The interpretation algorithm for TB Profiler output was developed by Microbial Diseases Laboratory (MDL) of California Department of Public Health for integration in the routine TB DST workflow and may not be suitable for all laboratories depending on the availability of confirmatory phenotypic testing and expertise in genetics of TB resistance. In the current version, the interpretation algorithm has not been CLIA-validated and is subject to change. For any inquiries please contact MDL Mycobacterial, Mycotic, and Parasitic Diseases Section, Dr. Varvara Kozyreva (varvara.kozyreva@cdph.ca.gov) and Dr. Matthew Sylvester (matthew.sylvester@cdph.ca.gov)

MDL WGS-DST interpretation logic

Last updated: 01/24/2024

In general, if there is an entry in the TBDB database for a mutation, that takes precedence and is applied first. TBDB database includes the default set as well as tbdb.other_annotations.csv. TBDB watchlist is active. Additionally mutations found in CDC's Varpipe database but not in TBDB were added into a custom TBDB; the confidence for the mutations that were present in TBDB but had a different interpretation in CDC Varpipe database was updated to match CDC's predicted phenotype in the final custom TBDB.

Below in the sections 1-4, please see the detailed description of parsing that is performed to generate interpretations for detected mutations, also referred to as "Laboratorian report" and contains broader information than is reported as a part of clinical report (see section 5).

1. For **genes *mmpR (Rv0678), atpE, pepQ, mmpL5, mmpS5, rrl, rplC (genes related to new drugs based on CDC expert rules)***:

1.1. IF mutation is in either of *mmpR, atpE, pepQ, mmpL5, mmpS5, rrl, rplC*
AND who_confidence has value for the corresponding drug,
THEN keep that value as in "confidence" column **AND** assign the following "Looker interpretation" & "MDL Interpretation" based on that confidence value:

Mutation classification (confidence field)	**Looker interpretation**	**MDL Interpretation**
WHO: associated with R [1]	R	R
WHO: associated with R- interim [2]	R-Interim	R
WHO: uncertain [3,4]	U	U
WHO: NOT associated w/R [5]	S	S
WHO: NOT associated w/R- interim [6]	S-Interim	S
WHO: synonymous [7]	S	S

[1] Exact value in WHO & TBDB "1) Assoc w R"
[2] Exact value in WHO & TBDB "2) Assoc w R – Interim"
[3] Exact value in WHO & TBDB "3) Uncertain significance"
[4] WHO also has value "NA" that is not found in TBDB
[5] Exact value in WHO "5) Not assoc w R". Not in TBDB
[6] Exact value in WHO & TBDB "4) Not assoc w R – Interim"
[7] Exact value in WHO "Synonymous". Not in TBDB

1.2. **IF** mutation is in either of *mmpR, atpE, pepQ, mmpL5, mmpS5, rrl, rplC*
AND no who_confidence available for the corresponding drug
THEN assign confidence= "No WHO annotation" **AND** assign the "Looker interpretation" & "MDL Interpretation" value based on table below:

Genes	Mutation location	Mutation type	Looker interpretation	MDL Interpretation
mmpR, atpE, pepQ, rplC	Target Promoter* (see coordinates below)	any	U	U
	Promoter (effect type= upstream_gene_variant, but not within the target promoter region)	any	U	S
	ORF (everything else)	NONsyn	U	U
		syn	S	S
mmpL5	Promoter (effect type= upstream_gene_variant)	any	U	S
	ORF (everything else)	NONsyn	U	U
		syn	S	S
mmpS5	Promoter (effect type= upstream_gene_variant)	any	U	S
	ORF (everything else)	NONsyn	U	U
		syn	S	S
rrl	nt positions 2003-2367 and 2449-3056 in rRNA	any	U	U
	outside of nt positions 2003-2367 and 2449-3056 in rRNA	any	U	S

* Promoter regions for different genes are defined here as:

	Promoter region coordinates
rplC	-18 to -1
mmpR	-84 to -1
atpE	-48 to -1
pepQ	-33 to -1

1.3. Only for genes *mmpR5 (Rv0678), mmpL5, mmpS5*- to standardize[8] the output and interpretation of mutations that may be listed in relation to any of those 3 genes:

1.3.1. Output mutations in "alternative_consequences" field to get information about these mutations in relation to the other three genes
Follow rule 1.2 to assign Looker interpretation and MDL interpretation.

