FOOD SAFETY LAB / MILK QUALITY IMPROVEMENT PROGRAM
Department of Food Science, Cornell University

MQIP
MILK QUALITY IMPROVEMENT PROGRAM

Title: User guide to management of *rpoB* database

Table of content

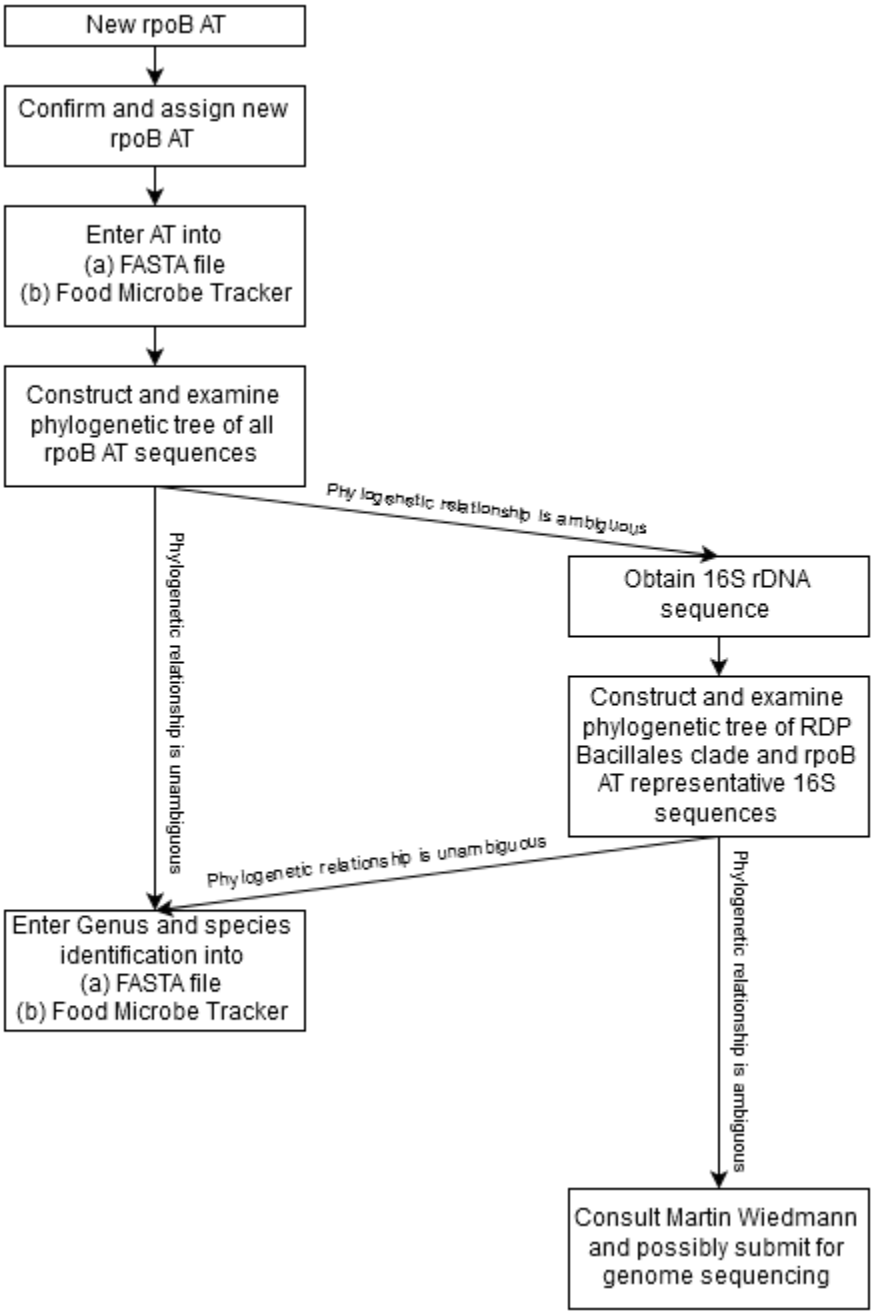


Figure 1: Flow Chart of steps needed to provide isolate with a new *rpoB* sequence type number (AT number) and Genus/species identification

1. INTRODUCTION

1.1. Purpose

- To identify novel rpoB sequence allelic types (AT) of bacteria isolated in the Milk Quality Improvement Program (MQIP) and Food Safety Laboratory (FSL).
- To provide detailed instructions for assigning new allelic types, genus and species names, managing, and updating the rpoB database.

1.2. Scope

This SOP applies to the MQIP and the FSL.

1.3. Definitions

- **AT:** allelic type; defined as one specific DNA sequence within a gene, in this case a 632 nucleotide region of the rpoB gene in Bacilleaceae
- **bp:** base pair
- **BLAST:** Basic Local Alignment Sequence Tool
- **Consensus:** a single sequence derived from a set of overlapping DNA segments originating from one genetic source
- **FMT:** Food Microbe Tracker; WWW-based tool for information exchange on bacterial subtypes and strains, containing a large amount of bacterial gene information
- **PCR:** Polymerase Chain Reaction, used to amplify a specific region within a DNA sequence.
- **Percentage sequence identity:** proportion of identical nucleotides between two sequences multiplied by 100.
- **Phylogeny** – The evolutionary history of taxonomic groups.
- **16S rRNA:** 16S ribosomal ribonucleic acid, a RNA component of the 30S small subunit of prokaryotic ribosomes
- **rpoB:** RNA polymerase beta subunit

