rMLST Species Identification

The rMLST species identification method is based on the principle of identifying the lowest common taxonomic node (LCTN) of each rMLST allele found in the bacterial genome under investigation. This LCTN information is combined for all the alleles identified within a query genome to deduce the species present.

What do we mean by lowest common taxonomic node (LCTN)?

For any rMLST allele, the allele has been identified in at least one genome sequence and defined in the rMLST sequence definition library. The genomes in the PubMLST Multispecies database have been scanned against the rMLST allele library and the genomic position of each allele variant is recorded along with the index number of the variant. This indexing allows us to observe which bacterial species the allele is present in by querying the database for all genomes that have been assigned to any particular rMLST allele.

The lowest common taxonomic node (LCTN) of each rMLST allele is calculated based upon these species annotations. For example, an allele observed in multiple *Neisseria* species is assigned an LCTN of *Neisseria* (a genus node), whereas an allele only observed in *Neisseria meningitidis* genomes is assigned an LCTN of *Neisseria meningitidis* (a species node).

The species identification process

The rMLST species identification process includes four main stages.

Firstly, the query genome is scanned against the rMLST allele library using BLASTN and exact allelic matches are recorded (i.e. locus name and allele index information).

Secondly, the lowest common taxonomic node of each observed rMLST allele is calculated dynamically based on the current species annotations of genomes linked to each allele in the PubMLST Multispecies database.

Thirdly, the LCTNs of all the matched alleles are mapped onto the nodes of the bacterial taxonomic tree and the lowest observed non-overlapping taxonomic nodes are identified and reported.

Finally, the 'allele support' of each reported taxonomic node is calculated, which is the number of alleles observed for the reported node divided by the total number of alleles observed across all reported nodes (expressed as a percentage). A single reported species node with an allele support above 90% indicates a high degree of confidence in that result.

A simple example

A *Neisseria* genome is scanned against the rMLST library and 55 exact allelic matches are observed. Ten of these alleles have a LCTN of *Neisseria meningitidis* and 45 have an LCTN of *Neisseria*. The *Neisseria* genus node is contained within the node path between the phylum node (Pseudomonadota) and the *Neisseria meningitidis* species node so the genus node is not reported. Ten out of 10 alleles ‘support’ the reported *Neisseria meningitidis* species node therefore the support will be 100%.

Another example

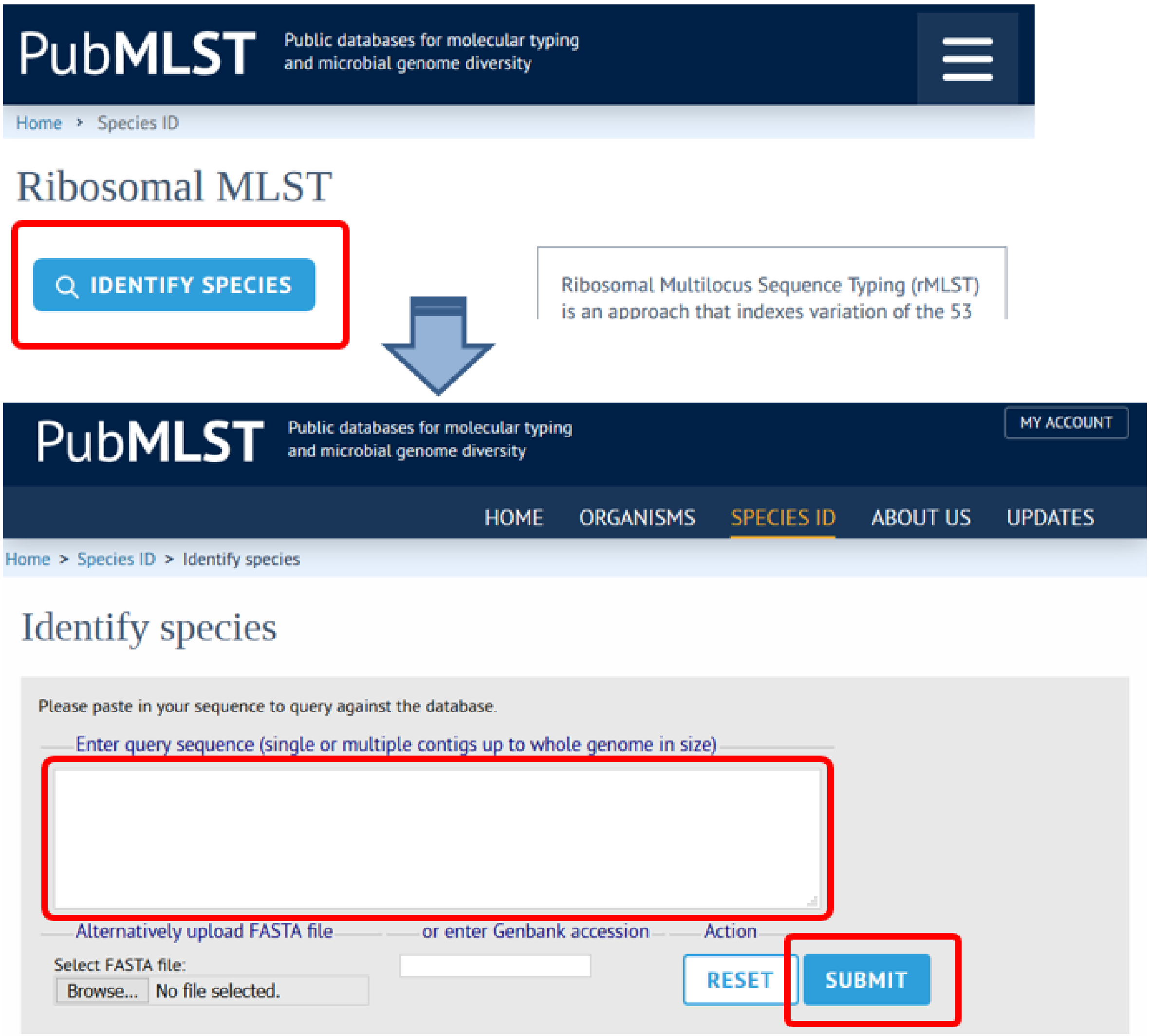
Another *Neisseria* genome is scanned against the rMLST library and 65 exact allelic matches are observed. This time, 15 alleles have a LCTN of *Neisseria gonorrhoeae*, 40 have an LCTN of *Neisseria* and 10 alleles have been found in various *Mycobacterium* species and the LCTN of these alleles is therefore calculated to be the *Mycobacterium* genus node. The *Mycobacterium* genus node does not lie within the phylum to species node path of *N. gonorrhoeae* and so this node is reported. In total 25 alleles support the two reported nodes, *Neisseria gonorrhoea* has 60% allele support (15/25) and *Mycobacterium* genus has 40% allele support (10/25). The reported presence of two nodes from different bacterial phyla often indicates that the input genome sequence contains DNA contamination.

