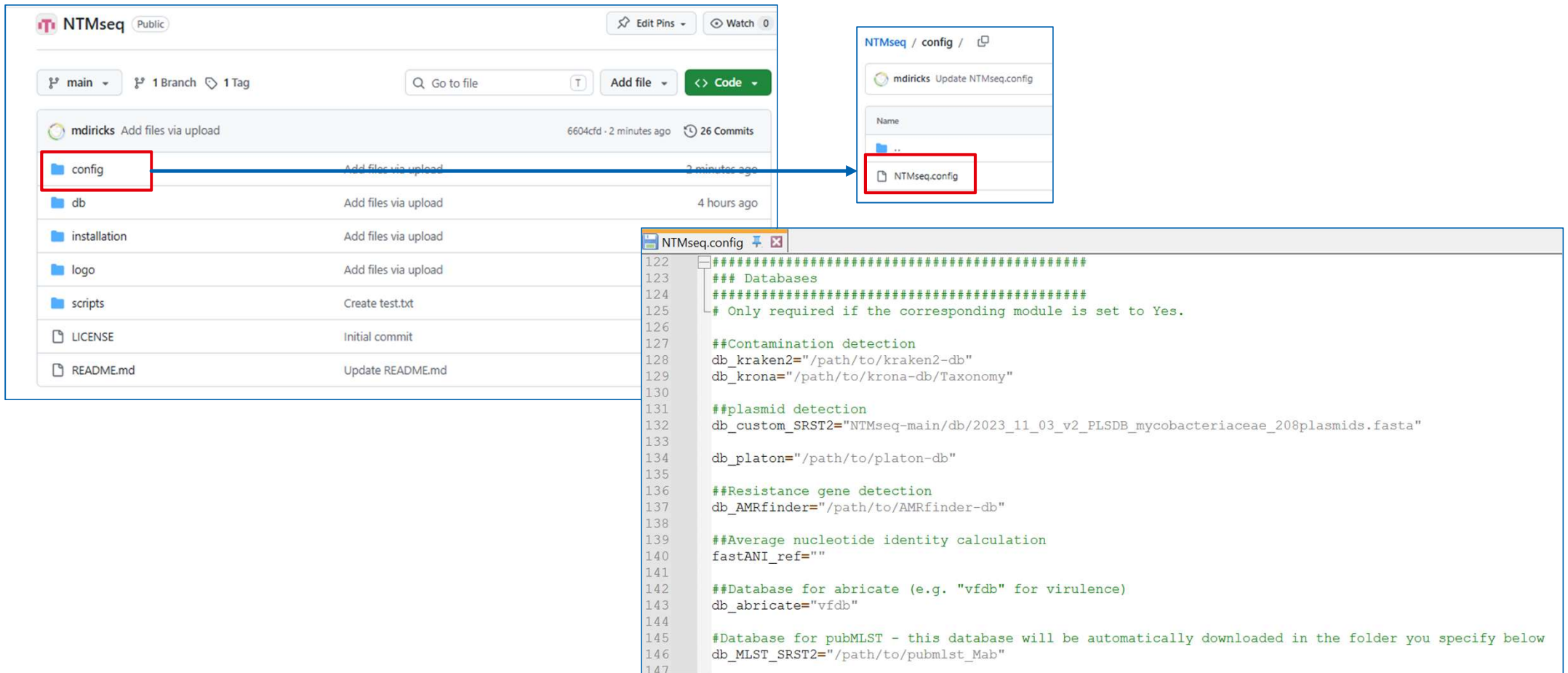
CONFIG



```
#####  
### Conda environment names - Can be left default if installed with Installation_NTMseq.sh  
#####  
# Only required if the corresponding module is set to Yes.  
  
conda_env_multiqc="NTMseq_multiqc"  
conda_env_seqkit="NTMseq_seqkit"  
conda_env_fastp="NTMseq_fastp"  
conda_env_fastANI="NTMseq_fastani"  
conda_env_kraken2="NTMseq_kraken2"  
conda_env_SRST2="NTMseq_srst2"  
conda_env_NTMprofiler="NTMseq_NTMprofiler"  
conda_env_shovill="NTMseq_shovill"  
conda_env_mashtree="NTMseq_mashtree"  
conda_env_AMRfinder="NTMseq_amrfinder"  
conda_env_abricate="NTMseq_abricate"  
conda_env_spades="NTMseq_spades"  
conda_env_platon="NTMseq_platon"
```

Run analyses: `$ bash starter_NTMseq.sh ../config/NTMseq.config`

CONFIG



The screenshot displays the GitHub repository for NTMseq. The repository is public and has 1 branch and 1 tag. The file list shows the following structure:

- config (highlighted with a red box)
- db
- installation
- logo
- scripts
- LICENSE
- README.md

The config directory is expanded, showing the file NTMseq.config (highlighted with a red box). The contents of NTMseq.config are as follows:

```
122 #####
123 ### Databases
124 #####
125 # Only required if the corresponding module is set to Yes.
126
127 ##Contamination detection
128 db_kraken2="/path/to/kraken2-db"
129 db_krona="/path/to/krona-db/Taxonomy"
130
131 ##plasmid detection
132 db_custom_SRST2="NTMseq-main/db/2023_11_03_v2_PLSDb_mycobacteriaceae_208plasmids.fasta"
133
134 db_platon="/path/to/platon-db"
135
136 ##Resistance gene detection
137 db_AMRfinder="/path/to/AMRfinder-db"
138
139 ##Average nucleotide identity calculation
140 fastANI_ref=""
141
142 ##Database for abricate (e.g. "vfdb" for virulence)
143 db_abricate="vfdb"
144
145 #Database for pubMLST - this database will be automatically downloaded in the folder you specify below
146 db_MLST_SRST2="/path/to/pubmlst_Mab"
147
```

Run analyses: `$ bash starter_NTMseq.sh ../config/NTMseq.config`

RUN ANALYSIS



Research Center Borstel
Leibniz Lung Center

```
(base) mdiricks@my-hostname:~$  
(base) mdiricks@my-hostname:~$ mkdir NTMseq  
(base) mdiricks@my-hostname:~$ cd NTMseq  
(base) mdiricks@my-hostname:~/NTMseq$ wget https://github.com/ngs-fzb/NTMseq/archive/refs/heads/main.tar.gz && tar -xzf main.tar.gz  
(base) mdiricks@my-hostname:~/NTMseq$ cd ./NTMseq-main/installation  
(base) mdiricks@my-hostname:~/NTMseq/NTMseq-main/installation$ bash Installation_NTMseq.sh  
(base) mdiricks@my-hostname:~$ cd ~/NTMseq/NTMseq-main/scripts  
(base) mdiricks@my-hostname:~$ bash starter_NTMseq ../config/NTMseq.config
```



- Quality control of reads
- Trimming
- Downsampling
- Contamination detection
- (Sub)species determination
- Strain typing (MLST)
- Assembly
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- Resistance prediction
- Virulence prediction
- Plasmid prediction

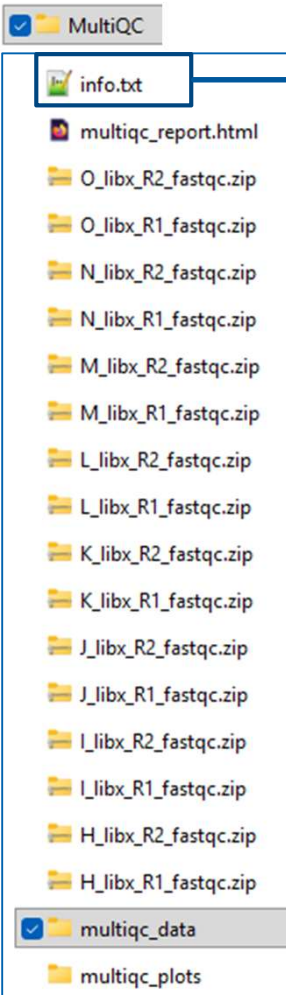
Path/to/output_folder (specified in config file)

☐ Name	Änderungsdatum	Typ
Abriicate	27.04.2026 07:59	Dateiordner
AMRFinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

- **Quality control of reads**
- Trimming
- Downsampling
- Contamination detection
- (Sub)species determination
- Strain typing (MLST)
- Assembly
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- Resistance prediction
- Virulence prediction
- Plasmid prediction

☐ Name	Änderungsdatum	Typ
Abriicate	27.04.2026 07:59	Dateiordner
AMRfinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

NTMSEQ: FASTQC/MULTIQC



So 26. Apr 23:11:22 CEST 2026
Version script: 1.0.1
fastqc version: FastQC v0.12.1
multiqc version: multiqc, version 1.34
Input files: path/to/Tutorial_Mab
Output files path/to/Tutorial_Mab/MultiQC
Amount of threads used: 8
Conda environment: NTMseq_multiqc

All result folders contain an info file with e.g. the version of the tool used

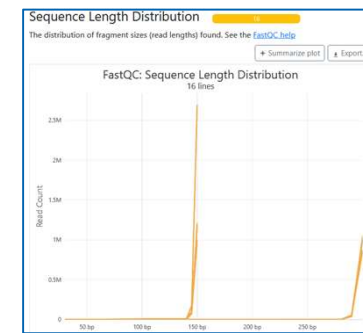
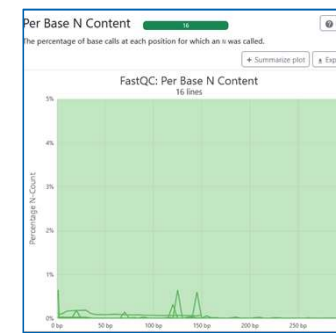
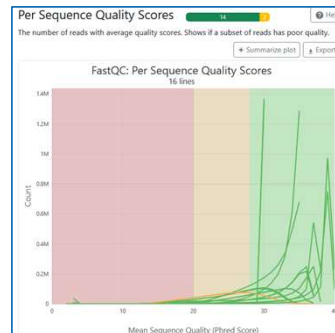
NTMSEQ: FASTQC/MULTIQC



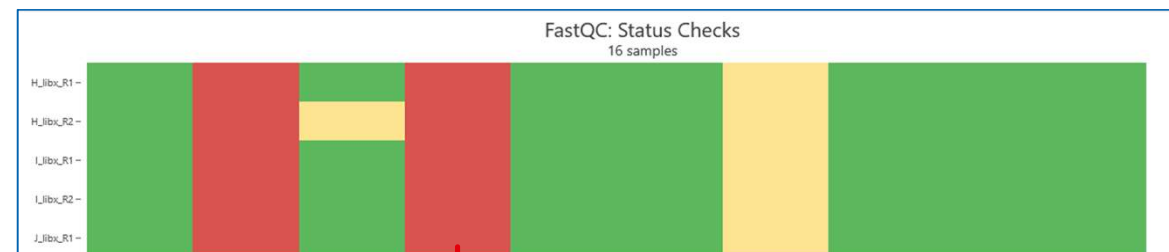
General Statistics

Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... Showing 16/16 rows and 4/6 columns. ↑ Summarize table

Sample Name	Dups	GC	Median len	Seqs
H_libx_R1	17.0 %	63.0 %	301 bp	0.9M
H_libx_R2	11.9 %	64.0 %	301 bp	0.9M
I_libx_R1	20.4 %	63.0 %	301 bp	1.1M
I_libx_R2	15.5 %	64.0 %	301 bp	1.1M



Sample	Filename	File type	Ending	Total Seqs	Total Base Sequence	%GC	Total	Reads	Seqs	Median	Subs	Stat	Per	Base	Per	Seqs
H_libx_R1	H_libx_R1_ConvSanger	II	932827	277.9	Mbp	0.35	301	63	82	95614	298	1405	301	pass	fail	warn
H_libx_R2	H_libx_R2_ConvSanger	II	932827	278.3	Mbp	0.35	301	64	88	1421	298	6405	301	pass	fail	warn
I_libx_R1	I_libx_R1_ConvSanger	II	935454	315.9	Mbp	0.35	301	63	75	59348	299	4802	301	pass	fail	pass
I_libx_R2	I_libx_R2_ConvSanger	II	1055454	316	Mbp	0.35	301	64	84	54113	299	6955	301	pass	fail	pass
J_libx_R1	J_libx_R1_ConvSanger	II	1092223	327.9	Mbp	0.35	301	63	78	10809	304	4282	301	pass	fail	pass
J_libx_R2	J_libx_R2_ConvSanger	II	1092223	328	Mbp	0.35	301	64	83	14869	300	5551	301	pass	fail	pass
K_libx_R1	K_libx_R1_ConvSanger	II	1070201	160.8	Mbp	0.35	151	63	71	51716	150	4763	151	pass	pass	pass
K_libx_R2	K_libx_R2_ConvSanger	II	1070201	160.7	Mbp	0.35	151	64	76	45702	150	488	151	pass	pass	pass
L_libx_R1	L_libx_R1_ConvSanger	II	2004664	434.8	Mbp	0.35	151	66	68	78099	149	9466	151	pass	pass	pass
L_libx_R2	L_libx_R2_ConvSanger	II	2004664	434.6	Mbp	0.35	151	66	75	14159	149	8714	151	pass	pass	pass
M_libx_R1	M_libx_R1_ConvSanger	II	932827	277.9	Mbp	0.35	301	63	82	95614	298	1405	301	pass	fail	pass
M_libx_R2	M_libx_R2_ConvSanger	II	932827	278.3	Mbp	0.35	301	64	88	1421	298	6405	301	pass	fail	warn
N_libx_R1	N_libx_R1_ConvSanger	II	1376793	302.3	Mbp	0.35	151	63	77	28669	147	1504	151	pass	pass	pass
N_libx_R2	N_libx_R2_ConvSanger	II	1376793	302.4	Mbp	0.35	151	63	76	54409	147	1448	151	pass	pass	pass
O_libx_R1	O_libx_R1_ConvSanger	II	1469686	213.6	Mbp	0.31	151	64	78	72932	145	5921	151	pass	pass	pass
O_libx_R2	O_libx_R2_ConvSanger	II	1469686	213.6	Mbp	0.31	151	65	80	94716	145	5793	151	pass	pass	pass



Per base sequence content

- Quality control of reads
- **Trimming**
- Downsampling
- Contamination detection
- (Sub)species determination
- Strain typing (MLST)
- Assembly
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- Resistance prediction
- Virulence prediction
- Plasmid prediction

☐ Name	Änderungsdatum	Typ
Abriicate	27.04.2026 07:59	Dateiordner
AMRFinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

NTMSEQ: FASTP



✓ FastP

Phred_scoring_scale.txt

O.html

N.html

M.html

L.html

K.html

J.html

info.txt

I.html

H.html

FastQ

Summary

General

fastp version:	1.3.2 (https://github.com/OpenGene/fastp)
sequencing:	paired end (301 cycles + 301 cycles)
mean length before filtering:	297bp, 298bp
mean length after filtering:	293bp, 290bp
duplication rate:	0.002787%
Insert size peak:	445

Before filtering

total reads:	1.865654 M
total bases:	556.295675 M
Q20 bases:	415.746159 M (74.734746%)
Q30 bases:	318.185050 M (57.197110%)
Q40 bases:	0 (0.000000%)
GC content:	64.061654%

After filtering

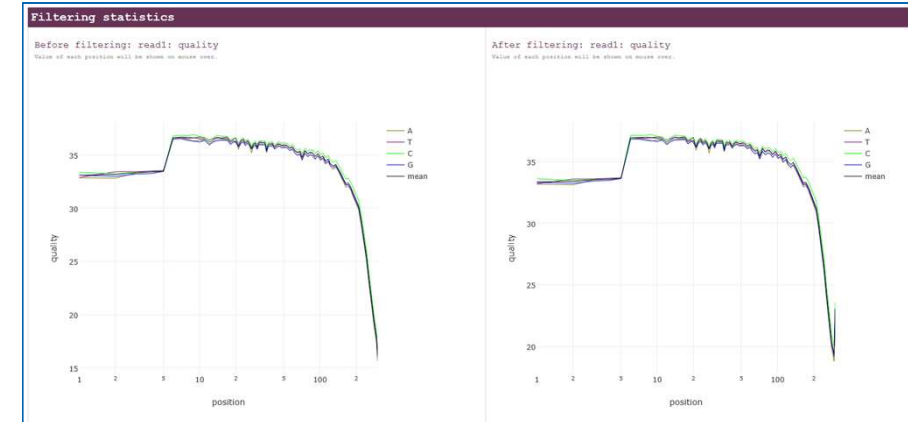
total reads:	1.590404 M
total bases:	464.386385 M
Q20 bases:	370.533258 M (79.789862%)
Q30 bases:	290.128928 M (62.475761%)
Q40 bases:	0 (0.000000%)
GC content:	64.076371%

H_libx_R1.fastq.gz_fastp_R1.fastq.gz

H_libx_R1.fastq.gz_fastp_R2.fastq.gz

I_libx_R1.fastq.gz_fastp_R1.fastq.gz

I_libx_R1.fastq.gz_fastp_R2.fastq.gz



- By default
 - Adapters are trimmed
 - Low quality bases are trimmed from end
 - --cut_tail
 - --cut_tail_window_size 1
- Can be changed in scripts/starter_fastp.sh
- Change path to fastQ in NTMseq.config if you want to use the trimmed fastQ files as input for other tools

- Quality control of reads
- Trimming
- **Downsampling**
- Contamination detection
- (Sub)species determination
- Strain typing (MLST)
- Assembly
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- Resistance prediction
- Virulence prediction
- Plasmid prediction

☐ Name	Änderungsdatum	Typ
Abriicate	27.04.2026 07:59	Dateiordner
AMRFinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

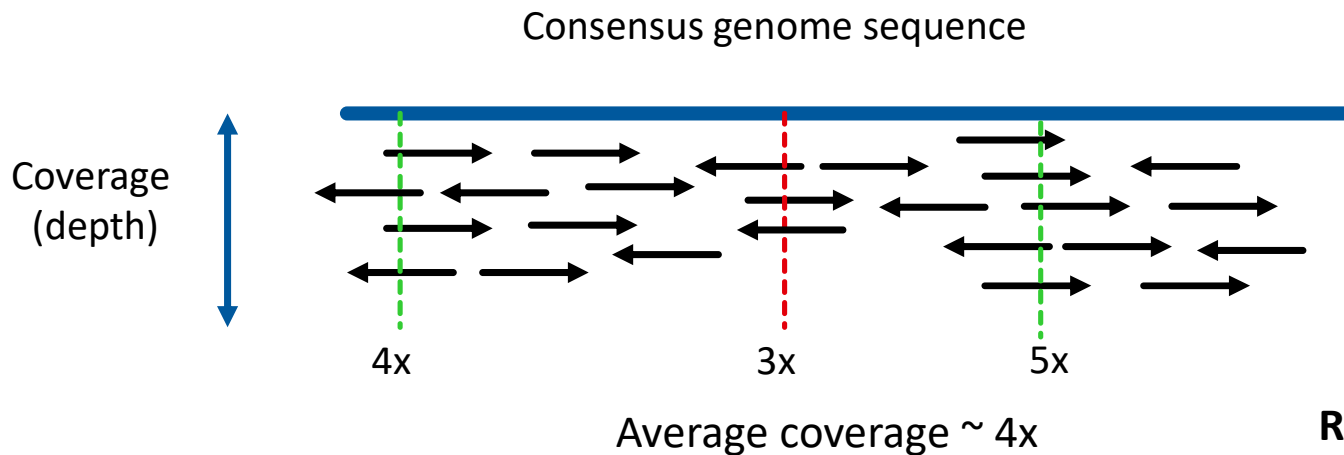
NTMSEQ: DOWNSAMPLING (SEQKIT)



- Coverage depth = how many times, on average, each base of a genome has been sequenced

$$C = \frac{N \times L}{G}$$

C = expected average coverage
N = Total number of reads (2x when paired-end reads)
L = read length
G = genome size



→ = Forward Sequence read

← = Reverse Sequence read

Recommended target coverage per isolate:
~50x-100x (Illumina/nanopore)
~15-30x (PacBio HiFi)