2. For **genes *katG, pncA, ethA, gid, rpoB (covered by WHO expert rules):***

2.1. IF mutation is in either of *katG*, *pncA*, *ethA*, *gid, rpoB*,
AND who_confidence has value for the corresponding drug,
THEN keep that value as in "confidence" column **AND** assign the following "Looker interpretation" & "MDL Interpretation" based on that confidence value:

Mutation classification (confidence field)	**Looker interpretation**	**MDL Interpretation**
WHO: associated with R	R	R
WHO: associated with R- interim	R-Interim	R
WHO: uncertain	U	U
WHO: NOT associated w/R	S	S
WHO: NOT associated w/R- interim	S-Interim	S
WHO: synonymous [9]	S	S

2.2. IF mutation found in either of *katG, pncA, ethA, gid, rpoB,*
AND no who_confidence available for the corresponding drug,
THEN apply WHO Expert rule:

2.2.1. IF mutation found in either of *katG, pncA, ethA, gid (*not rpoB*)*:

2.2.1.1. IF mutation represents a loss-of-function (mutation contains "del", "ins", "fs", "delins", "_" or ends with "*")
AND
Those mutations are found in ORF or within first 30 nucleotides upstream start codon
THEN assign confidence= "No WHO annotation" **AND** assign the "Looker interpretation" & "MDL Interpretation" = "U" for corresponding mutation.
ELSE assign confidence= "No WHO annotation" **AND** assign the "Looker interpretation" & "MDL Interpretation" value based on table below:

[8] This is done in interim until the TBDB is replaced with unmodified WHO database, after which the WHO gene assignment will be used as is, and in absence of WHO data the expert rule 1.3.2 will be followed.

[9] Not the same as synonymous mutations that are not found in WHO and labeled as such.

Mutation classification	Looker interpretation	MDL Interpretation
Detected mutation not in WHO AND **synonymous**	S	S
Detected mutation not in WHO AND **NONsynonymous** [10] OR **Promoter variants with effect type= upstream_gene_variant**	U	S

2.2.2.IF mutation found in *rpoB*:

2.2.2.1. **IF** mutation within *rpoB* codons 426-452 (even partially, if it's indel) **THEN** assign confidence= "No WHO annotation" **AND** assign the "Looker interpretation" & "MDL Interpretation" value based on table below:

Mutation classification	Looker interpretation	MDL Interpretation
Detected mutation not in WHO AND **synonymous**	S	S
Detected mutation not in WHO AND **NONsynonymous** [11]	R	R

2.2.2.2. **IF** mutation outside of codons 426-452 **THEN** assign confidence= "No WHO annotation" **AND** assign the "Looker interpretation" & "MDL Interpretation" value based on table below:

Mutation classification	Looker interpretation	MDL Interpretation
Detected mutation not in WHO and not covered by expert rule; **synonymous**	S	S

[10] Including indels in ORF

[11] Including indels in ORF

Detected mutation not in WHO and not covered by expert rule; **NONsynonymous** [12] OR **Promoter variants with effect type= upstream_gene_variant**	U	S

3. For genes **<u>other than</u>** *mmpR, atpE, pepQ, mmpL5, mmpS5, rrl, rplC, katG*, *pncA*, *ethA*, *gid, rpoB*:

3.1. IF mutation is in gene <u>other</u> than *mmpR, atpE, pepQ, mmpL5, mmpS5, rrl, rplC, katG*, *pncA*, *ethA*, *gid, rpoB*
AND who_confidence has value for the corresponding drug,
THEN keep that value as in "confidence" column
AND assign the following "Looker interpretation" & "MDL Interpretation" based on that confidence value:

Mutation classification (confidence field)	**Looker interpretation**	**MDL Interpretation**
WHO: associated with R	R	R
WHO: associated with R- interim	R-Interim	R
WHO: uncertain	U	U
WHO: NOT associated w/R	S	S
WHO: NOT associated w/R- interim	S-Interim	S
WHO: synonymous [13]	S	S

3.2. IF mutation is in gene <u>other</u> than *mmpR, atpE, pepQ, mmpL5, mmpS5, rrl, rplC, katG*, *pncA*, *ethA*, *gid, rpoB*
AND no who_confidence available for the corresponding drug