If the species identification result has two species in the same genus and species A has a high allele support (>90%) and species B has a very low allele support (<10%), then this can be interpreted as strong support for species A, rather than presence of any contaminating DNA.

Handling non-specific species annotations

There are species annotations in the PubMLST Multispecies database that are non-specific e.g. *Neisseria sp.* where the genus is known but the exact species is not described. If a non-specific annotation is observed along with a species annotation in the same genus, the calculated LCTN is the species e.g. *Neisseria meningitidis* and *Neisseria sp.*, the LCTN is *Neisseria meningitidis*. However if *Neisseria sp.* is observed along with a different genus then the lowest common node is calculated. e.g. *Neisseria sp.* and *Kingella oralis*, the calculated LCTN is the family node, Neisseriaceae.

Figure 1: Enter genome sequence and click submit



PubMLST
Public databases for molecular typing and microbial genome diversity
Home › Species ID

Ribosomal MLST

Ribosomal Multilocus Sequence Typing (rMLST) is an approach that indexes variation of the 53

PubMLST
Public databases for molecular typing and microbial genome diversity
MY ACCOUNT
HOME
ORGANISMS
SPECIES ID
ABOUT US
UPDATES
Home > Species ID > Identify species

Identify species

Please paste in your sequence to query against the database.

Enter query sequence (single or multiple contigs up to whole genome in size)
Alternatively upload FASTA file
Select FASTA file:
Browse...
No file selected.
or enter Genbank accession
Action
RESET
SUBMIT

Figure 2: Overview of the species identification process

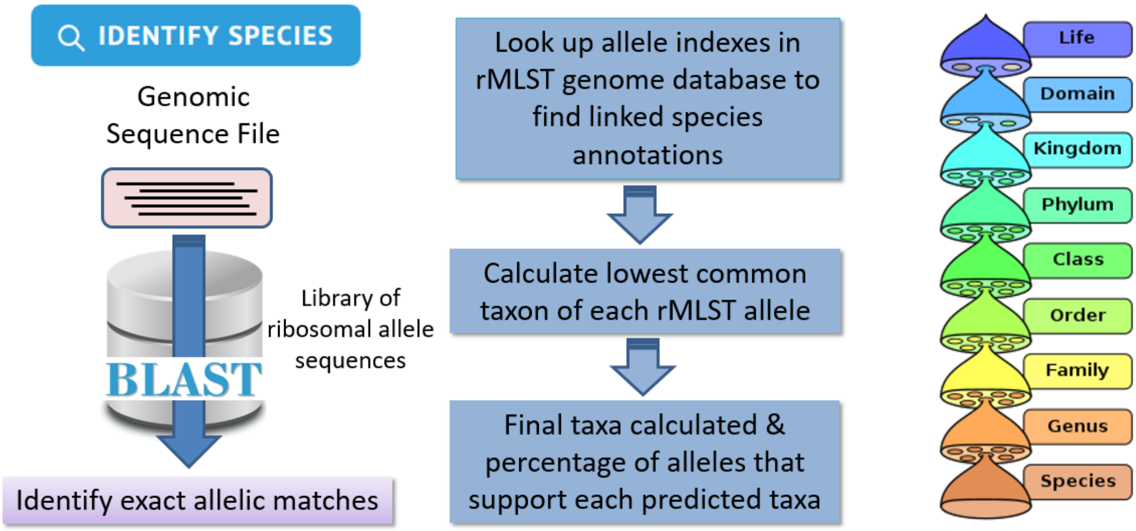
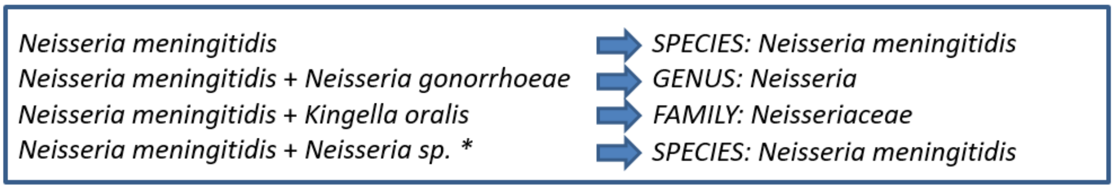