If >300x → downsample

NTMSEQ: DOWNSAMPLING



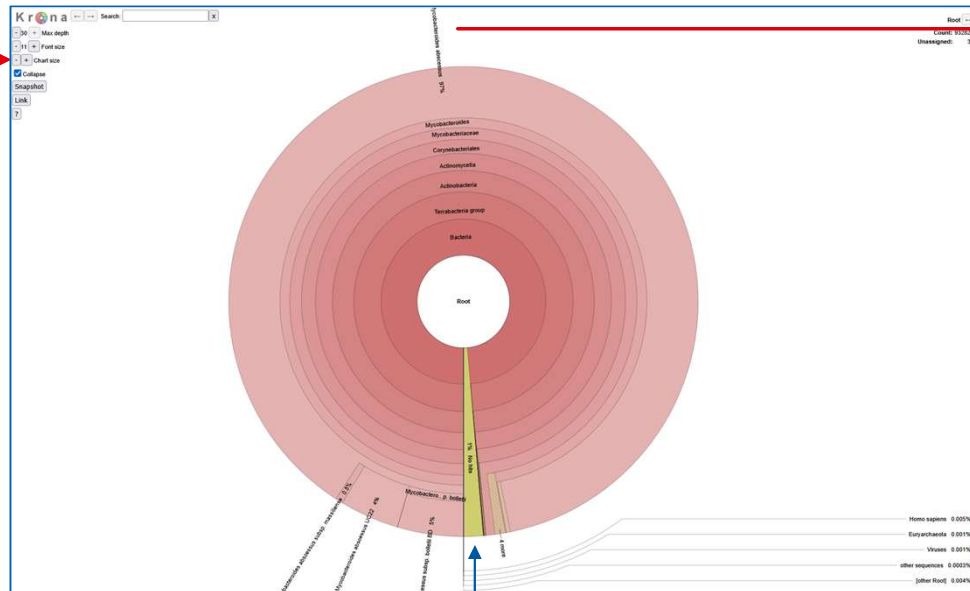
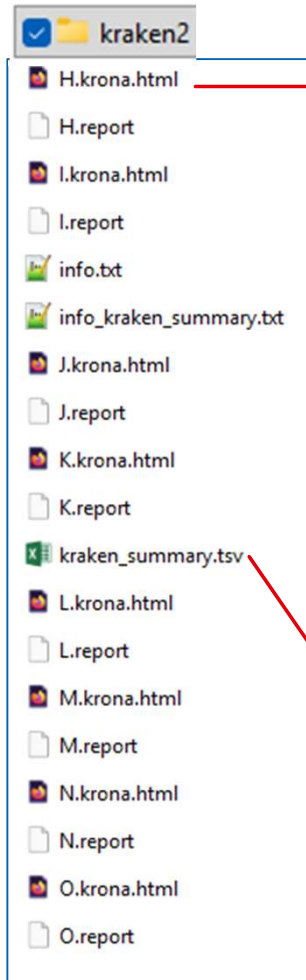
☐ Name	Änderungsdatum	Typ
Abriicate	27.04.2026 07:59	Dateiordner
AMRFinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

☒ FastQ_subsampled
SampleSet_readstats.txt
O_libx_R2.fastq.gz
O_libx_R1.fastq.gz
N_libx_R2.fastq.gz
N_libx_R1.fastq.gz
M-subsampled_R2.fastq.gz
M-subsampled_R1.fastq.gz
L-subsampled_R2.fastq.gz
L-subsampled_R1.fastq.gz
K_libx_R2.fastq.gz
K_libx_R1.fastq.gz
J-subsampled_R2.fastq.gz
J-subsampled_R1.fastq.gz
I-subsampled_R2.fastq.gz
I-subsampled_R1.fastq.gz
H-subsampled_R2.fastq.gz

- Quality control of reads
- Trimming
- Downsampling
- **Contamination detection**
- (Sub)species determination
- Strain typing (MLST)
- Assembly
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- Resistance prediction
- Virulence prediction
- Plasmid prediction

☐ Name	Änderungsdatum	Typ
Abriicate	27.04.2026 07:59	Dateiordner
AMRfinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

NTMSEQ: CONTAMINATION (KRAKEN2)



Mycobacteroides abscessus

Count: 902859
Unassigned: 817649
Taxon: [36809](#)
Rank: species

100% of
Mycobacteroides

98% of
Mycobacteriaceae

98% of
Corynebacteriales

98% of
Actinomycetia

98% of
Actinobacteria

98% of
Terrabacteria group

98% of
Bacteria

97% of
Root

<2-15%: OK!

<2%: OK!

>80%: OK!

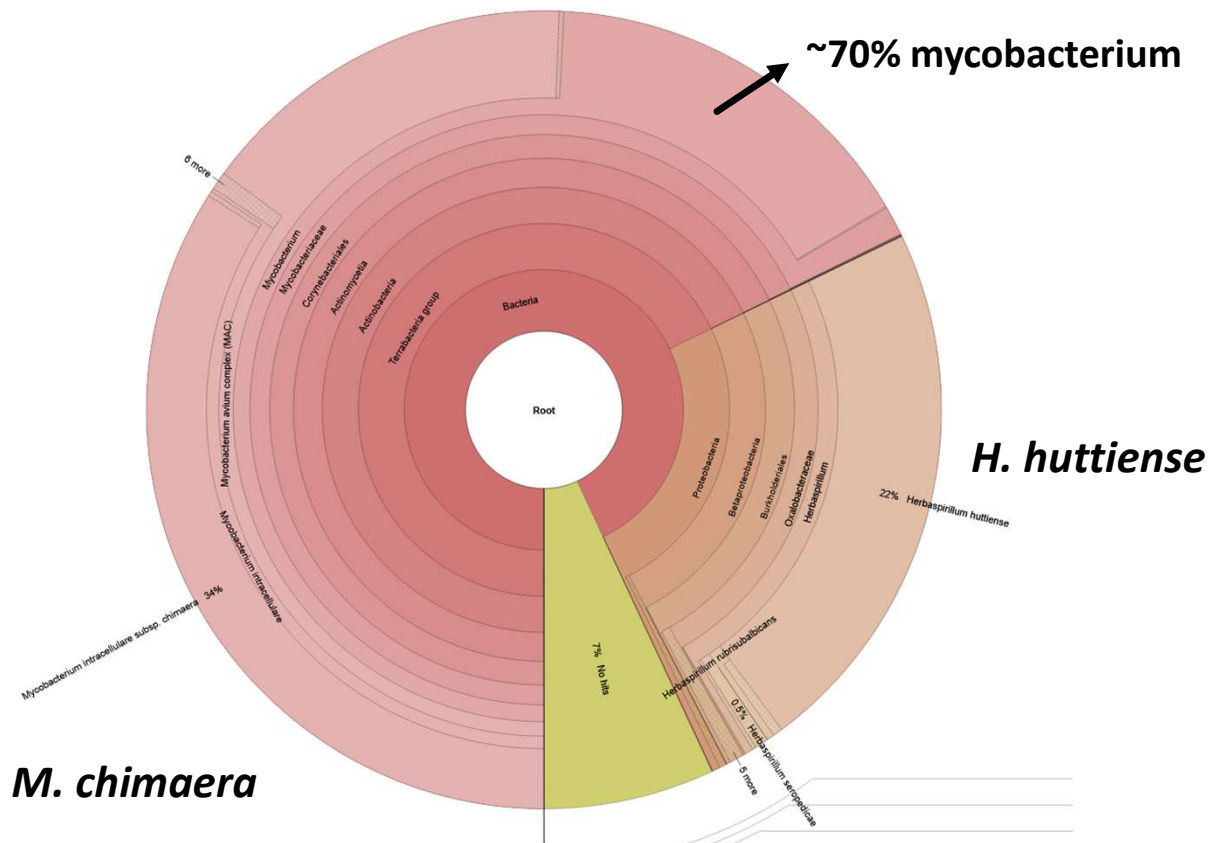
<5%: OK!

<5%: OK!

SampleID	Unclassified_perc	Unclassified_reads	Total_reads	Human_perc	Bacteria_perc	Mycobacterium_perc	Species	Species_perc	Species_other_max	Species_other_max_perc	Species_Myco_max	Species_Myco_max_perc	Species_nonMyco_max	Species_nonMyco_max_perc	Genus_max	Genus_max_perc	Genus_max_nonMyco	Genus_max_nonMyco_perc
H	1.33	12389	932827	0	98.66	0.31	teroides a	96.79	acterium c	0.16	teroides a	96.79	Pluralibacter gergoviae	0.01	cobacteroi	97.12	Nocardioides	0.03
I	1.2	12710	1055454	0.01	98.79	0.31	teroides a	96.97	acterium c	0.18	teroides a	96.97	Streptomyces sp. SGAir0924	0.01	cobacteroi	97.29	Nocardioides	0.03
J	1.28	13982	1092223	0.01	98.71	0.33	teroides a	96.86	acterium c	0.18	teroides a	96.86	Nocardioides sp. KUDC 5002	0.02	cobacteroi	97.19	Nocardioides	0.04
K	1.88	20090	1070201	0	98.11	1.03	teroides a	95.33	cteroides c	0.19	teroides a	95.33	Nocardioides sp. CF8	0.02	cobacteroi	95.69	Nocardioides	0.05
L	0.28	8257	2904664	0	99.72	99.17	teroides a	0	acterium c	52.78	acterium c	52.78	Nitrososphaera viennensis	0	/cobacterit	99.17	Homo	0
M	1.33	12389	932827	0	98.66	0.31	teroides a	96.79	acterium c	0.16	teroides a	96.79	Pluralibacter gergoviae	0.01	cobacteroi	97.12	Nocardioides	0.03
N	1.12	15447	1376793	0	98.86	0.01	teroides a	98.11	cteroides c	0.05	teroides a	98.11	Gordonia virus Lucky10	0	cobacteroi	98.31	Fromanvirus	0.01
O	2.16	31692	1469686	0	97.84	97.37	teroides a	0.02	acterium m	38.06	acterium m	38.06	Pseudomonas entomophila	0.01	/cobacterit	97.37	Streptomyces	0.03

NTMSEQ: CONTAMINATION (KRAKEN2)

- Nice to detect contamination but needs to be interpreted with caution for mycobacteria

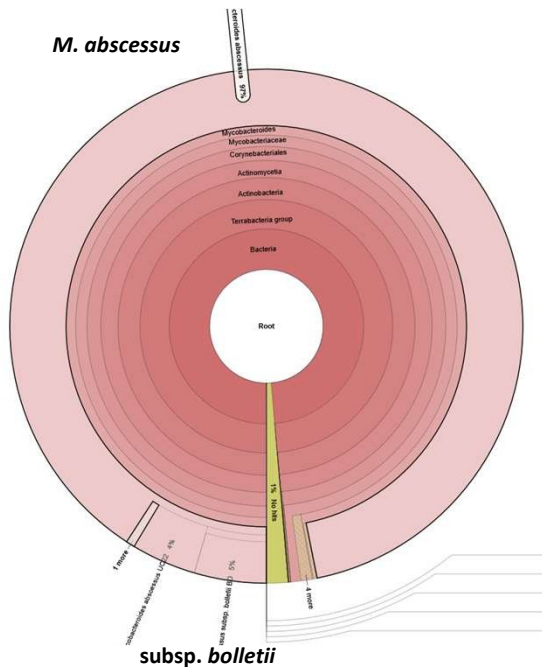


Herbaspirillum huttiense Contamination

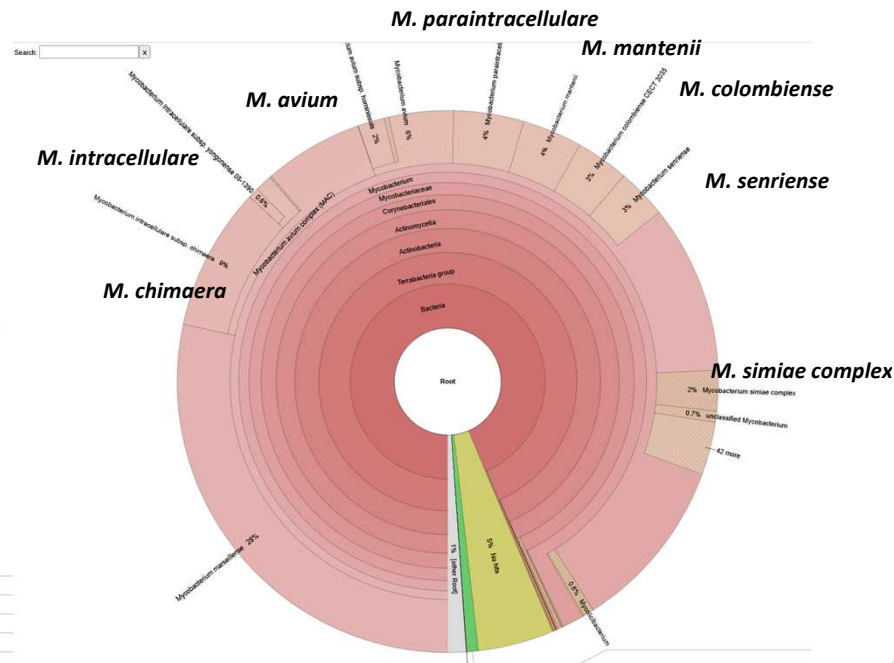
NTMSEQ: CONTAMINATION (KRAKEN2)



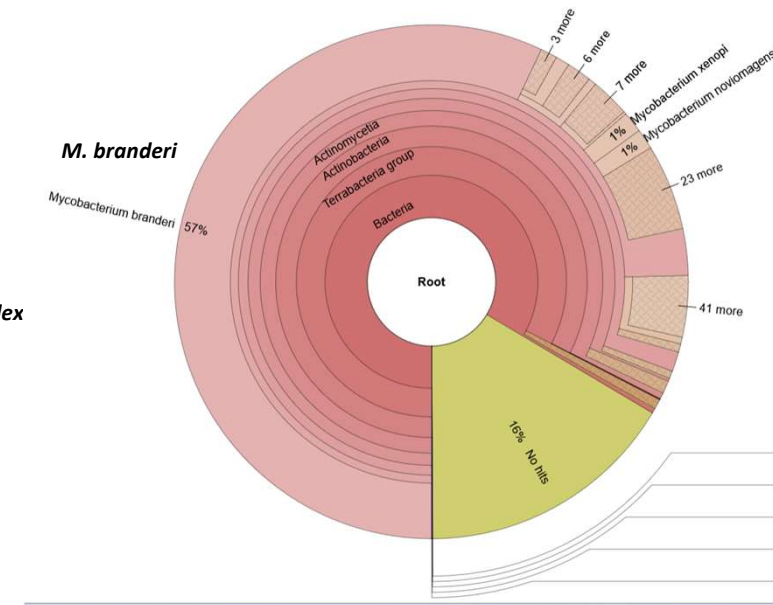
- Nice to detect contamination but needs to be interpreted with caution for mycobacteria
 - Cannot be used for Mab subspecies classification
 - Novel species → False detection of mixed samples; rare species → wrong species identification



M. abscessus subsp *abscessus*



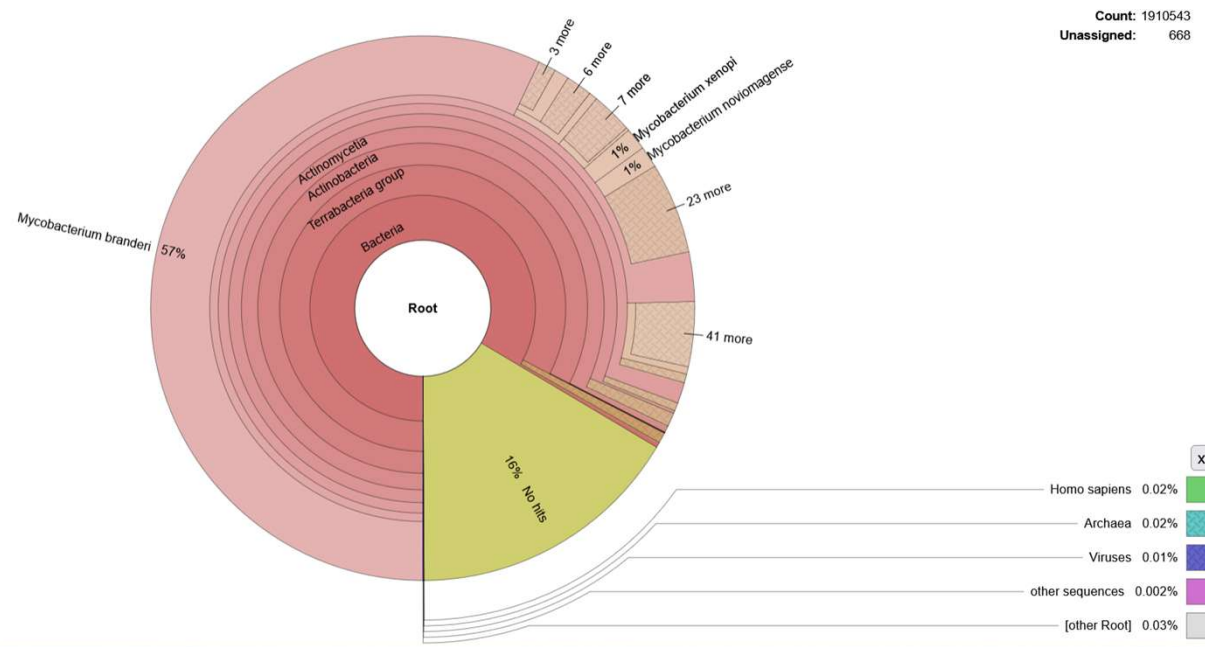
Novel species



M. kyorinense (rare)

NTMSEQ: CONTAMINATION (KRAKEN2)

- Kraken2 databases (containing the reference genomes) are extremely large but smaller laptop-compatible databases can also be downloaded



Full database 2023 ~80 Gb [<https://benlangmead.github.io/aws-indexes/k2>]

Database capped at 16 Gb 2026

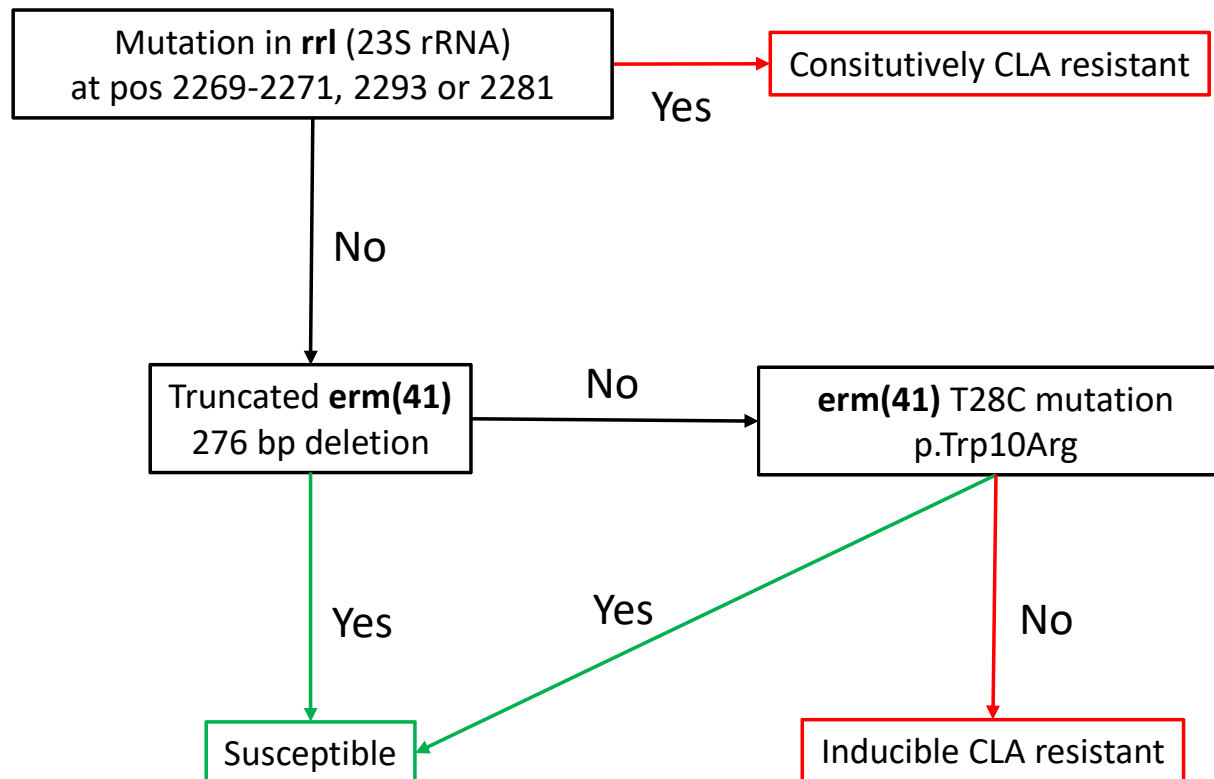
- Quality control of reads
- Trimming
- Downsampling
- Contamination detection
- **(Sub)species determination**
- Strain typing (MLST)
- Assembly
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- **Resistance prediction**
- Virulence prediction
- Plasmid prediction