3.2.1. IF Mutation is located within *rrs* gene region,

3.2.1.1. IF Mutations at *rrs* nucleotide positions 1401, 1402, or 1484 (positions in relation to NC_000962.3: 1473246, 1473247, or 1473329)
THEN assign:
confidence= "No WHO annotation";
"Looker interpretation" = U
"MDL Interpretation" = U

3.2.1.2. IF Mutations at *rrs* nucleotide positions OTHER THAN 1401, 1402, or 1484 (positions in relation to NC_000962.3: 1473246, 1473247, or 1473329),
THEN assign:

[12] Including indels in ORF
[13] Not the same as synonymous mutations that are not found in WHO and labeled as such.

confidence= "No WHO annotation";
"Looker interpretation" = U
"MDL Interpretation" = S

3.2.2.IF Mutation is located within *gyrA* coding gene (GyrA QRDR expert rule):

3.2.2.1. **AND** mutation is within region codons 88-94 of *gyrA* AND mutation is nonsynonymous
THEN
confidence= "No WHO annotation"
"Looker interpretation" = U
"MDL Interpretation" = U

3.2.3.IF Mutation is located within *gyrB* coding gene (GyrB QRDR expert rule):

3.2.3.1. **AND** mutation is within region codons 446-507 of *gyrB* AND mutation is nonsynonymous
THEN
confidence= "No WHO annotation"
"Looker interpretation" = U
"MDL Interpretation" = U

3.2.4. OTHERWISE (all remaining scenarios not covered above),
THEN assign confidence= "No WHO annotation"
AND
assign the "Looker interpretation" & "MDL Interpretation" value based on table below:

Mutation classification	**Looker interpretation**	**MDL Interpretation**
Detected mutation not in WHO and not covered by expert rule; **synonymous**	S	S
Detected mutation not in WHO and not covered by expert rule; **NONsynonymous** [14] OR **Promoter variants with effect type= upstream_gene_variant**	U	S

4. Reporting of remaining scenarios:

4.1. **IF** No mutations detected in given gene (WT)
AND
Gene locus passes QC based on coverage report
THEN
add that gene at the bottom of laboratorian report AND assign:
tbprofiler_variant_substitutions = "WT"
n_mutation, aa_mutation = "WT"
confidence, depth, frequency, read_support, rationale = "N/A"

[14] Including indels in ORF

antimicrobial = corresponding drug to which this gene locus confers resistance (may have to list the same locus more than once as "WT" for each of the drugs that it is associated with, since mutations in some loci cause R to more than one drug)
Looker interpretation = "S"
MDL Interpretation = "WT"

4.2. Quality filtering of final reportable results:

4.2.1. Breadth of coverage throughout the locus: 100% of the locus must be covered with at least 10x; reflected in coverage report

4.2.1.1. **IF** breadth of coverage at 10x is = 100% = PASS, report as is.

4.2.1.2. **IF** breadth of coverage at 10x is < 100% **AND** a deletion is present= PASS, report as is.

4.2.1.3. **IF** breadth of coverage at 10x is < 100% and a deletion is absent= FAIL

AND

4.2.1.3.1. No mutations detected (WT)
THEN
Write "Insufficient coverage for the locus" in the *Warning field* for the corresponding detected S, S-Interim, R-Interim, U mutation, if any (in addition to any other warning from the variation position QC; do not overwrite).
AND
Overwrite the following values for that gene (found at the bottom of laboratorian report):
Looker interpretation= "Insufficient Coverage"
MDL Interpretation= "Insufficient Coverage"
Do not change any other fields from what they normally would be for a WT locus.

OR

4.2.1.3.2. Only "S", "S-Interim", "R-Interim" or "U" mutations are detected in given gene (i.e., no "R" mutations detected based on Looker interpretation)
THEN
Write "Insufficient coverage for the locus" in the *Warning field* for the corresponding detected S, S-Interim, R-Interim, U mutation, if any (in addition to any other warning from the variation position QC; do not overwrite).
AND
Overwrite the following values:
Looker interpretation= "Insufficient Coverage"
MDL Interpretation= "Insufficient Coverage"
Do not change any other fields from what they normally would be for the detected mutation.