IDENTIFY SPECIES
Genomic Sequence File
Library of ribosomal allele sequences
BLAST
Identify exact allelic matches
Look up allele indexes in rMLST genome database to find linked species annotations
Calculate lowest common taxon of each rMLST allele
Final taxa calculated & percentage of alleles that support each predicted taxa
Life
Domain
Kingdom
Phylum
Class
Order
Family
Genus
Species

Figure 3: Example of allele matches, linked species annotations and calculation of the lowest common taxonomic node per allele for a subset of allelic matches observed for a *Neisseria meningitidis* query genome.

Query Genome
Allele Matches
Linked Species Annotations
Calculate Lowest Common Taxon

Locus	Allele	Linked data values	Lowest Common Taxon
BACT000001 (rpsA)	3	Neisseria meningitidis [n=5]	SPECIES: Neisseria meningitidis
BACT000002 (rpsB)	3	Neisseria meningitidis [n=4469]	SPECIES: Neisseria meningitidis
BACT000003 (rpsC)	3	Neisseria meningitidis [n=4651]	SPECIES: Neisseria meningitidis
BACT000004 (rpsD)	2	Neisseria meningitidis [n=7826]; Neisseria sp. [n=1]	SPECIES: Neisseria meningitidis
BACT000005 (rpsE)	1	Neisseria meningitidis [n=17211]; Neisseria gonorrhoeae [n=1]; Neisseria sp. [n=1]	GENUS: Neisseria
BACT000006 (rpsF)	3	Neisseria meningitidis [n=4536]	SPECIES: Neisseria meningitidis
BACT000007 (rpsG)	2	Neisseria meningitidis [n=4778]	SPECIES: Neisseria meningitidis
BACT000008 (rpsH)	2	Neisseria meningitidis [n=7495]	SPECIES: Neisseria meningitidis
..		..	..
BACT000062 (rpmG)	2	Neisseria meningitidis [n=4765]; Neisseria polysaccharea [n=13]	GENUS: Neisseria
BACT000063 (rpmH)	3	Neisseria meningitidis [n=8459]; Neisseria polysaccharea [n=17]; Neisseria sp. [n=10]; Neisseria cinerea [n=1]	GENUS: Neisseria
BACT000064 (rpmI)	2	Neisseria gonorrhoeae [n=15991]; Neisseria meningitidis [n=14709]; Neisseria bergeri [n=46]; Neisseria polysaccharea [n=30]; Neisseria sp. [n=13]	GENUS: Neisseria
BACT000065 (rpmJ)	1	Neisseria meningitidis [n=20313]; Neisseria sp. [n=1]	SPECIES: Neisseria meningitidis
BACT000065 (rpmJ)	1025	Neisseria meningitidis [n=9044]; Neisseria lactamica [n=642]; Neisseria bergeri [n=39]; Neisseria polysaccharea [n=22]; Neisseria sp. [n=17]; Neisseria gonorrhoeae [n=2]	GENUS: Neisseria

Figure 4: Simple overview of calculating the lowest common taxonomic node.



Neisseria meningitidis → *SPECIES: Neisseria meningitidis*

Neisseria meningitidis + Neisseria gonorrhoeae → *GENUS: Neisseria*

Neisseria meningitidis + Kingella oralis → *FAMILY: Neisseriaceae*

*Neisseria meningitidis + Neisseria sp. ** → *SPECIES: Neisseria meningitidis*

* Rule for ‘sp.’ annotations: When combined with a species-level annotation in the **same genus**, the species-level annotation is the result.

All lowest common taxons are mapped onto the nodes of the bacterial taxonomic tree and only the lowest observed node of each branch is reported

Allele support is defined as the percentage of alleles that have contributed directly to an individual reported taxa across all reported taxa