☐ Name	Änderungsdatum	Typ
Abriicate	27.04.2026 07:59	Dateiordner
AMRFinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

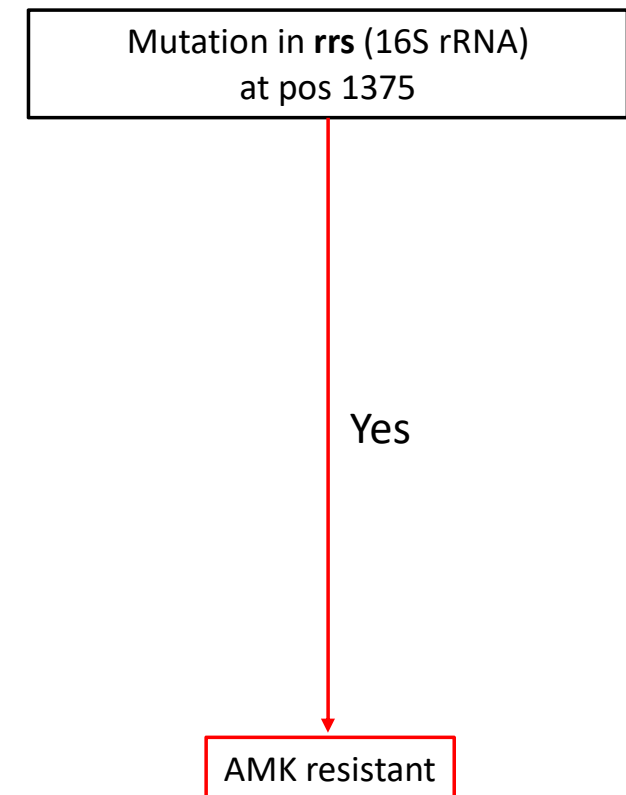
DECISION ALGORITHM RESISTANCE PREDICTION (*M. ABSCESSUS*)



Macrolides



Aminoglycosides



NTMSEQ: NTM-PROFILER



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

pathogen-profiler / ntm-db Public

<> Code Issues Pull requests Actions Projects Security Insights

Files

main

Go to file

- .github
- db
 - Mycobacterium_abscessus
 - Mycobacterium_avium
 - Mycobacterium_fortuitum
 - Mycobacterium_intracellulare
 - Mycobacterium_leprae
 - Mycobacterium_malmoense
 - Mycobacterium_marinum
 - species
- .gitignore
- .gitmodules
- README.md

ntm-db / db /

jodyphelan Merge pull request #28 from pathogen-profiler/Mycobacterium_abscessus

Name	Last commit message
..	
Mycobacterium_abscessus	update schema version
Mycobacterium_avium	update schema version
Mycobacterium_fortuitum	db-schema-version
Mycobacterium_intracellulare	update schema version
Mycobacterium_leprae	update schema version
Mycobacterium_malmoense	update version (#74)
Mycobacterium_marinum	update schema version
species	db-schema-version

Prediction of amikacin/macrolide resistance
Correlates well with in vitro pDST and clinical outcomes

Needs additional validation

github.com/pathogen-profiler/ntm-db/blob/main/db/Mycobacterium_abscessus/variants.csv

Files

main

Go to file

Preview Code Blame 29 lines (29 loc) · 2.55 KB

Search this file

Gene	Mutation	type	drug	literature	hain	E.coli-nomenda
rfrI	n.2269A>C	drug_resistance	macrolides	10.1016/j.jmoldx.2021.07.023	no	A2057C
rfrI	n.2269A>G	drug_resistance	macrolides	10.1016/j.jmoldx.2021.07.023	no	A2057G
rfrI	n.2269A>T	drug_resistance	macrolides	10.1016/j.jmoldx.2021.07.023	no	A2057T
rfrI	n.2270A>C	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2058C
rfrI	n.2270A>G	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2058G
rfrI	n.2270A>T	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2058T
rfrI	n.2271A>C	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2059C
rfrI	n.2271A>G	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2059G
rfrI	n.2271A>T	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2059T
rfrI	n.2281G>A	drug_resistance	macrolides	10.1038/s41467-021-25484-9	no	G2069A
rfrI	n.2281G>C	drug_resistance	macrolides	10.1038/s41467-021-25484-9	no	G2069C
rfrI	n.2281G>T	drug_resistance	macrolides	10.1038/s41467-021-25484-9	no	G2069T
rfrI	n.2293A>C	drug_resistance	macrolides	10.1038/s41467-021-25484-9	no	A2082C
rfrI	n.2293A>G	drug_resistance	macrolides	10.1038/s41467-021-25484-9	no	A2082G
rfrI	n.2293A>T	drug_resistance	macrolides	10.1038/s41467-021-25484-9	no	A2082T
rfrS	n.1373T>A	drug_resistance	amikacin	10.1016/j.jmoldx.2021.07.023	yes	T1406A
rfrS	n.1373T>C	drug_resistance	amikacin	10.1016/j.jmoldx.2021.07.023	no	T1406C
rfrS	n.1373T>G	drug_resistance	amikacin	10.1016/j.jmoldx.2021.07.023	no	T1406G
rfrS	n.1375A>C	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	A1408C
rfrS	n.1375A>G	drug_resistance	amikacin	10.1038/s41467-021-25484-9	yes	A1408G
rfrS	n.1375A>T	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	A1408T
rfrS	n.1376C>A	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	C1409A
rfrS	n.1376C>G	drug_resistance	amikacin	10.1038/s41467-021-25484-9	no	C1409G
rfrS	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://pathogen-profiler.github.io/ntm-db/>

NTM-DB Knowledge Base
NTM-DB
Mycobacterium abscessus
Mycobacterium avium
Mycobacterium fortuitum
Mycobacterium intracellulare
Mycobacterium leprae
Mycobacterium malmoense
Mycobacterium marinum
Species

Mycobacterium avium

General information

Key	value
Species	Mycobacterium avium
Reference sequence accession	CP000479.1
Subspecies detection	Yes
Resistance detection	Yes

Subspecies detection

The following subspecies are detected:

- subsp. hominissuis
- subsp. paratuberculosis

Drug resistance

Drug resistance is detected for:

- rifampicin
- macrolides

Genes of interest

Genes of interest

Gene	Drug	Literature
rpoB	rifampicin	10.1016/j.ijantimicag.2005.09.015
rrs	amikacin	10.1038/s41598-021-85721-5
rrl	macrolides	10.1128/AAC.39.12.2625
MAV_2512	bedaquiline	10.1183/13993003.00118-2019

Resistance mechanisms

Custom Search Builder
Add Condition

Column visibility ▼
Excel

Search:

Gene	Mutation	type	drug	literature	hain	E.coli-nomenclature	gene_id	so_term	cr
rrl	n.2278A>C	drug_resistance	macrolides	Realegeno et al	no	A2057C			
rrl	n.2278A>G	drug_resistance	macrolides	Realegeno et al	no	A2057G			
rrl	n.2278A>T	drug_resistance	macrolides	Realegeno et al	no	A2057T			
rrl	n.2279A>C	drug_resistance	macrolides	10.1038/s41467-021-25484-9	yes	A2058C			

NTMSEQ: NTM-PROFILER



☒
NTMprofiler

☐ Name

info.txt

itol_NTMprofiler_Species.txt

itol_NTMprofiler_Subspecies.txt

ntmprofiler.collate.txt

ntmprofiler.collate.txt.variants.csv

O.results.json

O.results.txt

N.results.json

N.results.txt

N.delly.bcf

N.bam.bai

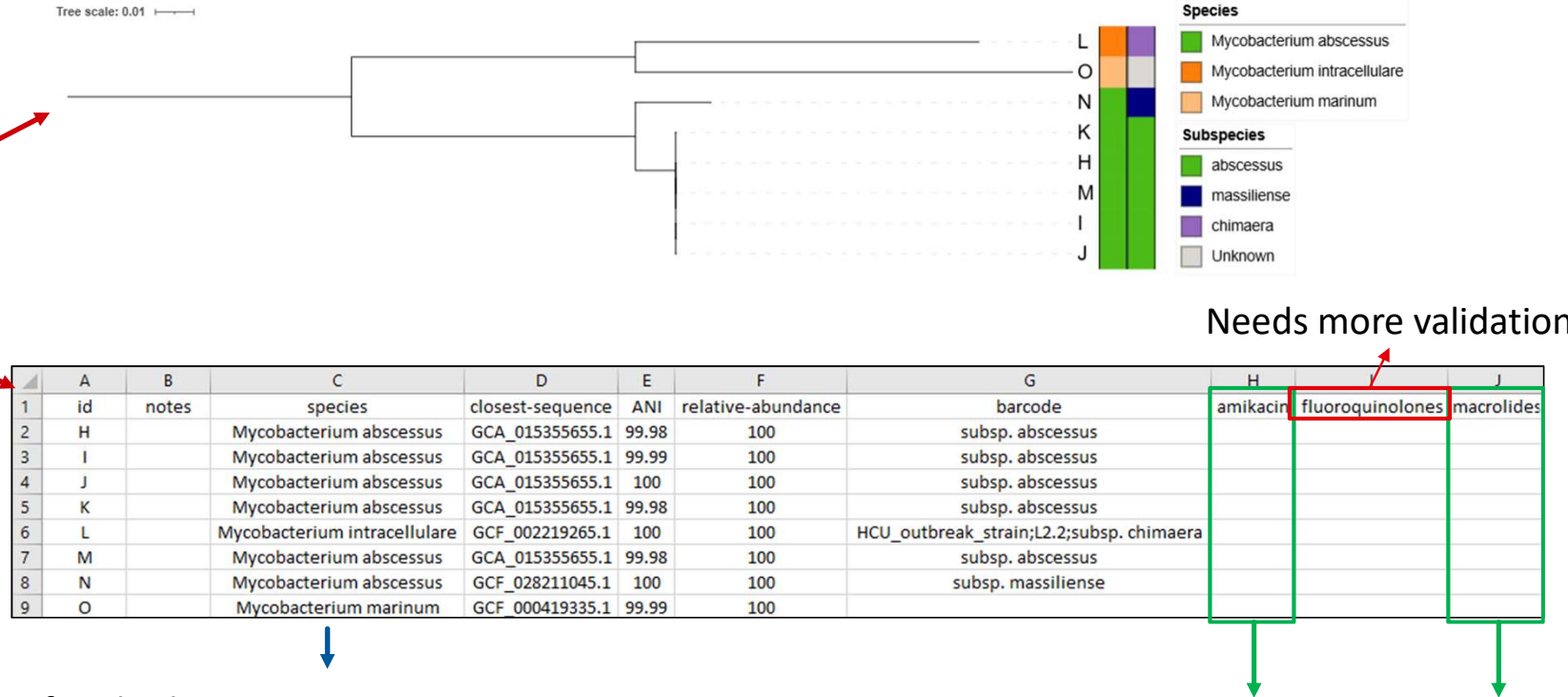
M.results.json

M.results.txt

M.delly.bcf

M.bam.bai

L.results.json



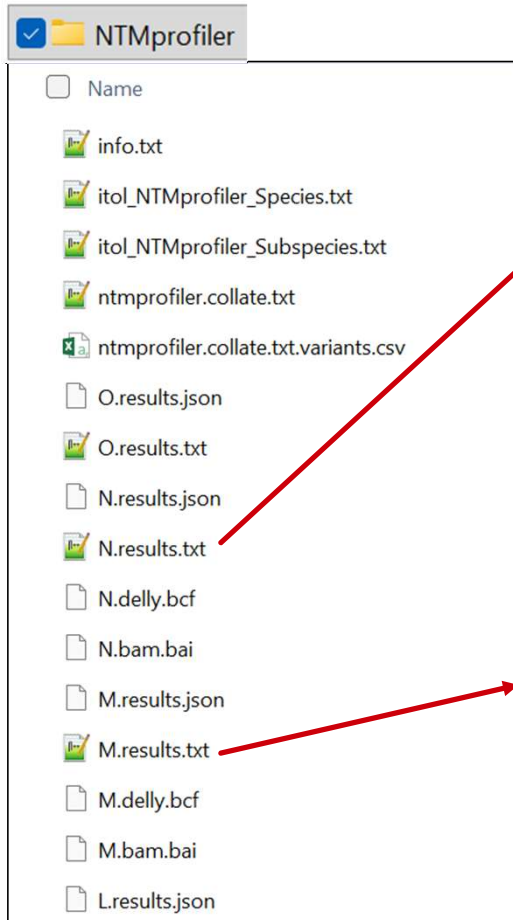
If multiple NTM are present in sample, all species names will be displayed here

Nothing reported in these columns → likely susceptible against amikacin and macrolides

Example if strain would have resistance-associated mutations:

rrl n.2270A>G (1.00); erm(41) (resistance_gene:functionally_normal)

NTMPROFILER



NTM-Profiler report
=====

The following report has been generated by NTM-Profiler.

Summary

ID: N
Date: Mon Apr 27 00:17:23 2026

Notes

Species report

Notes:

Species Accession	Ani	Abundance	Relative Abundance
Mycobacterium abscessus	GCF_028211045.1	100.000	100.000 100.000

Subspecies report

Taxa	Frequency
subsp. massiliense	100.000

2346017	MAB_2297	erm(41)	frameshift_variant	c.64_65delCG	0.982
2346039	MAB_2297	erm(41)	missense_variant	p.Val29Phe	1.000
2346044	MAB_2297	erm(41)	synonymous_variant	c.90C>T	1.000
2346063	MAB_2297	erm(41)	missense_variant	p.Asp37Asn	1.000
2346077	MAB_2297	erm(41)	synonymous_variant	c.123A>G	1.000
2346112	MAB_2297	erm(41)	frameshift_variant	c.159_432del	1.000
2346420	MAB_2297	erm(41)	missense_variant	p.Ala156Thr	1.000
3587645	MAB_3542c	rshA	synonymous_variant	c.270C>T	1.000
3587735	MAB_3542c	rshA	synonymous_variant	c.180T>C	1.000
3587792	MAB_3542c	rshA	synonymous_variant	c.123T>C	1.000
3587822	MAB_3542c	rshA	synonymous_variant	c.93C>G	1.000
3587909	MAB_3542c	rshA	synonymous_variant	c.6C>T	1.000
3587971	MAB_3542c	rshA	upstream_gene_variant	c.-57T>C	1.000
3588028	MAB_3542c	rshA	upstream_gene_variant	c.-114C>T	1.000

Coverage report

Target	Percent_Depth	Pass	Median_Depth
gyrB	95.670	78.000	
gyrA	100.000	76.000	
MAB_0185c	100.000	68.000	
rrs	100.000	89.000	
rsl	100.000	92.000	
erm(41)	62.140	46.000	
rshA	100.000	82.000	

Truncated erm(41) gene

1464318	MAB_r5052	rsl	non_coding_transcript_exon_variant	n.111C>T	0.986
1464484	MAB_r5052	rsl	non_coding_transcript_exon_variant	n.277C>T	1.000
1464540	MAB_r5052	rsl	non_coding_transcript_exon_variant	n.333C>T	1.000
1464557	MAB_r5052	rsl	non_coding_transcript_exon_variant	n.350G>A	1.000
1464841	MAB_r5052	rsl	non_coding_transcript_exon_variant	n.634C>T	1.000
1465142	MAB_r5052	rsl	non_coding_transcript_exon_variant	n.935T>C	0.981
2345982	MAB_2297	erm(41)	loss_of_function_variant	p.Trp10Arg	0.982
2346113	MAB_2297	erm(41)	synonymous_variant	c.159T>C	1.000
2346192	MAB_2297	erm(41)	missense_variant	p.Ile80Val	1.000
2346284	MAB_2297	erm(41)	synonymous_variant	c.330A>C	1.000

Coverage report

Target	Percent_Depth	Pass	Median_Depth
gyrB	95.710	61.000	
gyrA	100.000	64.000	
MAB_0185c	100.000	54.000	
rrs	100.000	61.000	
rsl	100.000	64.000	
erm(41)	100.000	43.000	
rshA	100.000	71.000	

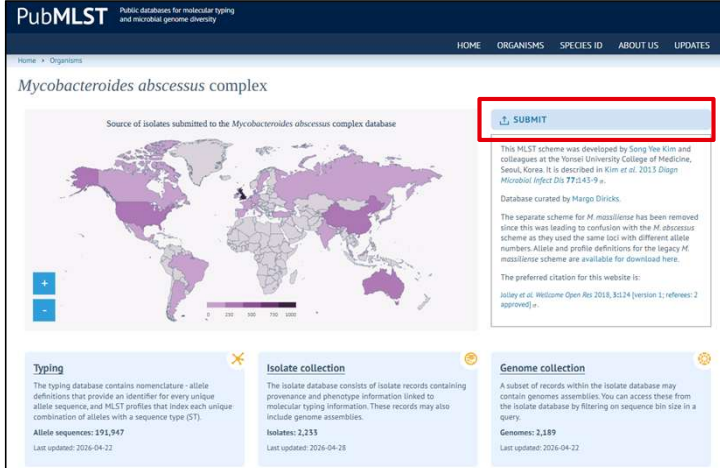
Full length erm(41) gene

Look for low frequency mutations/novel mutations in known resistance-associated genes
BUT
Not all mutations in resistance-associated genes result in resistance!

- Quality control of reads
- Trimming
- Downsampling
- Contamination detection
- (Sub)species determination
- **Strain typing (MLST)**
- Assembly
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- Resistance prediction
- Virulence prediction
- Plasmid prediction

☐ Name	Änderungsdatum	Typ
Abriicate	27.04.2026 07:59	Dateiordner
AMRfinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

MLST_SRST2	
5316_mlst_Mycobacteroides_abscessus_results.txt	15.02.2024 11:53
5830.log	15.02.2024 11:51
5830_mlst_Mycobacteroides_abscessus_results.txt	15.02.2024 11:51
5831.log	15.02.2024 11:49
5831_mlst_Mycobacteroides_abscessus_results.txt	15.02.2024 11:49
6004.log	15.02.2024 11:51
6004_mlst_Mycobacteroides_abscessus_results.txt	15.02.2024 11:51
info.txt	15.02.2024 12:00
LogAvailable.txt	15.02.2024 11:44
MLST_SRST2_SampleSet_compiledResults.txt	15.02.2024 12:00



pubMLST: Public databases for molecular typing and microbial genome diversity

Home > Organisms

Mycobacteroides abscessus complex

Source of isolates submitted to the Mycobacteroides abscessus complex database

SUBMIT

This MLST scheme was developed by Sung Yon Kim and colleagues at the Yonsei University College of Medicine, Seoul, Korea. It is described in Kim et al. 2015 Diagn Microbiol Infect Dis 77:141-9 e.

Database curated by Margo Diricks.

The separate scheme for *M. abscessus* 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. abscessus* scheme are available for download here.

The preferred citation for this website is:
 Jolley et al. Wellcome Open Res 2018, 13:124 (version 1; referees: 2 approved) e.