OR

4.2.1.3.3. R mutation is detected in given gene (based on mutation interpretation in "Looker interpretation" field)
THEN
No additional action, report as if locus coverage QC ok

4.2.2. QC in the position where mutation is detected (see in Laboratorian report)

4.2.2.1. **IF** Detected mutation failed QC in that position in any of the following QC parameters:
Total Read Depth ("depth" column in Lab report) < 10X
Variant Read Depth ("read_support" column in Lab report) < 10X
Percent Alt Allele ("frequency" column in Lab report) < 10%
THEN
Do not output the corresponding mutation on LIMS report and treat the sequence in this position as WT
ELSE report as usual

4.3. Additional analysis notes:

4.3.1. Mutations (indels) that start outside of target regions that are a subject of an expert rule, but span into the region of interest, are interpreted as meeting the corresponding rule. E.g. a deletion in rplC occurs over -24 to -17, part of it would be in the promoter region coordinates for that gene (-18 to -1). Or indel starts outside of RRDR but continues into RRDR region. This rule covers indels that completely encompass region of interest.

4.3.2. If sample has less than 100% of the locus covered with at least 10x (failed breadth of coverage), subject matter expert must review the sequence manually to confirm absence of large deletions that may be missed by the pipeline. If a deletion of >=50 bp detected in the gene of interest, SME may report such mutation as "U".

4.3.3. All antimicrobial drugs associated with a gene are reported. This includes WHO annotation and the gene_associated_drugs field. I.e., when there is a mutation in the gene associated with two drugs (e.g. Rv0678 for BDQ and CFZ) and TBDB only has confidence listed for one drug and not another drug, the following logic was implemented:

4.3.3.1. Generate interpretation based on confidence in TBDB for the drug that is listed in TBDB;

4.3.3.2. For the drug that is not listed in the TBDB for that mutation- follow applicable expert rule.

5. LIMS report parsing:

5.1. Below is description of how a separate output file is generated for ingest into LIMS[15] by parsing results of described above interpretation output (aka "Laboratorian report") for resistance reporting (5.2-5.3), parsing of TB Profiler output for species ID (5.4).

5.2. Outputting ***drug interpretations*** in LIMS report (e.g. M_DST_B01_INH): an overall value based on the highest-severity mutation (from high to low: R > U > S > WT) is generated. Assign text value for corresponding drug according to the table below.

Information coming from Laboratorian report			**In LIMS output of TB Profiler**
MDL Interpretation (overall value based on the highest severity mutation within all targets associated with given drug)			**Drug (e.g. M_DST_B01_INH)**
R	R mutations in *rpoB* gene	Following mutations in *rpoB* gene: Leu430Pro Asp435Tyr His445Asn His445Ser His445Leu His445Cys Leu452Pro Ile491Phe AND No other R mutations in *rpoB* gene	Type "Predicted low-level resistance to rifampin. May test susceptible by phenotypic methods"
		All other R mutations in *rpoB* gene	Type "Predicted resistance to rifampin"
	R- for all other cases		Type "Mutation(s) associated with resistance to XXX detected"
U			Type "The detected mutation(s) have uncertain significance. Resistance to XXX cannot be ruled out"
S	S for all drugs, except RIF (see below)		Type "No mutations associated with resistance to XXX detected"
	S interpretation for RIF	Synonymous mutation present within *rpoB* codons 426-452 (RRDR) AND No other mutations interpreted as R or U are detected in *rpoB* (anywhere in the gene)	Type "Predicted susceptibility to rifampin. The detected synonymous mutation(s) do not confer resistance". " (If other *rpoB* mutations have MDL interpretation "U" or "R", then report

[15] LIMS output was specifically designed for CalLIMS

		P.S.: if synonymous mutations outside of *rpoB* codons 426-452 follow the rule for "S" above.	drug predicted susceptibility based on those mutations' interpretation)
		All other cases when S mutation detected in *rpoB* AND No other mutations interpreted as R or U are detected in *rpoB* (anywhere in the gene)	Type "Predicted susceptibility to rifampin"
WT (no mutations detected in the corresponding targets)			Type "No mutations associated with resistance to XXX detected"
"Insufficient Coverage" AND No other genes associated with given drug has mutation interpreted as R			Pending Retest (If other genes associated with the same drug have mutations with MDL interpretation "R", then report drug predicted susceptibility based on those targets)

5.2.1. The Drug interpretation is based only on the set of genes that is being reported. So for LIMS, if a gene is not on the reportable list, we should ignore it and only generate overall Drug interpretation based on mutations found (or not) in the genes listed in LIMS export.