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: 191,947
 Last updated: 2026-04-22

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: 2,253
 Last updated: 2026-04-28

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: 2,189
 Last updated: 2026-04-22

Sample	ST	argH	cya	gnd	murC	pta	purH	rpoB	mismatches	uncertainty	depth	maxMAF
H_libx	222	1	2	3	2	2	25	4	0	-	31.55814	0.230769
I_libx	222	1	2	3	2	2	25	4	0	-	42.67529	0.166667
J_libx	222	1	2	3	2	2	25	4	0	-	48.78514	0.166667
K_libx	222*	1	2	3	2	2	25*	4	purH_25/1snp	-	15.38471	0.416667
M_libx	222	1	2	3	2	2	25	4	0	-	31.55814	0.230769
N_libx	33	30	31	17	27	21	31	18	0	-	35.66257	0.136364

Might be a novel sequence type, minority variant or a technical artifact → check assembly in pathogenwatch or pubMLST

New sequence type? Submit to pubMLST!

Note: recent change of data access policy → only ST's submitted until 31/12/2024 included

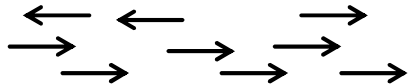
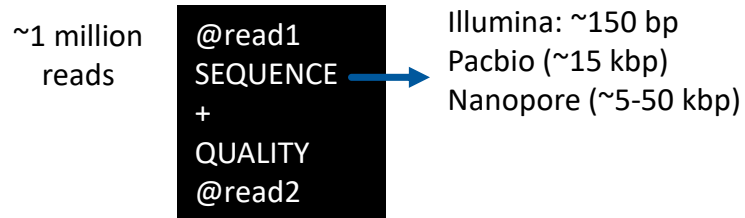
- Quality control of reads
- Trimming
- Downsampling
- Contamination detection
- (Sub)species determination
- Strain typing (MLST)
- **Assembly**
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- Resistance prediction
- Virulence prediction
- Plasmid prediction

☐ Name	Änderungsdatum	Typ
 Abricate	27.04.2026 07:59	Dateiordner
 AMRFinder	27.04.2026 07:59	Dateiordner
 Assemblies	27.04.2026 01:50	Dateiordner
 customDB_SRST2	27.04.2026 07:41	Dateiordner
 fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
 FastP	26.04.2026 23:23	Dateiordner
 FastQ_subsampled	26.04.2026 23:22	Dateiordner
 kraken2	27.04.2026 00:00	Dateiordner
 logs	27.04.2026 07:59	Dateiordner
 Mashtree	27.04.2026 01:54	Dateiordner
 MLST_SRST2	27.04.2026 00:23	Dateiordner
 MultiQC	26.04.2026 23:16	Dateiordner
 NTMprofiler	28.04.2026 13:08	Dateiordner
 Plasmidspades	27.04.2026 06:51	Dateiordner
 Platon	27.04.2026 07:58	Dateiordner

SEQUENCE DATA FILE FORMATS



FastQ



```
@HWI-ST911:111:C0N4WACXX:5:1101:2249:2216 1:N:0:TTAGGC CGATC:??FF
NATGGCACCATTAAAAAGATGTTTATATGTTGTGAGAAGGACAAAGCTGAAGAAGAAATTTAGTCTGCACTTGATGTTGCAATGCAAAGAAA
+
#2A2<CCFHIIIIIIIIIIIGCCHIIIGIIIFHIIIDGHIGIIIIIIICGIIIGGCECEGICFHCECEFFFDDEEEEDDDDDCDDDDDBCB
@HWI-ST911:111:C0N4WACXX:5:1101:2509:2197 1:N:0:TTAGGC CGATC:1+4=B
NATGAGATAATCAATTGCTTAAATGAAGTACAGCTTTTGAATAATGAGTTTTGAACCTCTTGCAACTTTTGGAACTTTAAAGTTTGAATG
+
#4A2<AADHIIIIIIIIIIHHIIIIIIIIIFGIIIGIIIIIDHEHIIIIIIIIIIICHHIIHHEEDFFFFEEEEECEEEADDFC
@HWI-ST911:111:C0N4WACXX:5:1101:3746:2179 1:N:0:TTAGGC CCATC:+11+A
NATGTCATCCATCTTTTCTATCTAAAAAAGAATCAAAAAAGGATAGTACAGAGGGAAGTTCAATCCAGAGGACGATGAAACACTGATTGATGG
+
```

e.g. Isolate1_R1.fastq.gz

200 MB

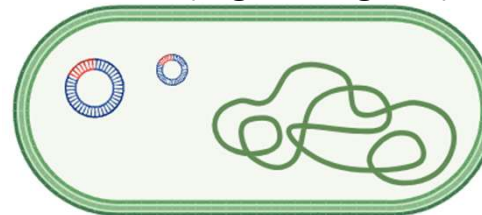


Short reads:
SPAdes, skesa,...

De novo assembly

Long reads:
flye, canu,...

Plasmids (e.g. contig 2-3)



Chromosome (e.g. contig 1)

FastA

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

>contig1
SEQUENCE
>contig2
SEQUENCE

Illumina:
~1.5 Mbp
Pacbio/Nanopore:
5 Mbp

TACGACGTCGTCGTCG GTCTA GTC
Contig 1 Contig 2 Contig 3

Draft genome

```
>contig00001 len=1551651 cov=62.3 corr=0 origname=Contig_11_6
CGCCACCGCCGCCCTCCGCCGGTGACGACGCATGAGGCGCCGCCCGGTGACTACCA
CGACGACACCGCCGCCCTCCGCCCGACGACGACGACCGGCGGGCCACGGCAGA
TCACGTACTCGGTGACCGGTTCCAAGGCTCCGCTCGATCGCATCTCGATCACCTGGACCG
ACGGTTCGGGACGACCCGGGTGAACCCGAATGTGTACATCCCGTGGTTCGATCACGGTTA
CCCCATTTCGAATTCGGAGATCGGATCGGTGTGCGGAGTAGCTTCTGCGGCTGAGTC
AGCTCAACTGCACGATCACCACGACGCGGTCAGGTTTTGTCTCCAACAACAACAAATT
CGGCGCAGGCAACCTGCTGATGCCAGCAGGTCGATTGGCCAAGCCACTGGAGGCAATCGC
GGGCCGCTGCGGCGGTATCGCCGAGTCCGTTGACCGGATGCTCATCGGCTTGTGCGC
```

e.g. Isolate1.fasta

5 MB

Shovill

Assemble bacterial isolate genomes from Illumina paired-end reads

Introduction

The SPAdes genome assembler has become the *de facto* standard *de novo* genome assembler for Illumina whole genome sequencing data of bacteria and other small microbes. SPAdes was a major improvement over previous assemblers like Velvet, but some of its components can be slow and it traditionally did not handle overlapping paired-end reads well.

Shovill is a pipeline which uses SPAdes at its core, but alters the steps before and after the primary assembly step to get similar results in less time. Shovill also supports other assemblers like SKESA, Velvet and Megahit, so you can take advantage of the pre- and post-processing the Shovill provides with those too.

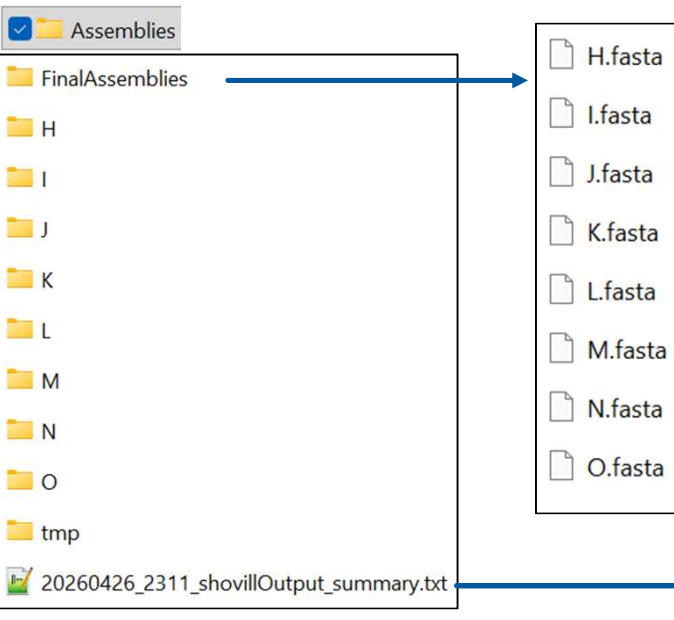
⚠ Shovill is for isolate data only, primarily small haploid organisms. It will **NOT** work on metagenomes or larger genomes. Please use [Megahit](#) directly instead.

Main steps

1. Estimate genome size and read length from reads (unless `--gsz` provided)
2. Reduce FASTQ files to a sensible depth (default `--depth 100`)
3. Trim adapters from reads (with `--trim` only)
4. Conservatively correct sequencing errors in reads
5. Pre-overlap ("stitch") paired-end reads
6. Assemble with SPAdes/SKESA/Megahit with modified kmer range and PE + long SE reads
7. Correct minor assembly errors by mapping reads back to contigs
8. Remove contigs that are too short, too low coverage, or pure homopolymers
9. Produce final FASTA with nicer names and parseable annotations

Too much contigs, assembly length too big → indication of contamination

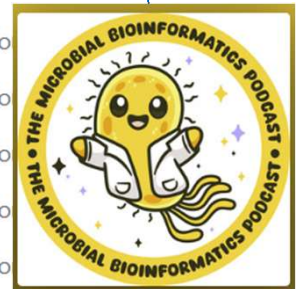
Expected genome length
+/- 20%



Sample	Estimate d_depth(x)	Read_ma x_len	Read_avg _len	Read_mi n_len	Surviving _Read_P airs_trim perc	Fw_survi ving_trim perc	Rv_surviv ing_trim perc	Dropped _trimmo matic_pe rc	Flash_Co mbinedP airs_perc	Walltime	Contig Amount	MinContig Length	Assembly Length	AssemblyR el%(5.1M)	GC%	Amount_ N	Assembler
H	109	[s] max_lers]	avg_lenats]	min_le	98,66	1,33	0,01	0,00	41.65	2 min 2 se	25	17	5004445	-1.87	64.1		skesa
I	123	[s] max_lers]	avg_lenats]	min_le	99,71	0,28	0,01	0,00	42.9	1 min 27 se	25	17	5005506	-1.85	64.1		skesa
J	128	[s] max_lers]	avg_lenats]	min_le	99,94	0,06	0,01	0,00	41.3	1 min 49 se	26	17	5007620	-1.81	64.1		skesa
K	63	[s] max_lers]	avg_lenats]	min_le	99,94	0,03	0,03	0,01	1	1 min 31 se	50	17	4989676	-2.16	64.1		skesa
L	170	[s] max_lers]	avg_lenats]	min_le	99,75	0,22	0,00	0,03	6.16	3 min 31 se	209	17	6263557	22.81	67.5		skesa
M	109	[s] max_lers]	avg_lenats]	min_le	98,66	1,33	0,01	0,00	41.65	1 min 53 se	25	17	5004445	-1.87	64.1		skesa
N	79	[s] max_lers]	avg_lenats]	min_le	99,35	0,62	0,00	0,03	37.02	1 min 23 se	82	17	5230450	2.56	64.1		skesa
O	83	[s] max_lers]	avg_lenats]	min_le	99,87	0,10	0,00	0,02	44.6	1 min 32 se	152	15	6223926	22.04	65.8		skesa

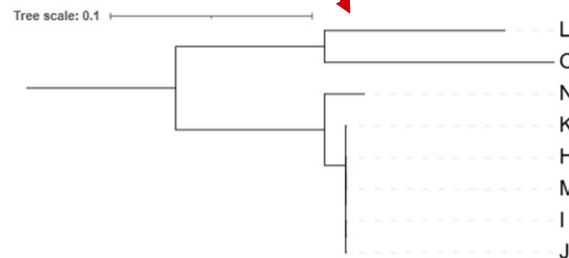
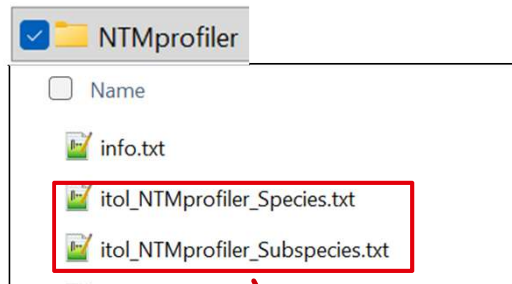
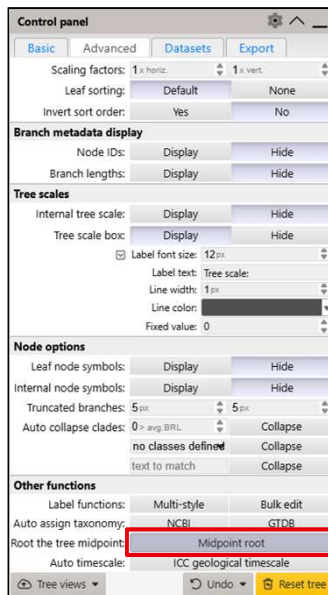
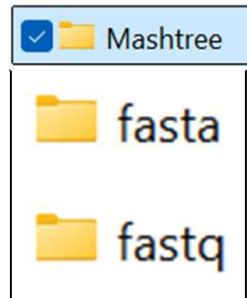
- Quality control of reads
- Trimming
- Downsampling
- Contamination detection
- (Sub)species determination
- Strain typing (MLST)
- Assembly
- **Tree (Dendrogram)**
- Genome similarity estimation (ANI)
- Resistance prediction
- Virulence prediction
- Plasmid prediction

☐ Name	Änderungsdatum	Typ
Abriicate	27.04.2026 07:59	Dateiordner
AMRfinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

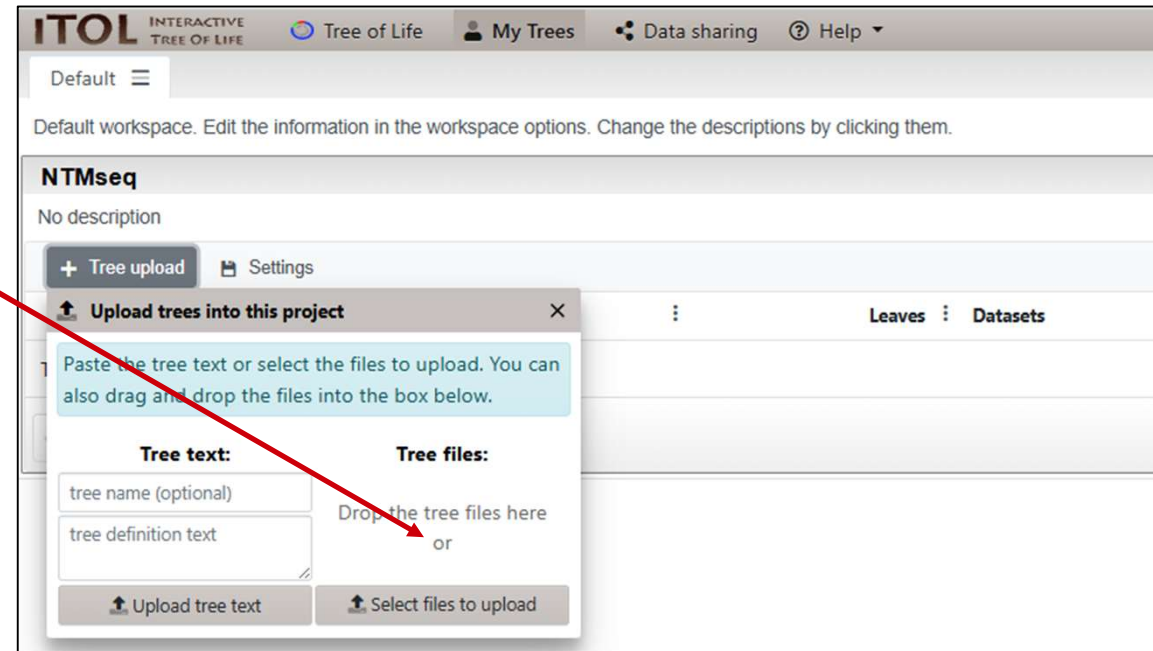


Micro binfie podcast
Microbial Bioinformatics

NTMSEQ: MASHTREE



Species
Mycobacterium abscessus
Mycobacterium intracellulare
Mycobacterium marinum
Subspecies
abscessus
massiliense
chimaera
Unknown



<https://itol.embl.de/login.cgi>

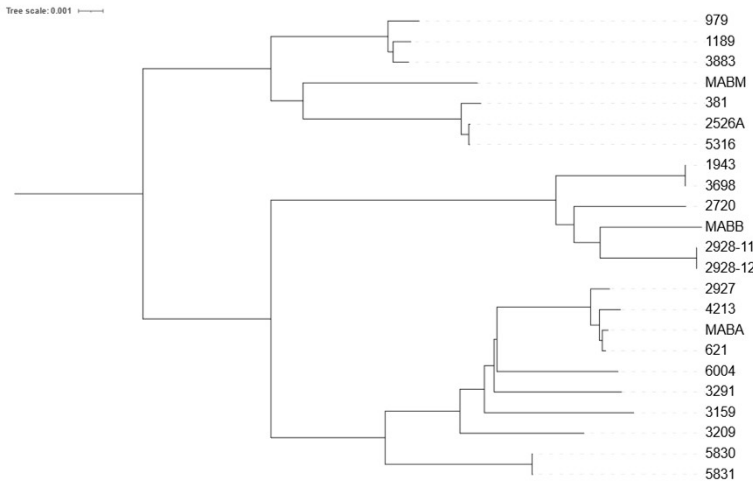
*

	O	I	K	M	L	J	H	N
O		0	0.270747	0.270747	0.270747	0.202933	0.270747	0.287309
I	0.270747		0	5.25E-05	1.19E-05	0.248764	1.19E-05	0.03014
K	0.270747	5.25E-05		0	5.49E-05	0.248764	5.96E-05	0.030159
M	0.270747	1.19E-05	5.49E-05		0	0.248764	1.91E-05	0
L	0.202933	0.248764	0.248764	0.248764		0	0.248764	0.250557
J	0.270747	1.19E-05	5.96E-05	1.91E-05	0.248764		0	1.91E-05
H	0.270747	1.19E-05	5.49E-05		0	0.248764		0
N	0.287309	0.03014	0.030159	0.03014	0.250557	0.030149	0.03014	

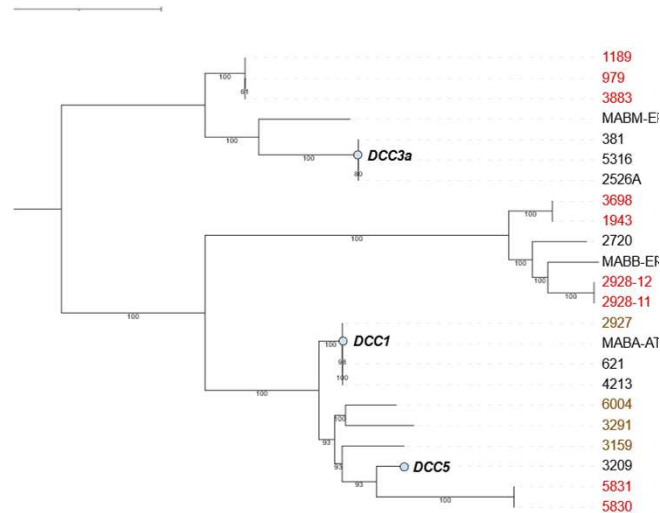
DIFFERENCES IN TREE RECONSTRUCTION ALGORITHMS



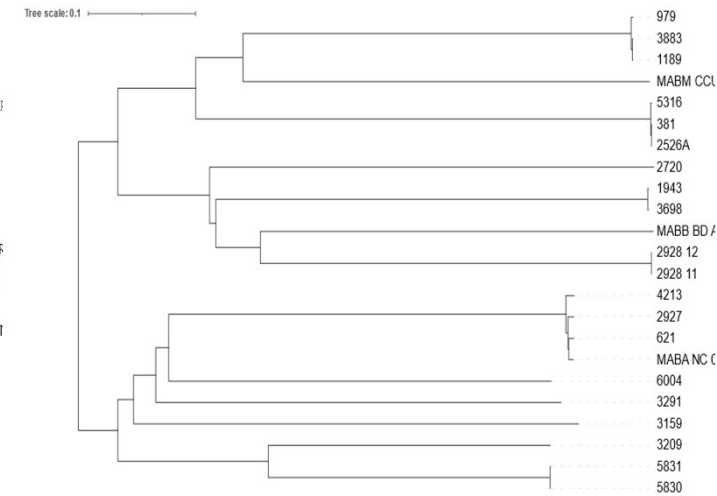
- Differences in tree?



Mashtree (NJ)

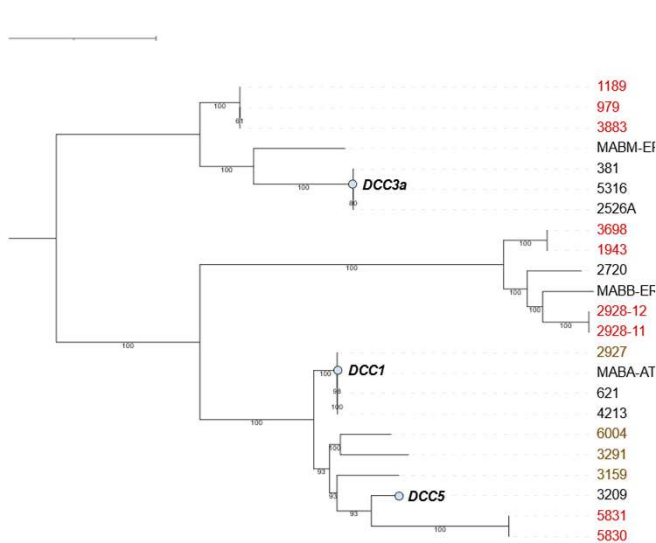


IQtree (ML)

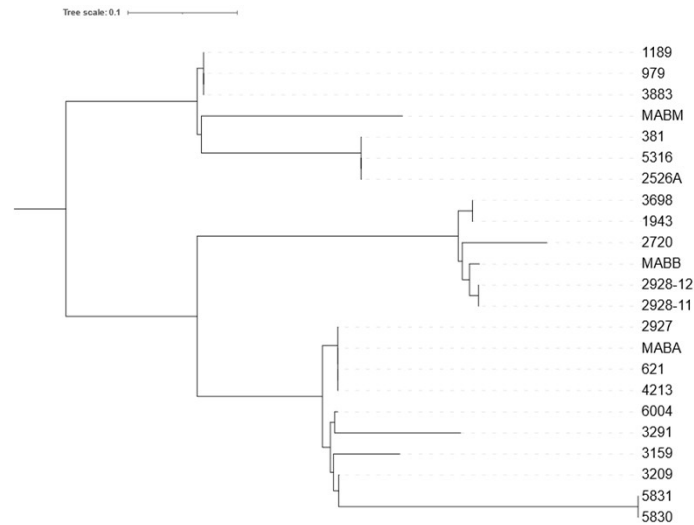


cgMLST (NJ)

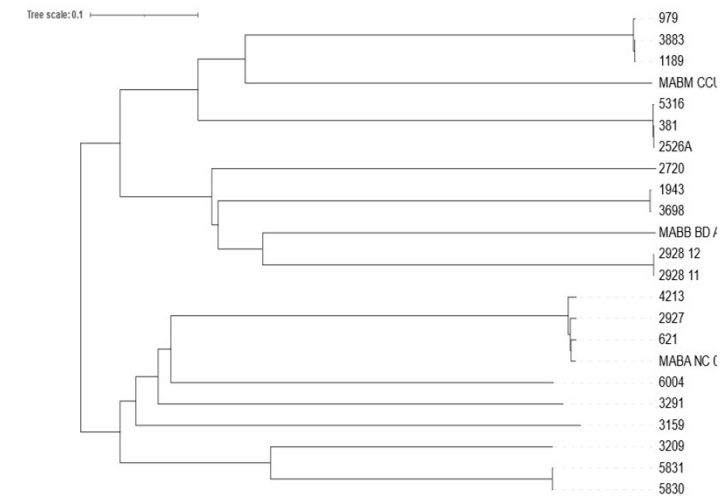
DIFFERENCES IN TREE RECONSTRUCTION ALGORITHMS



IQtree (ML)

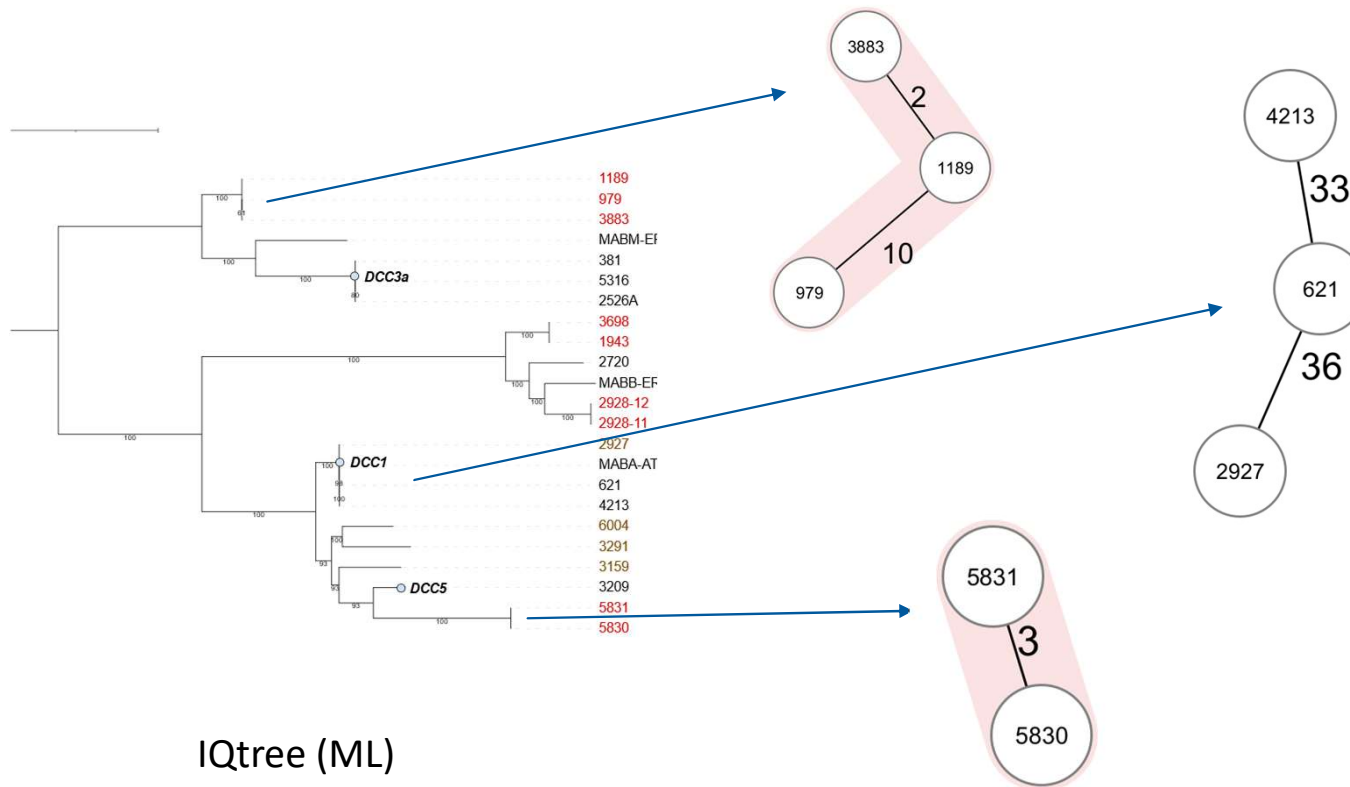


removal of recombination
IQtree (ML)



cgMLST (NJ)

TRANSMISSION/OUTBREAK ANALYSIS

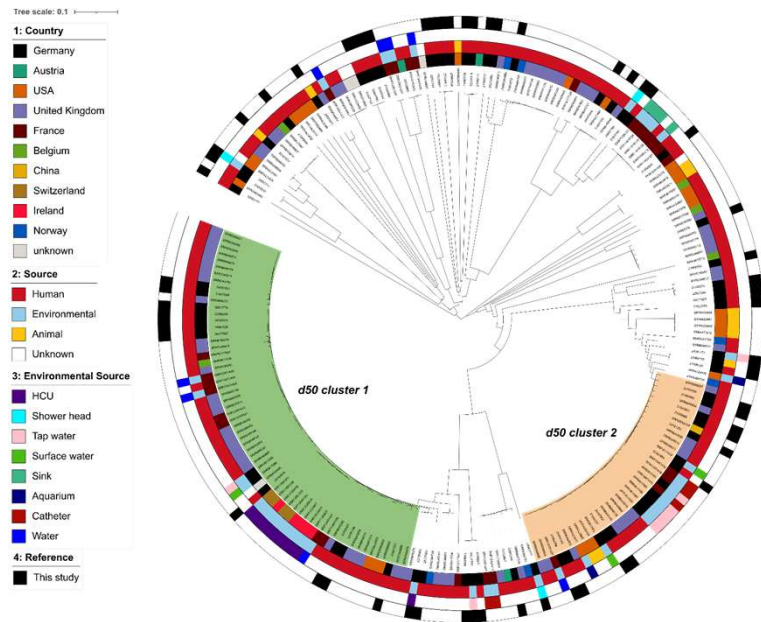


**Short branches in overall
phylogenies do not always
imply direct transmission or an
outbreak**

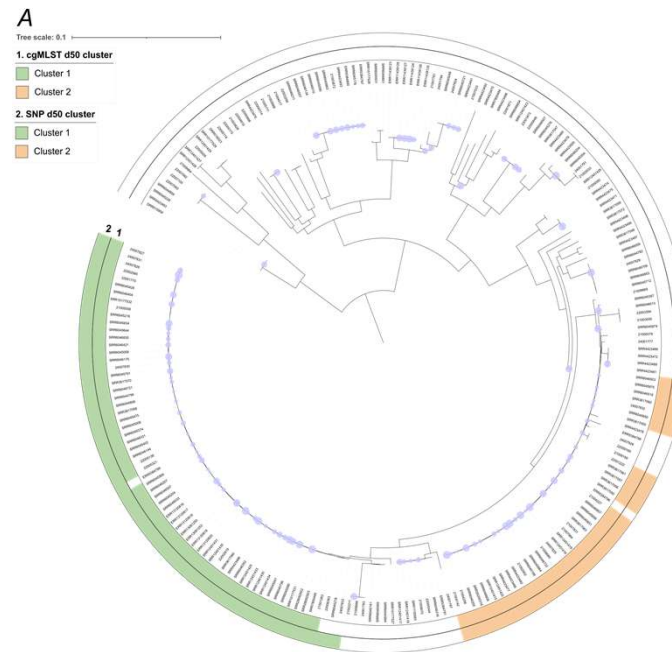
DIFFERENCES IN TREE RECONSTRUCTION ALGORITHMS



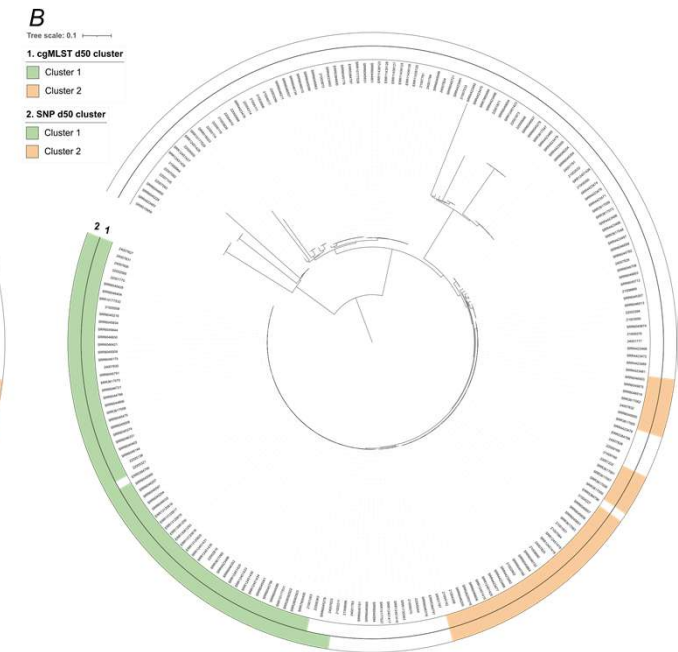
• *M. chelonae*



cgMLST



cgSNP (MTBseq w12)

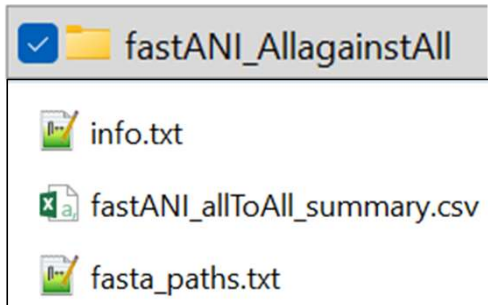


cgSNP (MTBseq + ClonalFrameML)

- Quality control of reads
- Trimming
- Downsampling
- Contamination detection
- (Sub)species determination
- Strain typing (MLST)
- Assembly
- Tree (Dendrogram)
- **Genome similarity estimation (ANI)**
- Resistance prediction
- Virulence prediction
- Plasmid prediction

☐ Name	Änderungsdatum	Typ
Abriicate	27.04.2026 07:59	Dateiordner
AMRfinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

NTMSEQ: AVERAGE NUCLEOTIDE IDENTITY (ANI)



ANI > 95% → same species

	A	B	C	D	E	F
1	Isolate 1	Isolate 2	ANI	# shared fragments	# total fragments	Alignment fraction
2	H	H	100	1654	1656	99.9
3	H	M	100	1654	1656	99.9
4	H	J	99.9965	1648	1656	99.5
5	H	I	99.9956	1649	1656	99.6
6	H	K	99.9806	1630	1656	98.4
7	H	N	96.8833	1455	1656	87.9
8	H	L	77.7967	381	1656	23.0
9	L	L	100	1977	1989	99.4
10	L	O	79.9887	899	1989	45.2
11	L	N	77.9698	371	1989	18.7
12	L	K	77.6031	397	1989	20.0
13	L	M	77.5241	408	1989	20.5
14	L	H	77.5241	408	1989	20.5
15	L	J	77.5231	411	1989	20.7
16	L	I	77.5068	409	1989	20.6
17	N	N	100	1703	1706	99.8
18	N	K	96.9427	1454	1706	85.2
19	N	J	96.9392	1459	1706	85.5
20	N	I	96.9368	1458	1706	85.5
21	N	H	96.9323	1461	1706	85.6
22	N	M	96.9323	1461	1706	85.6
23	N	L	77.7628	398	1706	23.3

Shared fragments/total fragments*100

M. abscessus vs *M. abscessus*

subsp. *abscessus* vs *abscessus*

subsp. *abscessus* vs *massiliense*

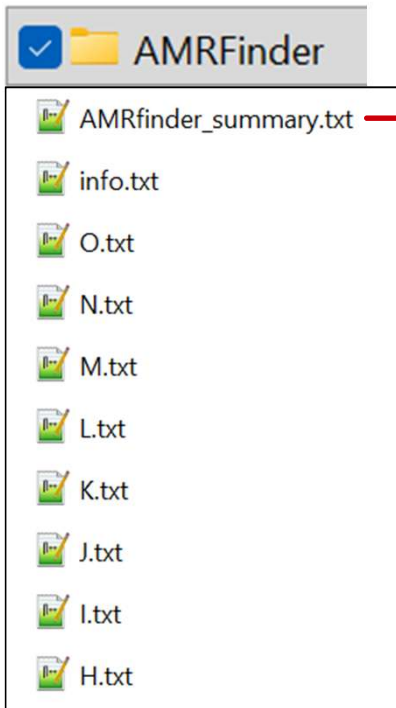
M. abscessus vs *M. chimaera*

Note: Genome pairs with insufficient homology (<~70% ANI) are not reported by FastANI (e.g. H: *M. abscessus* vs O: *M. marinum*)

- Quality control of reads
- Trimming
- Downsampling
- Contamination detection
- (Sub)species determination
- Strain typing (MLST)
- Assembly
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- **Resistance prediction**
- **Virulence prediction**
- Plasmid prediction

☐ Name	Änderungsdatum	Typ
 Abricate	27.04.2026 07:59	Dateiordner
 AMRFinder	27.04.2026 07:59	Dateiordner
 Assemblies	27.04.2026 01:50	Dateiordner
 customDB_SRST2	27.04.2026 07:41	Dateiordner
 fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
 FastP	26.04.2026 23:23	Dateiordner
 FastQ_subsampled	26.04.2026 23:22	Dateiordner
 kraken2	27.04.2026 00:00	Dateiordner
 logs	27.04.2026 07:59	Dateiordner
 Mashtree	27.04.2026 01:54	Dateiordner
 MLST_SRST2	27.04.2026 00:23	Dateiordner
 MultiQC	26.04.2026 23:16	Dateiordner
 NTMprofiler	28.04.2026 13:08	Dateiordner
 Plasmidspades	27.04.2026 06:51	Dateiordner
 Platon	27.04.2026 07:58	Dateiordner

NTMSEQ: AMRFINDER



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Name	Protein id	Contig id	Start	Stop	Strand	Gene sym	Sequence Scope	Element t	Element s	Class	Subclass	Method	Target len	Reference %	Coverage %	Identity	Alignment	Accession	Name of c	HMM id	HMM desc	
2	H	NA	contig0001	378122	378628	+	erm(37)	23S rRNA i core	AMR	AMR	LINCOSAM	AZITHRON	BLASTX	169	179	92.74	59.76	169	WP_0039C	23S rRNA i	NA	NA	
3	H	NA	contig0001	474775	475521	+	aph(3'')	aminoglyc core	AMR	AMR	AMINOGL	STREPTON	BLASTX	249	272	92.65	57.71	253	WP_01237	aminoglyc	NA	NA	
4	H	NA	contig0001	555232	557328	-	katP	catalase/ç plus	VIRULENC	VIRULENC	NA	NA	BLASTX	699	736	95.24	55.95	706	BAA31832	catalase/ç	NA	NA	
5	H	NA	contig0001	645140	646153	+	acr3	arsenite e plus	STRESS	METAL	ARSENIC	ARSENITE	BLASTX	338	346	97.69	53.85	338	CAB14520	arsenite e	NA	NA	
6	H	NA	contig0001	963160	964026	+	blaMAB	MAB fami core	AMR	AMR	BETA-LAC	CARBAPE	EXACTX	289	289	100	100	289	WP_00505	MAB fami	NA	NA	
7	H	NA	contig0001	1118821	1119951	-	cmrA	chlorampl core	AMR	AMR	PHENICOL	CHLORAM	BLASTX	377	391	96.42	55.97	377	WP_06384	chlorampl	NA	NA	
8	H	NA	contig0001	1125306	1125941	+	catA	type A-5 c core	AMR	AMR	PHENICOL	CHLORAM	BLASTX	212	220	95.91	59.43	212	WP_06384	type A-5 c	NA	NA	
9	H	NA	contig0001	4096	6069	+	mdt	Rv1877 fai plus	AMR	AMR	NA	NA	BLASTX	658	687	95.78	52.64	663	CNC42950	Rv1877 fai	NA	NA	
10	H	NA	contig0001	385991	386404	-	arr	NAD(+)-r core	AMR	AMR	RIFAMYCI	RIFAMYCI	BLASTX	138	143	96.5	67.39	138	WP_06384	NAD(+)-r	NA	NA	
11	H	NA	contig0001	57448	58719	-	abaF	fosfomyci plus	AMR	AMR	FOSFOMY	FOSFOMY	BLASTX	424	429	98.14	53.54	424	ABO11759	fosfomyci	NA	NA	
12	H	NA	contig0001	139029	139550	-	aac(2'')	aminoglyc core	AMR	AMR	AMINOGL	GENTAMI	BLASTX	174	195	92.31	62.22	180	WP_00388	aminoglyc	NA	NA	
13	I	NA	contig0001	378185	378691	+	erm(37)	23S rRNA i core	AMR	AMR	LINCOSAM	AZITHRON	BLASTX	169	179	92.74	59.76	169	WP_0039C	23S rRNA i	NA	NA	
14	I	NA	contig0001	474838	475584	+	aph(3'')	aminoglyc core	AMR	AMR	AMINOGL	STREPTON	BLASTX	249	272	92.65	57.71	253	WP_01237	aminoglyc	NA	NA	
15	I	NA	contig0001	555295	557391	-	katP	catalase/ç plus	VIRULENC	VIRULENC	NA	NA	BLASTX	699	736	95.24	55.95	706	BAA31832	catalase/ç	NA	NA	
16	I	NA	contig0001	645203	646216	+	acr3	arsenite e plus	STRESS	METAL	ARSENIC	ARSENITE	BLASTX	338	346	97.69	53.85	338	CAB14520	arsenite e	NA	NA	
17	I	NA	contig0001	963223	964089	+	blaMAB	MAB fami core	AMR	AMR	BETA-LAC	CARBAPE	EXACTX	289	289	100	100	289	WP_00505	MAB fami	NA	NA	
18	I	NA	contig0001	1118884	1120014	-	cmrA	chlorampl core	AMR	AMR	PHENICOL	CHLORAM	BLASTX	377	391	96.42	55.97	377	WP_06384	chlorampl	NA	NA	
19	I	NA	contig0001	1125369	1126004	+	catA	type A-5 c core	AMR	AMR	PHENICOL	CHLORAM	BLASTX	212	220	95.91	59.43	212	WP_06384	type A-5 c	NA	NA	