5.2.2. In genes that are associated with more than 1 drug, when a mutation is listed for one drug but not listed for another:

5.2.2.1. For the drug that it is listed for this mutation in TBDB: List this mutation under the with the corresponding drug interpretation;

5.2.2.2. For another drug with which it is associated but does not have interpretation for in TBDB: do not list this mutation under that drug, i.e. results for that gene would be as if it was WT ("No mutations detected") and the drug interpretation would also state that no mutations associated w/R to XXX detected.

5.2.3. For Looker Drug interpretations (that Looker matrix table), we should take into consideration all genes for the corresponding drugs, because we will be importing information for all genes into the Looker.

5.3. Outputting ***individual mutations*** in gene target fields in LIMS report (e.g. M_DST_B02_katG):

5.3.1. Only output mutations classified as "R" or "U" in "MDL interpretations" field of Laboratorian report, for the exception of RRDR region of rpoB:

5.3.1.1. Within *rpoB* codons 426-452, output ALL individual mutations (including synonymous);

5.3.1.2. After the synonymous mutation output text "[synonymous]", e.g.: "c.2630G>A (p.Asp877Asp) [synonymous]" (RIF interpretation output is changed accordingly in case if the only mutation detected in rpoB RRDR is synonymous; see above)

5.3.2. See instructions on output format for individual mutations below.

<table>
<tr><th colspan="2">Information coming from Laboratorian report</th><th colspan="2">In LIMS output of TB Profiler</th></tr>
<tr><td colspan="2" rowspan="2">MDL Interpretation (for each given mutation)</td><td colspan="2">Gene Loci (e.g. M_DST_B02_katG), i.e. where individual mutations that are listed for each target</td></tr>
<tr><td>Format</td><td>Mutation should be listed in LIMS report</td></tr>
<tr><td colspan="2">R</td><td>Report the detected mutation in format "c.2066C>T (p.Ala689Val)" [16]</td><td>Yes</td></tr>
<tr><td colspan="2">U</td><td>Report the detected mutation in format "c.2066C>T (p.Ala689Val)"</td><td>Yes</td></tr>
<tr><td rowspan="2">S</td><td>S for mutations in RRDR region of rpoB gene:
If synonymous mutation present within rpoB codons 426-452
(P.S.: do not report synonymous mutations outside this region)</td><td>Report the detected synonymous mutation in format "c.2066C>T (p.Ala689Ala) [synonymous]"
(Report whether synonymous mutation is detected alone in RRDR or there are other R mutations in RRDR)</td><td>Yes</td></tr>
<tr><td>S- for all other cases</td><td>Mutation is not displayed on LIMS report. If the only mutation detected in given target gene was S, type "No high confidence mutations detected"</td><td>No</td></tr>
<tr><td colspan="2">WT (no mutations detected in the corresponding target)</td><td>Type "No mutations detected" for the corresponding gene target</td><td>N/A</td></tr>
<tr><td colspan="2">"Insufficient Coverage"</td><td>Type "No sequence" for the corresponding gene target</td><td>N/A</td></tr>
</table>

[16] If multiple mutations in the gene, list separated by semicolon: "c.2066C>T (p.Ala689Val); c.169C>G (p.His57Asp)"

5.3.3. If a mutation in the same position (usually a deletion) is listed under different annotation types on Laboratorian report, use the one with highest read_support to decide which deletion to report in LIMS. Leave the alternative mutation annotation on Laboratorian report but do not use it for evaluation of predicted drug resistance.

5.4. Parse TB Profiler original output that contains lineage prediction information to populate LIMS report field "M_DST_A01_ID" for species ID. Assign following values in M_DST_A01_ID depending on text in "main_lin" field of TBProfiler:

5.4.1. IF ALL genes on the coverage report that are ALSO on the LIMS report have >=90% breadth of coverage at 10x, **THEN**

5.4.1.1. If main_lin field contains "lineage": "DNA of Mycobacterium tuberculosis species detected"

5.4.1.2. If main_lin field or sublin field contains "BCG": "DNA of Mycobacterium bovis BCG detected"

5.4.1.3. If those fields do **NOT** contain "BCG", but **DO** contain "bovis" or "La1": "DNA of Mycobacterium bovis (not BCG) detected"'

5.4.1.4. If main_lin field blank **OR** "NA" **OR** non-existent: "DNA of Mycobacterium tuberculosis complex detected"

5.4.2. ELSE: "DNA of Mycobacterium tuberculosis complex NOT detected"