- aac and aph → do not cause phenotypic resistance against amikacin!!
- blaMAB = hydrolyzes penicillins and cephalosporins, although penems (e.g. imipenem) and cefoxitin appear to be more resistant
- Synergistic effect of beta-lactam inhibitors like avibactam or durlobactam (shin 2024)
- erm(37) = erm (41) → with default values only reported when full-length (not truncated)
- Arr: causes resistance against rifampicin
- Virulence and stress genes reported but not well validated and not complete

- Quality control of reads
- Trimming
- Downsampling
- Contamination detection
- (Sub)species determination
- Strain typing (MLST)
- Assembly
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- Resistance prediction
- **Virulence prediction**
- Plasmid prediction

☐ Name	Änderungsdatum	Typ
Abricate	27.04.2026 07:59	Dateiordner
AMRfinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

NTMSEQ: VIRULENCE (ABRICATE)



Abriicate

Exploratory, clinical significance remains unclear

info.txt

O.tab

Summary_abriicate.tab

N.tab

M.tab

L.tab

K.tab

J.tab

I.tab

H.tab

#FILE	NUM_FOUND	PE18	PE19	PE5	Rv1796	Rv1794	Rv2952	aftB	aftD	ahpC
H	4	.	.	.	.	.	.	.	.	.
I	4	.	.	.	.	.	.	.	.	.
J	4	.	.	.	.	.	.	.	.	.
K	4	.	.	.	.	.	.	.	.	.
L	43	.	100	.	.	100	.	97.08	98.33	99.66
M	4	.	.	.	.	.	.	.	.	.
N	5	.	.	.	.	.	.	.	.	95.75
O	53	100	100	100	80.39	100	98.03	99.2	.	99.83

phoP	pkS1	pkS15	plcC	ppsE	relA	sigA/rpoV	sigE	sigF	sigH	sodA	whiB3
91.13	.	.	.	.	95.36	.	.	.	.	.	.
91.13	.	.	.	.	95.36	.	.	.	.	.	.
91.13	.	.	.	.	95.36	.	.	.	.	.	.
91.13	.	.	.	.	95.36	.	.	.	.	.	.
93.82	.	.	.	.	96.67	93.64	90.05	100	91.71	100	95.47
91.13	.	.	.	.	95.36	.	.	.	.	.	.
91.13	.	.	.	.	95.36	.	.	.	.	.	.
98.25	98.06	94.63	99.8	98.99	96.84	.	93.28	99.87	100	100	98.71

eccA1	eccA3	eccA5	eccC3	eccC5	eccCa1	eccCb1	eccD5	embC	espR	esxA	esxB	esxG	esxH	esxM	esxN
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	94.25	98.25	98.17	99.33	.	.	.	97.87	95.99	.	.	.	100	100	100
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
99.94	99.26	100	99.5	99.52	99.73	99.94	99.87	98.17	99.75	99.65	98.20	100.00	100.00	99.60	97.64

VFDB virulence factors of Pathogenic Bacteria

Home Search Category Compounds Status Download Contacts

JS interface

Acinetobacter
Aeromonas
Anaplasma
Bacillus
Bacteroides
Bartonella
Bordetella
Borrelia
Brucella
Burkholderia
Campylobacter
Chlamydia
Clostridium
Corynebacterium
Coxiella
Enterobacter
Enterococcus
Escherichia
Francisella
Haemophilus
Helicobacter
Klebsiella
Legionella
Listeria
Mycobacterium
Mycoplasma
Neisseria
Pseudomonas
Rickettsia
Salmonella
Shigella
Staphylococcus
Streptococcus
Yersinia

About VFDB

The virulence factor database (VFDB) is an integrated and comprehensive online resource for curating various medically significant bacterial pathogens.

The motivation for constructing VFDB was twofold:

- First, to provide in-depth coverage major virulence factors of the best-characterized bacterial pathogens, and cause disease.
- Second, to provide current knowledge of the wide variety of mechanisms used by bacterial pathogens for treatment and prevention of infectious diseases.

Definitions

- A bacterial pathogen is usually defined as any bacterium that has the capacity to cause disease. Its ability to cause disease is determined by its virulence factors.
- Virulence provides a quantitative measure of the pathogenicity or the likelihood of causing disease.
- Virulence factors refer to the properties (i.e., gene products) that enable a microorganism to establish, mediate bacterial attachment, cell surface carbohydrates and proteins that protect a bacterium, and hydrolyze host tissues.

About anti-virulence compounds

The rapid emergence of adaptive mutations and the widespread dissemination of antibiotic resistance essential cellular functions of bacteria, such as cell wall, replication, transcription and translation, which there is an urgent need for novel antibacterial strategies to combat drug-resistant bacterial infections.

Disarming bacterial pathogens by targeting virulence factors is a promising alternative to classic antibiotic resistant strains, and even further diminish the risk of commensal bacteria elimination so as to be less harmful to the host.

To provide valuable reference resources for drug design, repurposing, and target selection, the VFDB has integrated existing knowledge of bacterial VFs. By cross-linking the current knowledge of bacterial VFs with information on innovative and effective antibacterial therapies to combat bacterial infections and address antibiotic resistance.

About VF category

Traditional microbiological studies generally rely on laboratory bacterium isolation and cultivation, which in-depth mining of the panbacterial microbiome data in terms of pathogenesis requires a well-organized database leads to considerable confusion and hampers follow-up panbacterial analyses. Therefore, a well-defined database to better organize and present bacterial VFs in the database, the VFDB proposed an individual simplification of different bacteria are generally independent of each other, although there are some similarities.

<https://www.mgc.ac.cn/VFs/main.htm>

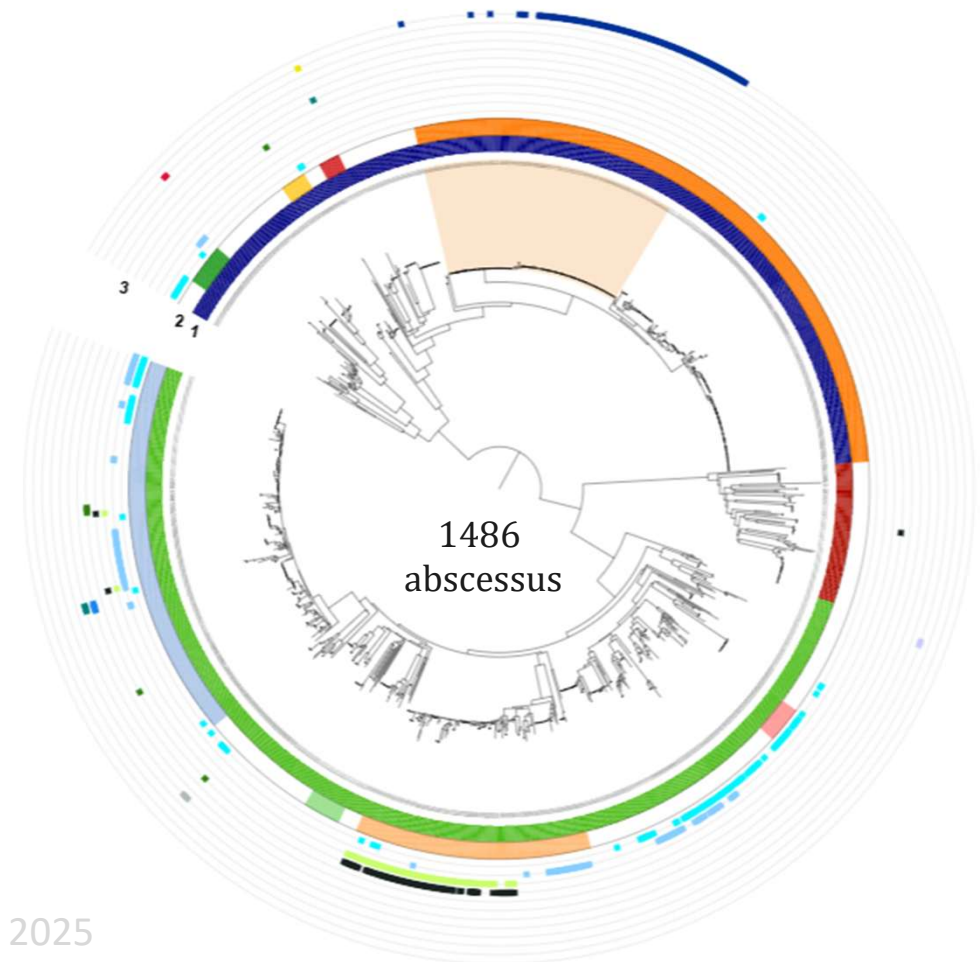
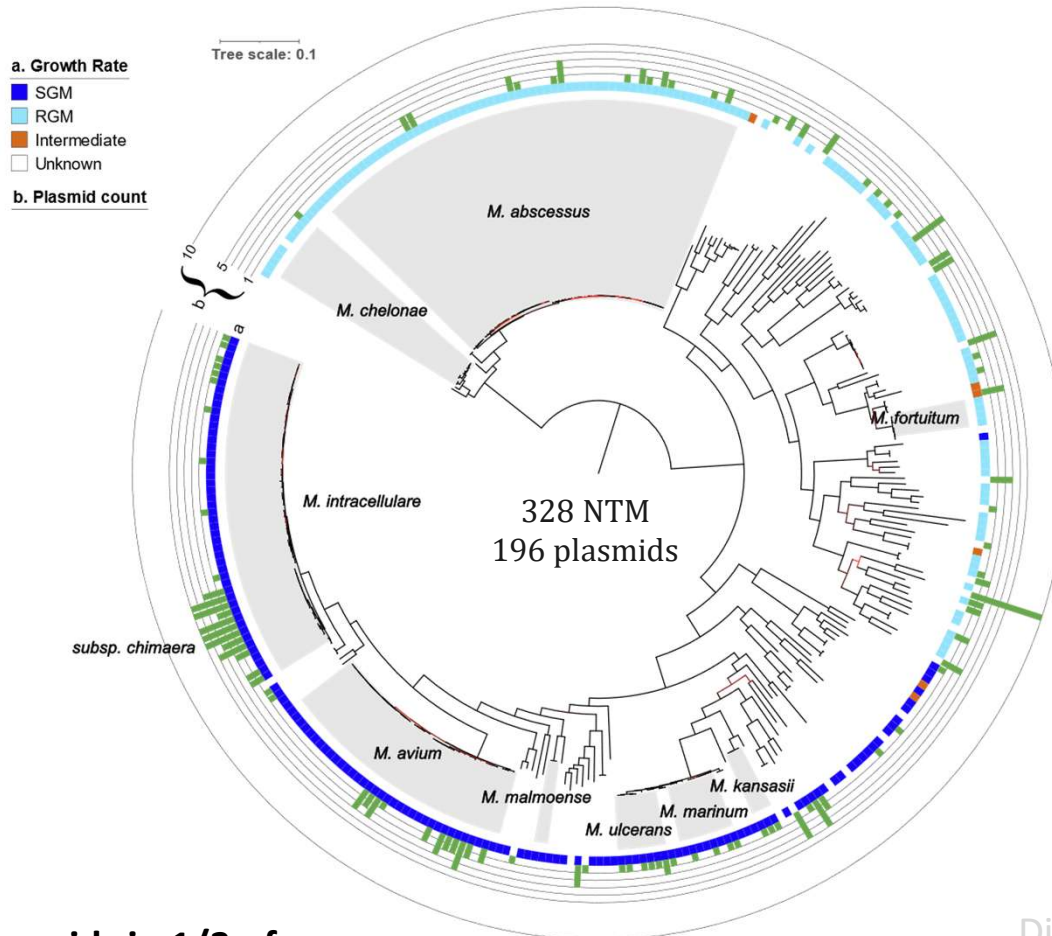
- Quality control of reads
- Trimming
- Downsampling
- Contamination detection
- (Sub)species determination
- Strain typing (MLST)
- Assembly
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- Resistance prediction
- Virulence prediction
- **Plasmid prediction**

☐ Name	Änderungsdatum	Typ
 Abricate	27.04.2026 07:59	Dateiordner
 AMRfinder	27.04.2026 07:59	Dateiordner
 Assemblies	27.04.2026 01:50	Dateiordner
 customDB_SRST2	27.04.2026 07:41	Dateiordner
 fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
 FastP	26.04.2026 23:23	Dateiordner
 FastQ_subsampled	26.04.2026 23:22	Dateiordner
 kraken2	27.04.2026 00:00	Dateiordner
 logs	27.04.2026 07:59	Dateiordner
 Mashtree	27.04.2026 01:54	Dateiordner
 MLST_SRST2	27.04.2026 00:23	Dateiordner
 MultiQC	26.04.2026 23:16	Dateiordner
 NTMprofiler	28.04.2026 13:08	Dateiordner
 Plasmidspades	27.04.2026 06:51	Dateiordner
 Platon	27.04.2026 07:58	Dateiordner

- No plasmids in *M. tuberculosis* or *M. leprae*
- Plasmids common in NTM
 - Virulence plasmid in *M. ulcerans* (pMUM001)
 - Virulence plasmid in *M. avium*
 - Resistance plasmids in *M. abscessus*, *M. marinum* and *M. scrofulaceum*
 - Post-surgical outbreak strains carrying plasmids
 - *M. abscessus* (Brazilian skin/soft tissue infection epidemic)
 - *M. chimaera* (global „heater-cooler unit“ – cardiac surgery outbreak)
- Knowledge about NTM plasmids remains limited



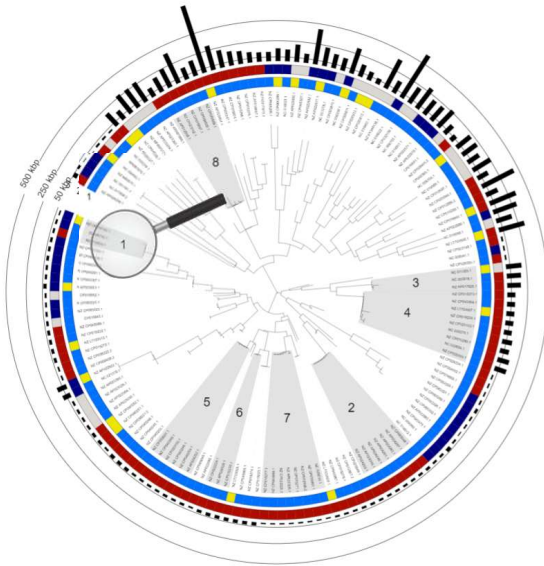
WHICH CLINICALLY RELEVANT NTMS HARBOR PLASMIDS?



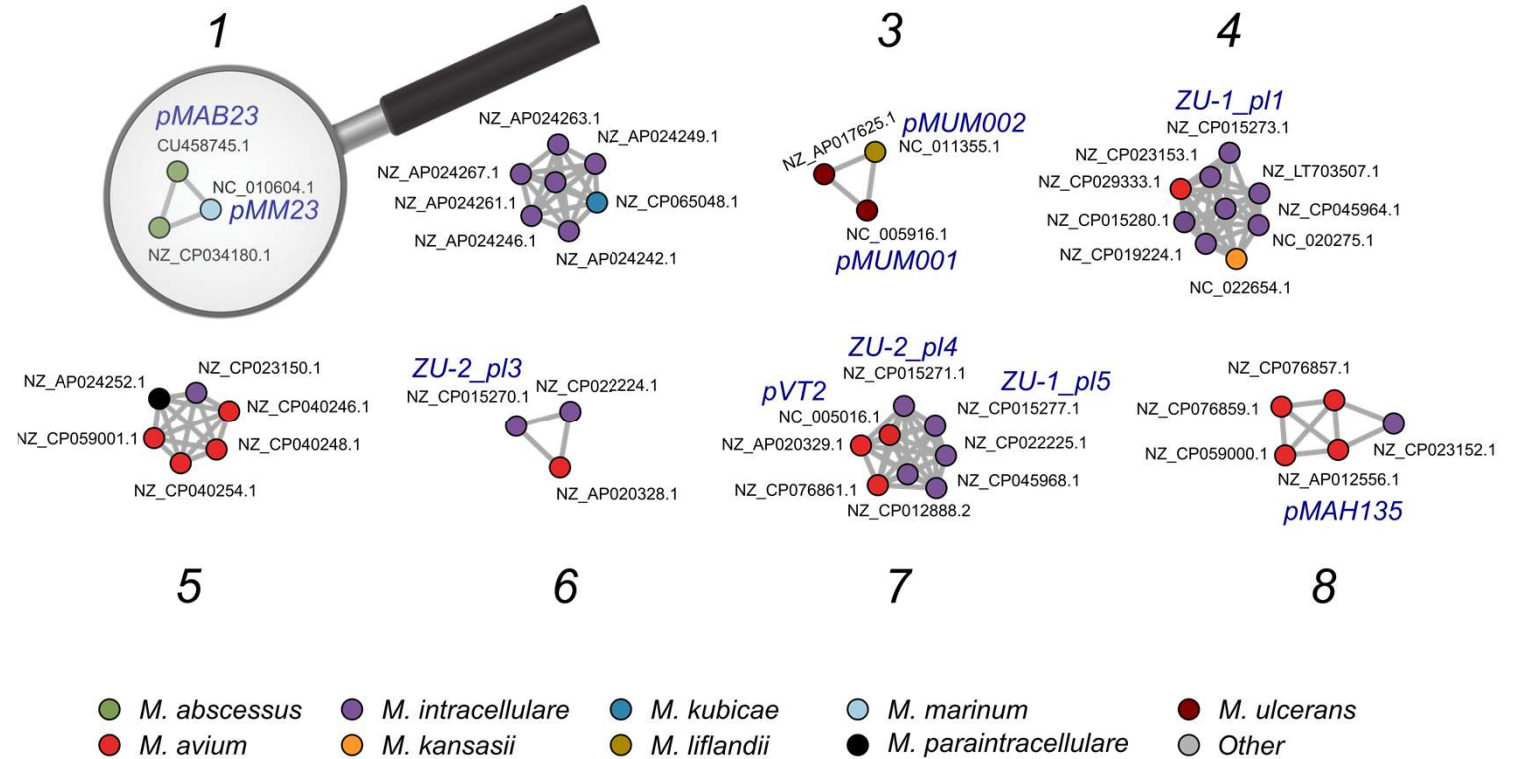
Diricks 2025

Plasmids in 1/3 of genomes

WHAT IS THE GENETIC DIVERSITY OF NTM PLASMIDS?



Length: 10 kbp-400 kbp
Linear and circular

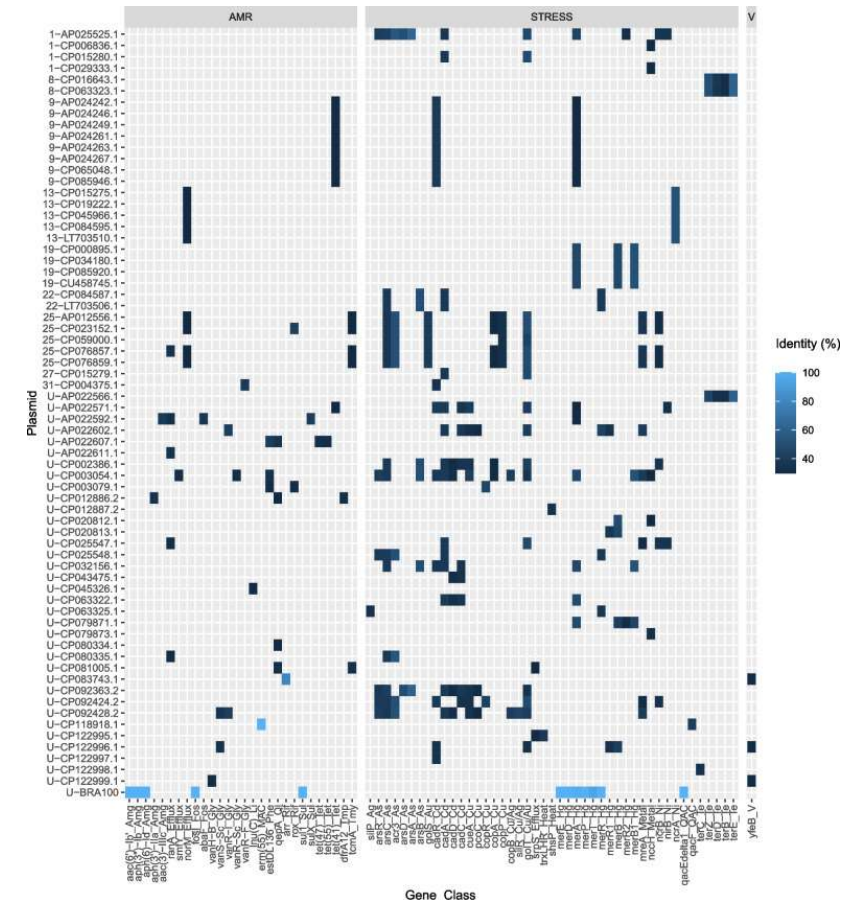


NTM plasmids can move across species barriers!

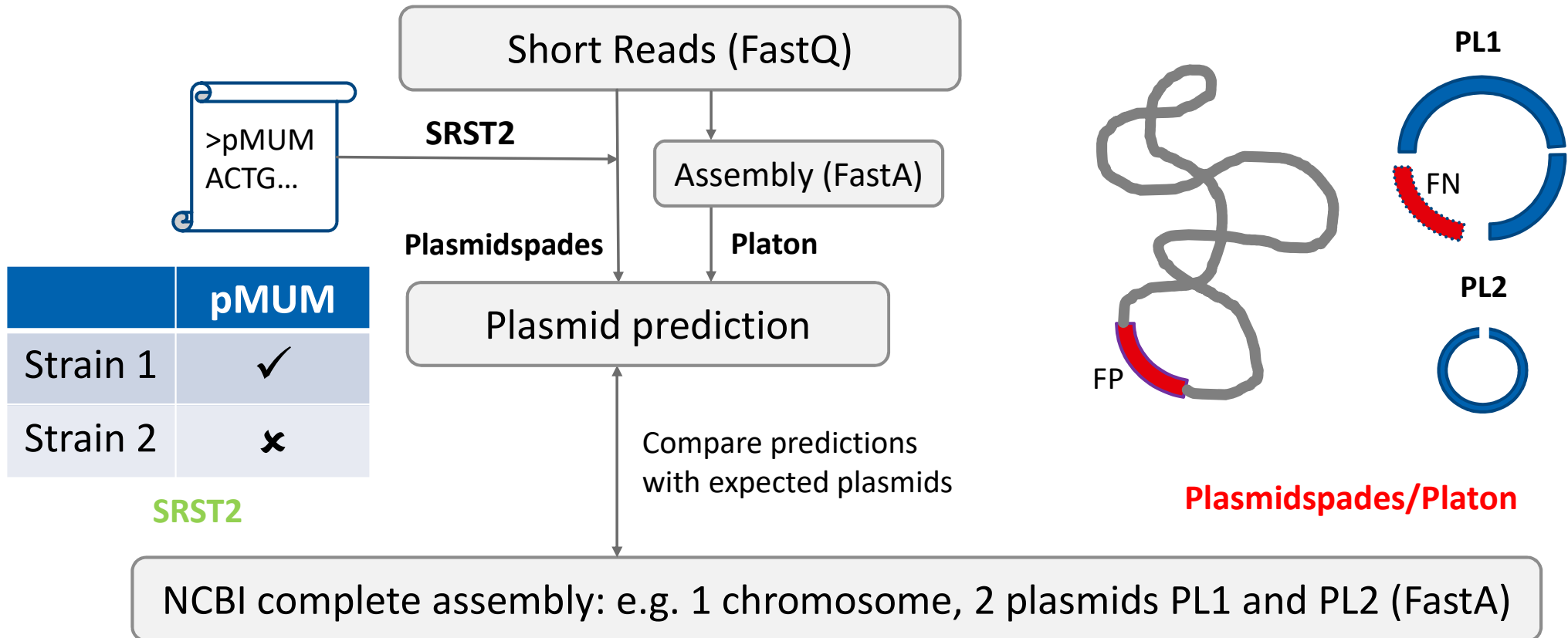
WHICH GENES CAN BE FOUND ON NTM PLASMIDS?



- 17,547 protein-coding genes
 - No single gene present in all NTM plasmids
 - Most genes: hypothetical proteins (8168; 46.5%)
- Core genes
 - **Replication** genes (e.g. helicases *repA*)
 - **Segregation** genes (e.g. *parA/B*) and putative **toxin/antitoxin** genes
 - **Mobility** (e.g. relaxase *recD2*) and **conjugation** genes (e.g. T4SS and T7SS)
- Accessory genes
 - **Antimicrobial resistance** genes
 - **Stress resistance** genes
 - **Virulence** genes



CAN WE PREDICT PLASMIDS FROM SHORT READ NGS DATA



- Quality control of reads
- Trimming
- Downsampling
- Contamination detection
- (Sub)species determination
- Strain typing (MLST)
- Assembly
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- Resistance prediction
- Virulence prediction
- **Plasmid prediction**

☐ Name	Änderungsdatum	Typ
Abriicate	27.04.2026 07:59	Dateiordner
AMRfinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

The screenshot shows the NTMseq web interface. At the top, the breadcrumb navigation is 'NTMseq / db /' followed by a folder icon. Below this is a search bar containing 'mdiricks' and the text 'Delete db/test.txt'. The main area is a file browser with a header 'Name'. It lists three items: a folder icon followed by '..', a document icon followed by '2023_11_03_v2_PLSDb_mycobacteriaceae_208plasmids.fasta', and a document icon followed by 'Plasmid_metadata_208.xlsx'. Two blue arrows point to the document icons of the two files.

	A	B	C	D	E	F	G
1	NUCCORE_AC	Strain ID/plasmid name	Species	Mash_distance group 0.0	NUCCORE_CreatedDate	UCCOR	BIOSAMPLE_Host
2	NZ_OY726396.1	pMu0050	M. sp._MU0050	98	02.09.2023	0.646038	
3	NZ_M60875.1	pAL5000	M. fortuitum	74	02.12.2020	0.648621	Homo sapiens
4	NZ_CP126903.1	unnamed2	M. vanbaalenii	118	21.09.2023	0.654664	
5	NZ_CP126902.1	unnamed1	M. vanbaalenii	57	21.09.2023	0.672224	
6	NZ_AP022578.1	pJCM15296	M. aubagnense	24	02.03.2020	0.648804	
7	NZ_CP097321.1	pJCM15296	Candidatus_M. methanotrophicum	117	23.05.2022	0.655335	
8	NZ_CP076860.1	pS2b	M. avium	75	13.07.2021	0.642545	Homo sapiens
9	NZ_CP076859.1	pS2a	M. avium	19	13.07.2021	0.664962	Homo sapiens
10	NZ_CP076861.1	pS2c	M. avium	8	13.07.2021	0.663639	Homo sapiens
11	NZ_CP081005.1	p20-157661	M. sp._IDR2000157661	43	20.02.2022	0.650887	Homo sapiens
12	NZ_AP025528.1	p20KCMC460_04	M. sp._20KCMC460	97	12.02.2022	0.665006	Homo sapiens
13	NZ_AP025527.1	p20KCMC460_03	M. sp._20KCMC460	62	12.02.2022	0.650806	Homo sapiens
14	NZ_AP025526.1	p20KCMC460_02	M. sp._20KCMC460	9	12.02.2022	0.646194	Homo sapiens
15	NZ_AP025525.1	p20KCMC460_01	M. sp._20KCMC460	2	12.02.2022	0.656573	Homo sapiens
16	NZ_CP040246.1	pFLAC0181_CHOP101	M. avium	36	23.05.2019	0.653293	Homo sapiens
17	NZ_CP040245.1	pMARIA	M. avium	34	23.05.2019	0.64551	Homo sapiens
18	NZ_CP040249.1	pMARIA	M. avium	34	23.05.2019	0.64545	Homo sapiens
19	NZ_CP040248.1	pFLAC0181_CHOP101_1	M. avium	36	23.05.2019	0.651739	Homo sapiens
20	NZ_CP040254.1	pFLAC0181_CHOP101_2	M. avium	36	23.05.2019		
21	NZ_CP040253.1	pMARIA_2	M. avium	34	23.05.2019		
22	NZ_CP040252.1	unnamed2	M. avium	41	23.05.2019		
23	NZ_CP040251.1	unnamed1	M. avium	41	23.05.2019		
24	NZ_KR997898.1	pMA100	M. avium	123	03.06.2021		
25	NC_005916.1	pMUM001	M. ulcerans	7	05.02.2004		



NTMSEQ: PLASMID PREDICTION FROM READS (SRST2)



☒

customDB_SRST2

info.txt

20260426_2311_custom-SRST2_plasmid_presence_itol.txt

customDB_SRST2_20260426_2311_compiledResults.txt

Plasmid_hits.txt

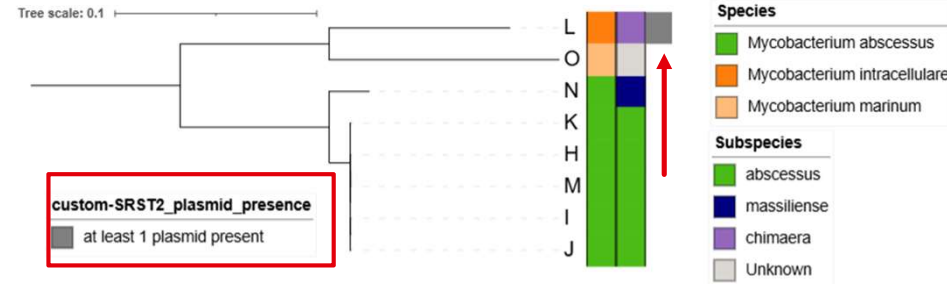
L.log

L_fullgenes_2023_11_03_v2_PLSDb_mycobacteriaceae_208...

L_genes_2023_11_03_v2_PLSDb_mycobacteriaceae_208pla...

J.log

J_genes_2023_11_03_v2_PLSDb_mycobacteriaceae_208pla...



1	Mash group	8	8	2	23	1	27	8	1	23	2
2	Plasmid ID	p3-JPH1	unnamed3	ZU-1 pl1	ZU-1 pl2	ZU-1 pl3	ZU-1 pl4	ZU-1 pl5	pDHQP2015227	pDHQP20	pDHQP20
3	Sample	NZ_AP020329.1	NZ_CP012888.2	NZ_CP015273.1	NZ_CP015274.1	NZ_CP015275.1	NZ_CP015276.1	NZ_CP015277.1	NZ_CP019222.1	NZ_CP019	NZ_CP019
4	L_libx	NZ_AP020329.1*	NZ_CP012888.2*	NZ_CP015273.1*	NZ_CP015274.1*	NZ_CP015275.1*	NZ_CP015276.1*	NZ_CP015277.1	NZ_CP019222.1	NZ_CP019	NZ_CP019

File only reports samples for which at least one known plasmid was found.

Tutorial: only *M. chimaera*

*mismatches

Retain only one per mash group to find unique plasmids

	A	B	C	D	E	F
1	Mash group	2	23	1	27	8
2	Plasmid ID	ZU-1 pl1	ZU-1 pl2	ZU-1 pl3	ZU-1 pl4	ZU-1 pl5
3	Sample	NZ_CP015273.1	NZ_CP015274.1	NZ_CP015275.1	NZ_CP015276.1	NZ_CP015277.1
4	L_libx	NZ_CP015273.1*	NZ_CP015274.1*	NZ_CP015275.1*	NZ_CP015276.1*	NZ_CP015277.1

Only predicts known plasmids included in the custom database! Not able to detect new plasmids! Not able to determine location (sequence might be integrated in the chromosome)

NTMseq / db /

mdiricks Delete db/test.txt

Name

..

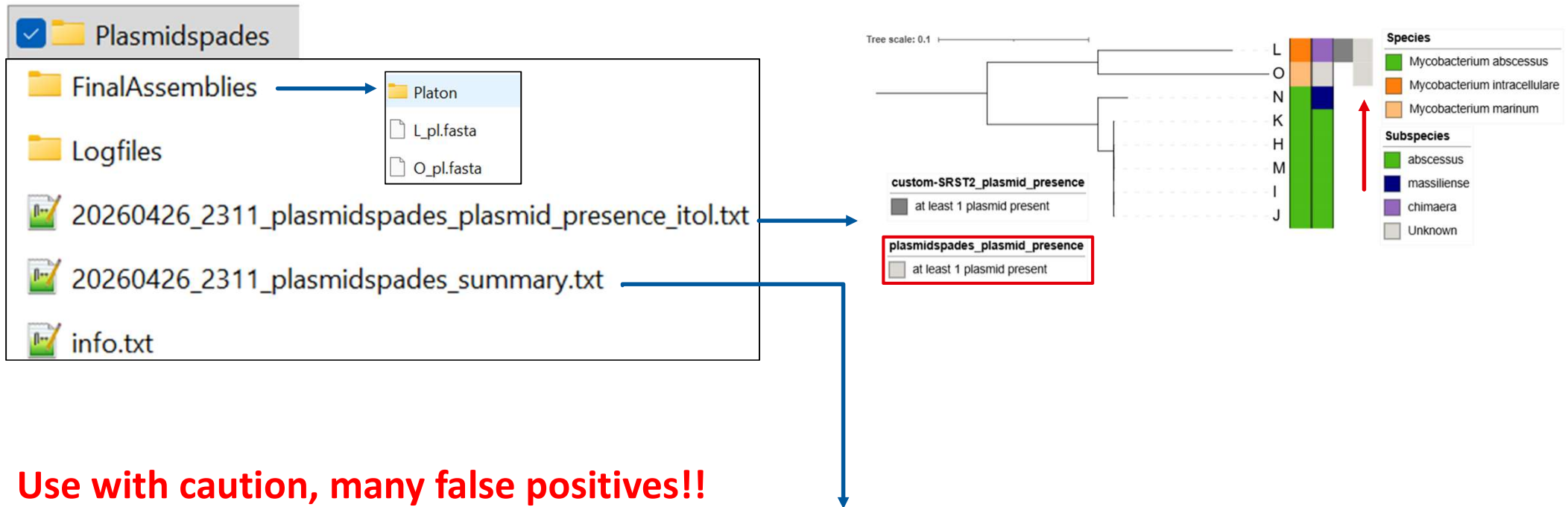
2023_11_03_v2_PLSDb_mycobacteriaceae_208plasmids.fasta

Plasmid_metadata_208.xlsx

- Quality control of reads
- Trimming
- Downsampling
- Contamination detection
- (Sub)species determination
- Strain typing (MLST)
- Assembly
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- Resistance prediction
- Virulence prediction
- **Plasmid prediction**

☐ Name	Änderungsdatum	Typ
Abriicate	27.04.2026 07:59	Dateiordner
AMRfinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

NTMSEQ: DE NOVO PLASMID PREDICTION FROM READS (PLASMIDSPADES)



Use with caution, many false positives!!

	A	B	C
	Sample	#Putative Mobile Elements	Mobile Elements (node_length-bp_cov)
1	L	12	1_65095_52.677564;2_32264_91.505772;3_21250_70.088955;4_19014_28.450733;5_13585_114.171051;6_13356_25.562628;7_4615_30.600267;8_4478_37.656401;9_1316_6.267452;10_1170_46.511026;11_624_20.941650;12_552_20.604706;
2	O	177	1_7856_16.889378;2_6339_17.737766;3_6288_34.134556;4_5323_14.931486;5_3624_21.381184;6_3079_18.642615;7_3049_12.204312;8_3028_18.768356;9_1847_14.024419;10_1803_16.485084;11_1779_24.377119;12_1627_19.277333;13_1413_

- Quality control of reads
- Trimming
- Downsampling
- Contamination detection
- (Sub)species determination
- Strain typing (MLST)
- Assembly
- Tree (Dendrogram)
- Genome similarity estimation (ANI)
- Resistance prediction
- Virulence prediction
- **Plasmid prediction**

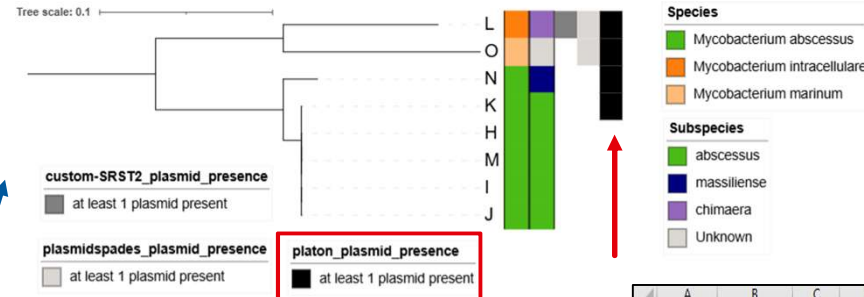
☐ Name	Änderungsdatum	Typ
 Abricate	27.04.2026 07:59	Dateiordner
 AMRfinder	27.04.2026 07:59	Dateiordner
 Assemblies	27.04.2026 01:50	Dateiordner
 customDB_SRST2	27.04.2026 07:41	Dateiordner
 fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
 FastP	26.04.2026 23:23	Dateiordner
 FastQ_subsampled	26.04.2026 23:22	Dateiordner
 kraken2	27.04.2026 00:00	Dateiordner
 logs	27.04.2026 07:59	Dateiordner
 Mashtree	27.04.2026 01:54	Dateiordner
 MLST_SRST2	27.04.2026 00:23	Dateiordner
 MultiQC	26.04.2026 23:16	Dateiordner
 NTMprofiler	28.04.2026 13:08	Dateiordner
 Plasmidspades	27.04.2026 06:51	Dateiordner
 Platon	27.04.2026 07:58	Dateiordner

NTMSEQ: PLASMID PREDICTION (PLATON)



Platon

- FinalAssemblies
 - K.plasmid.fasta
 - L.plasmid.fasta
 - N.plasmid.fasta
 - O.plasmid.fasta
- H
- I
- J
- K
- L
- M
- N
- O
- 20260426_2311_platon_plasmid_presence_itol.txt
- info.txt
- mobGenes.tsv
- Platon_Output_RefSeqHits_20260426_2311.tsv
- Platon_Output_summary_20260426_2311.tsv
- Platon_Output_summary2_20260426_2311.tsv



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Sample	ID	Length	Coverage	# ORFs	RDS	Circular	Inc Type(s)	Replicatio	Mobilizatic	# OriT	Conjugatic	# AMRs	# rRNAs	Plasmid H
2	K	contig00009	241197	NA	226	-7.5	no	0	1	0	0	0	0	0	0
3	L	contig00018	87562	NA	70	-6.6	no	0	0	1	0	1	0	0	0
4	L	contig00061	32137	NA	29	0.6	no	0	0	1	0	0	0	0	1
5	L	contig00062	31223	NA	36	1.8	no	0	0	0	0	0	0	0	1
6	L	contig00076	24483	NA	15	0.9	no	0	0	1	0	0	0	0	1
7	L	contig00089	21123	NA	22	0.1	no	0	0	1	0	0	0	0	1
8	L	contig00092	19189	NA	22	1.7	no	0	0	0	0	0	0	0	1
9	L	contig00104	15421	NA	22	1.8	no	0	0	0	0	0	0	0	1
10	L	contig00106	15102	NA	17	-2.2	no	0	1	0	0	0	0	0	0
11	L	contig00111	13996	NA	11	-1.9	no	0	1	0	0	0	0	0	0
12	L	contig00113	13767	NA	15	0.8	no	0	0	0	0	0	0	0	1
13	L	contig00114	13371	NA	18	1	no	0	0	1	0	0	0	0	1
14	L	contig00148	5517	NA	5	1.3	no	0	0	0	0	0	0	0	1
15	L	contig00156	4522	NA	4	1.6	no	0	0	0	0	0	0	0	1
16	N	contig00017	94633	NA	76	-7.2	no	0	1	0	0	0	0	0	0
17	O	contig00002	280357	NA	245	-2.6	no	0	1	0	0	0	0	0	0

	A	B
1	Sample	#putative mobile elements (platon)
2	H	0
3	I	0
4	J	0
5	K	1
6	L	13
7	M	0
8	N	1
9	O	1

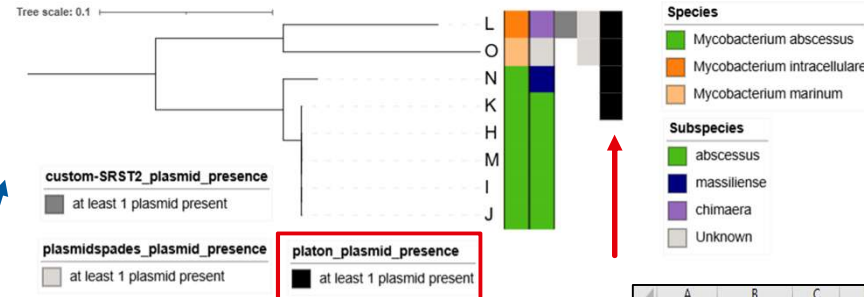
Use with caution, many false positives!!

NTMSEQ: PLASMID PREDICTION (PLATON)



Platon

- FinalAssemblies
 - K.plasmid.fasta
 - L.plasmid.fasta
 - N.plasmid.fasta
 - O.plasmid.fasta
- H
- I
- J
- K
- L
- M
- N
- O
- 20260426_2311_platon_plasmid_presence_itol.txt
- info.txt
- mobGenes.tsv
- Platon_Output_RefSeqHits_20260426_2311.tsv
- Platon_Output_summary_20260426_2311.tsv
- Platon_Output_summary2_20260426_2311.tsv



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Sample	ID	Length	Coverage	# ORFs	RDS	Circular	Inc Type(s)	Replicatio	Mobilizatic	#Ort	Conjugatic	#AMRs	#rRNAs	Plasmid H
2	K	contig00009	241197	NA	226	-7.5	no	0	1	0	0	0	0	0	0
3	L	contig00018	87562	NA	70	-6.6	no	0	0	1	0	1	0	0	0
4	L	contig00061	32137	NA	29	0.6	no	0	0	1	0	0	0	0	1
5	L	contig00062	31223	NA	36	1.8	no	0	0	0	0	0	0	0	1
6	L	contig00076	24483	NA	15	0.9	no	0	0	1	0	0	0	0	1
7	L	contig00089	21123	NA	22	0.1	no	0	0	1	0	0	0	0	1
8	L	contig00092	19189	NA	22	1.7	no	0	0	0	0	0	0	0	1
9	L	contig00104	15421	NA	22	1.8	no	0	0	0	0	0	0	0	1
10	L	contig00106	15102	NA	17	-2.2	no	0	1	0	0	0	0	0	0
11	L	contig00111	13996	NA	11	-1.9	no	0	1	0	0	0	0	0	0
12	L	contig00113	13767	NA	15	0.8	no	0	0	0	0	0	0	0	1
13	L	contig00114	13371	NA	18	1	no	0	0	1	0	0	0	0	1
14	L	contig00148	5517	NA	5	1.3	no	0	0	0	0	0	0	0	1
15	L	contig00156	4522	NA	4	1.6	no	0	0	0	0	0	0	0	1
16	N	contig00017	94633	NA	76	-7.2	no	0	1	0	0	0	0	0	0
17	O	contig00002	280357	NA	245	-2.6	no	0	1	0	0	0	0	0	0

	A	B
1	Sample	#putative mobile elements (platon)
2	H	0
3	I	0
4	J	0
5	K	1
6	L	13
7	M	0
8	N	1
9	O	1

Use with caution, many false positives!!

	A	B	C	D
1	id	rep	mob_genes	conj_genes
2	H			
3	I			
4	J			
5	K	replication_PCLA_725460		
6	L	replication_PCLA_725460,replication_PCLA_725460	profile_MOBF,profile_MOBF,profile_MOBF,profile_MOBF,profile_MOBF	conjugation_PCLA_2783275
7	M			
8	N	replication_PCLA_725460		
9	O	replication_PCLA_3267610		

- From short reads
 - Works well if you know what you are looking for (e.g. virulence plasmid)
 - De novo prediction (e.g. with plasmidspades/platon) → many false positives, sometimes false negatives
- Long read sequencing (e.g. Pacbio/Nanopore) much better suited to resolve plasmid sequences!
- Wet-lab validation (e.g. plasmid extraction, junction PCR to check circularity,...)

NTMSEQ: TROUBLESHOOTING



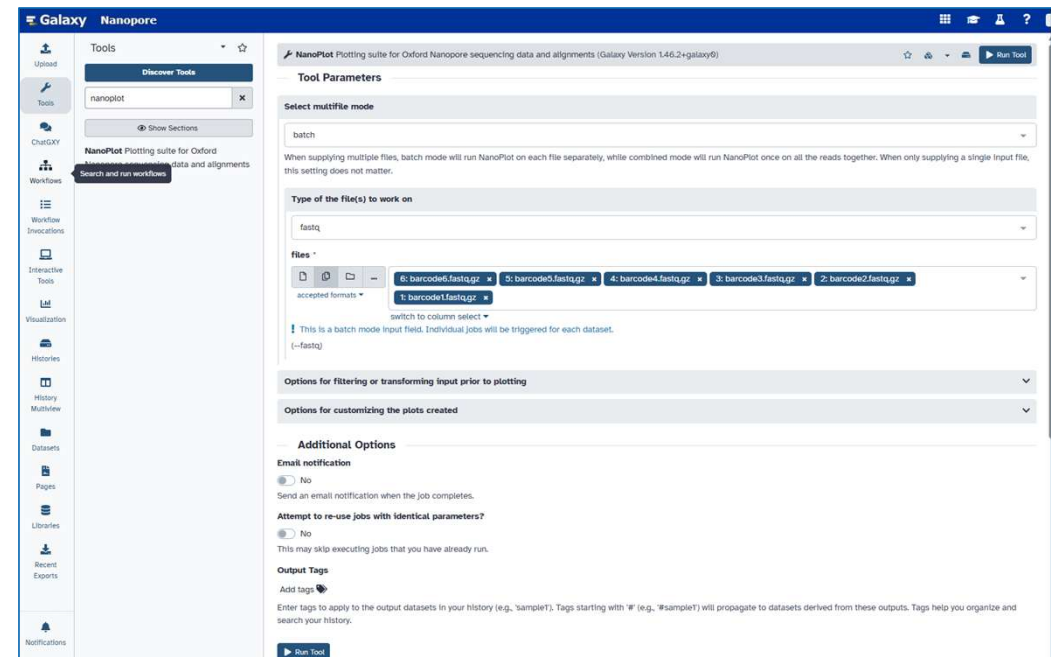
☐ Name	Änderungsdatum	Typ
Abriicate	27.04.2026 07:59	Dateiordner
AMRFinder	27.04.2026 07:59	Dateiordner
Assemblies	27.04.2026 01:50	Dateiordner
customDB_SRST2	27.04.2026 07:41	Dateiordner
fastANI_AllagainstAll	27.04.2026 01:54	Dateiordner
FastP	26.04.2026 23:23	Dateiordner
FastQ_subsampled	26.04.2026 23:22	Dateiordner
kraken2	27.04.2026 00:00	Dateiordner
logs	27.04.2026 07:59	Dateiordner
Mashtree	27.04.2026 01:54	Dateiordner
MLST_SRST2	27.04.2026 00:23	Dateiordner
MultiQC	26.04.2026 23:16	Dateiordner
NTMprofiler	28.04.2026 13:08	Dateiordner
Plasmidspades	27.04.2026 06:51	Dateiordner
Platon	27.04.2026 07:58	Dateiordner

☒ logs
abricate.log
AMRFinderPlus.log
assembly_shovill.log
Contamination_detection_kraken2.log
fastANI_all_against_all.log
mash_on_FastA_files.log
mash_on_FastQ_files.log
MLST_on_FastQ_files_SRST2.log
NTM_species_and_resistance_detection_NTM...
parse_kraken2_results.log
plasmid_assembly_using_plasmidspades.log
plasmid_detection_with_platon.log
plasmid_detection_with_SRST2_using_custo...
platon_on_plasmidspades_assemblies.log
preprocess_reads_fastp.log
Quality_control_on_reads_fastQCmultiQC.log

TOOLS FOR NANOPORE DATA



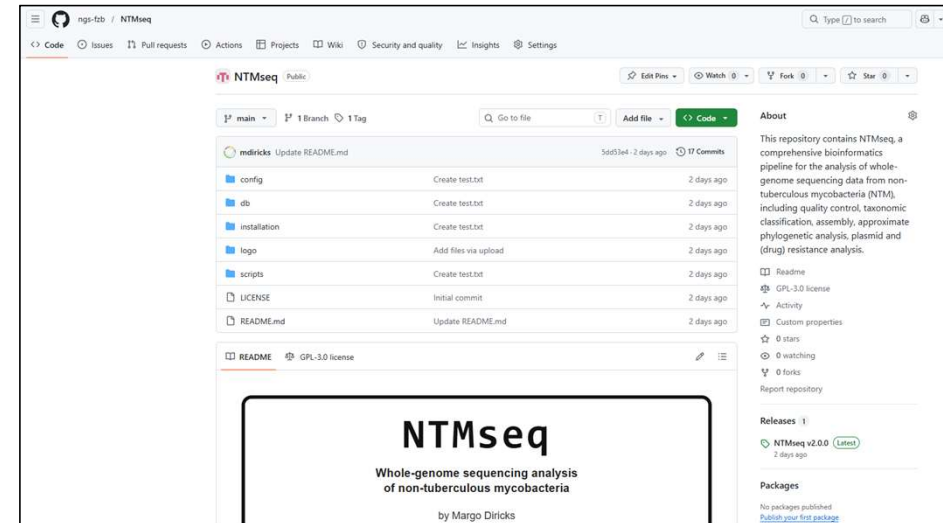
- Nanoplot for quality control
- Autocycler (slow, highly accurate) or Dragonfly (fast) for assembly → use output as input for NTMseq
- NTMprofiler (works on nanopore fastQ files or assemblies)
- Nanogalaxy



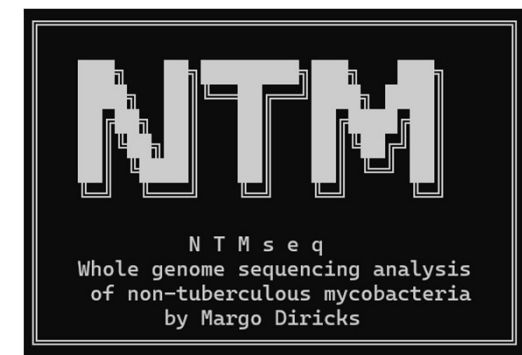
INSTALLATION AND RUNNING NTMSEQ



- <https://github.com/ngs-fzb/NTMseq>
- Optimized for short-read sequencing data (Illumina)
- Requirements
 - Linux-based system (e.g. VM, wsl,...)
 - Basic command-line experience
 - Conda (e.g. miniconda) or mamba for faster installation
- Steps
 - Connect to a linux-based system (e.g. to a VM via putty)
 - Download repository (wget)
 - Install required tools: `$ bash Installation_NTMseq.sh`
 - Install databases: instructions at github
 - Configure your analyses: in NTMseq.config
 - Run analyses: `$ bash starter_NTMseq.sh ../config/NTMseq.config`



NTMseq is for Research Use Only!



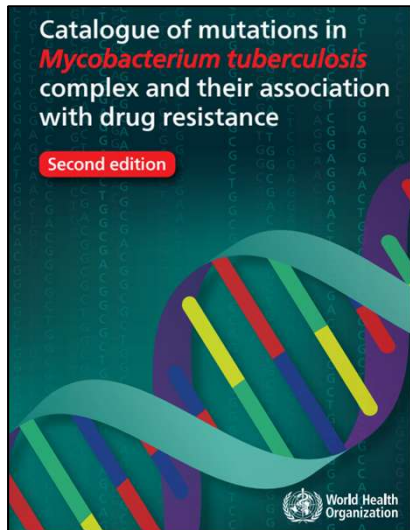
CONCLUSIONS



- NTM are an emerging public health threat
- Accurate (sub)species identification is essential for treatment
- Environmental reservoirs play a key role in transmission → environmental sampling!
- WGS enables high-resolution species identification, outbreak detection and source tracing
- Interpretation requires integration with epidemiological data
- WGS-based tools and mutations catalogues are rapidly developing
 - No one best solution, all tools have advantages/disadvantages → use multiple!
 - Require continuous validation
 - Ask help from NTM experts
- Collaboration and standardisation are critical for future surveillance



RESISTANCE CATALOGUES FOR MYCOBACTERIA



Drug	Variant	MUT Present_pheno S	MUT Absent_pheno S	MUT Present_pheno R	MUT Absent_pheno R	Sensitivity	Specificity
RIF	rpoB_p.Ser450Leu	226	30643	10859	6002	64.4%	99.3%
RIF	rpoB_p.Asp435Val	17	30515	1154	15656	6.9%	99.9%
RIF	rpoB_p.His445Asp	10	30522	608	16202	3.6%	100.0%
RIF	rpoB_p.His445Tyr	11	30521	593	16217	3.5%	100.0%

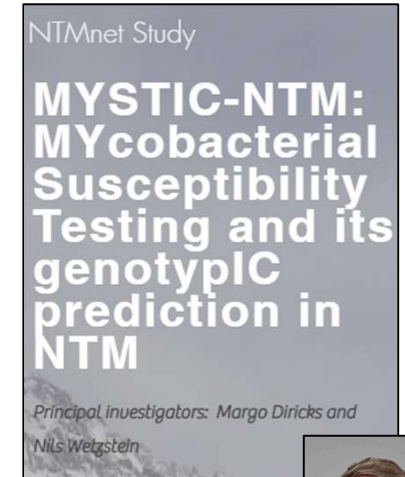
Associated with resistance

Not associated with resistance

MYSTIC NTM



<https://www.ntmnet.org/mystic-ntm>



- Goal 1: to improve and extend current resistance catalogues for NTM
- Goal 2: to provide a benchmark dataset for resistance prediction tools (e.g. NTMprofiler)
- Goal 3: to improve surveillance and monitor trends
- > 17 institutions across 13 countries expressed interest
 - Data sharing (preparations) have already begun for a few of them



pDST



gDST

PubMLST

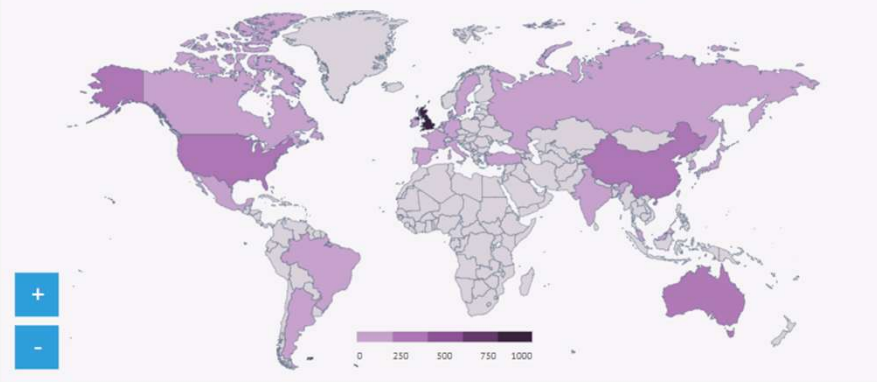
Public databases for molecular typing and microbial genome diversity

[HOME](#)
[ORGANISMS](#)
[SPECIES ID](#)
[ABOUT US](#)
[UPDATES](#)

Home > Organisms

Mycobacteroides abscessus complex

Source of isolates submitted to the *Mycobacteroides abscessus* complex database



[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 [a](#).

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] [a](#).

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: 191,947

Last updated: 2026-04-22

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: 2,233

Last updated: 2026-04-28

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: 2,189

Last updated: 2026-04-22

- Submit your sequences to this database!
- Important for surveillance
- Some tools only available with submitted samples

- Exception: iTol tool

PubMLST

Public databases for molecular typing and microbial genome diversity

[HOME](#)
[ORGANISMS](#)
[SPECIES ID](#)
[ABOUT US](#)
[UPDATES](#)

[Home](#) > [Organisms](#) > [Mycobacteroides abscessus complex](#) > [Mycobacteroides abscessus complex isolates](#) > [Plugins](#) > iTOL

 Logged in
  Help
  Footnotes

iTOL - Interactive Tree of Life - *Mycobacteroides abscessus* complex isolates



This plugin uploads data for analysis within the Interactive Tree of Life online service:

iTOL is developed by: Ivica Letunic (1) and Peer Bork (2,3,4)

1. Biobyte solutions GmbH, Bothestr 142, 69126 Heidelberg, Germany
2. European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany
3. Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany
4. Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany

Web site: <https://itol.embl.de/>

Publication: Letunic & Bork (2021) Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**(W1):W293-6.

This tool will generate neighbor-joining trees from concatenated nucleotide sequences or allelic profiles. For sequence trees, only DNA loci that have a corresponding database containing allele sequence identifiers, or DNA and peptide loci with genome sequences, can be included. Please check the loci that you would like to include. Alternatively select one or more schemes to include all loci that are members of the scheme.

Analysis is limited to 5,000 records (2,000 records for ClustalW).

If generating a tree by aligning sequences then it is also limited to 2,500,000 sequences (records x loci).

Isolates

1813
1881
1882
1883
1884
2072
2073

[Clear](#)
[List all](#)

User genomes

Optionally include data not in the database.

Upload assembly FASTA file (or zip/tar.gz/tar file containing multiple FASTA files - one per genome): 

[Datei auswählen](#) [Keine ausgewählt](#) 

Loci

Select options

[Paste list](#)

Schemes

 All loci
  MLST
  Ribosomal MLST
  cgMLST

Tree generation

Generate trees from:

☐ profiles

☒ sequences

Tree algorithm (sequences):

☒ FastTree (Approximate ML)

☐ ClustalW (NJ)

iTOL datasets

Select to create data overlays 

isolate

iTOL data type

☒ text labels

☐ coloured strips

Action

[SUBMIT](#)