2. MATERIALS

- **Computer** with multicore processor
- **Partial rpoB gene sequences**: From PCR products of isolates; sequences are obtained after PCR products are sent to the BRC facility (Biotechnology Research Center). Raw data files are in .ab1 format and edited consensus sequences are saved as *.fas files.
- **Internet Access:** For accessing Food Microbe Tracker and Ribosomal Database Project.
- Geneious: Comprehensive Software for Molecular Biology. Geneiuos can be used in all steps. The lab has subscription through Cornell BRC. To use Geneious exclusively, several plugins and workflows need to be installed (see detail later).
- **Mesquite**: Evolutionary biology software that can be used to make alignments. Can be downloaded. Mesquite is used in Section 3.2 for use in database management and multiple sequence alignments.
- **Muscle**: Multiple-alignment algorithm and program. It can be downloaded : It must be inputted into Mesquite during your first use of the program. Muscle is used with Mesquite in Section 3.2 for database management.
- **RaxML**: Algorithm and program used to create phylogenetic trees based on maximum likelihood and bootstrapping. RaxML GUI is the graphical user interface version of RaxML.
- **rpoB database FASTA file**: File that is kept up to date with rpoB allelic typing entries, which is subsequently used to create phylogenetic trees, only to be updated by database managers \\cornell.edu\ag\FOOD\FOOD-MQIP\rpoB database\Current database file. This file is available upon request.
- **FigTree**: Program used to view phylogenetic tree files, used in conjunction with RaxML.
- **BioEdit**: Biological sequence alignment editor that can be used for blasting.

3. PROCEDURE

3.1 Confirming New rpoB Allelic Types Using Food Microbe Tracker

Monthly or bi-weekly, collect rpoB sequences that have been sent by individuals in the lab working on rpoB gene sequencing. They will have sent an email to the rpoB database manager and should have added raw and consensus (final) sequence data to \\cornell.edu\ag\FOOD\FOOD-MQIP\rpoB database\Possible new ATs.

Confirm the new rpoB ATs by BLASTing the sequence against the current rpoB AT database as described in 3.1.1.

Users who submitted the potentially new rpoB allelic type should have checked the ab1 files to make sure the single nucleotide polymorphism(s) are legitimate in Sequencher or Geneious, but the database manager must also confirm the SNPs in the raw files.

3.1.1 Introduction

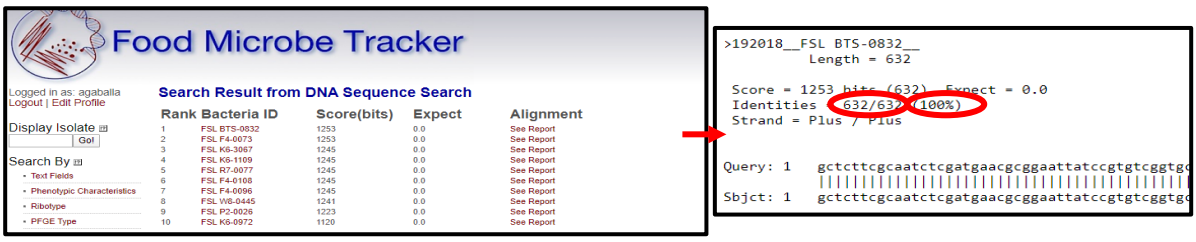
In order to confirm new allelic types, search each new sequence against the rpoB database in Food Microbe Tracker. The “Search By DNA Sequence” feature uses the BLAST algorithm to search a given sequence (query) against a database (the rpoB AT database), and returns the existing ATs with the best possible alignments to the provided query sequence. By doing this search, you will confirm that the existing AT with the best possible alignment is not a 100% match, and hence that the query sequence is a new allelic type. In extreme circumstances with many new ATs, the Food Microbe Tracker search tool may be difficult to use, and USEARCH (see Appendix B) may be used instead.

3.1.2 Using Food Microbe Tracker (few isolates):

1. Obtain edited (final) *rpoB* sequence data (Creating a consensus DNA sequence from ABI sequence data using Sequencher) from the user. These sequences should be at a minimum, 632 bp in length for *rpoB*, but are often longer.
2. Open website for Food Microbe Tracker: http://www.foodmicrobetracker.com
    a. Log-in, or request account for Log-in.

b. In Food Microbe Tracker (FMT), on the left-hand side of the main page, under "Search By", click on "DNA Sequence".
c. Once on this page use the pull-down menus to adjust your search parameters:
    i. "Number of Results": *default=10*, you may wish to increase/decrease this.
    ii. "Genus": default=Unspecified.
    iii. "Species": default=Unspecified.
    iv. "Sequence Type": *default=Unspecified*, **this must be changed to "*rpoB* allelic typing"** in the pull-down menu (from the pull-down menu, make sure that you are using ***rpoB* allelic typing** and **not *rpoB***)

3. Open the *rpoB* consensus (final) sequence file (.fas). This can be done in either Notepad or Sequencher.
4. Copy and paste the *rpoB* sequence into the space labeled "Enter DNA sequence".
5. Click Submit. Once your results page ("Search Results from DNA Sequence Search") has loaded, choose the first Alignment file by clicking on "See Report" in red.
6. When the new page/tab has appeared in your web browser, click to view it and review some key details:



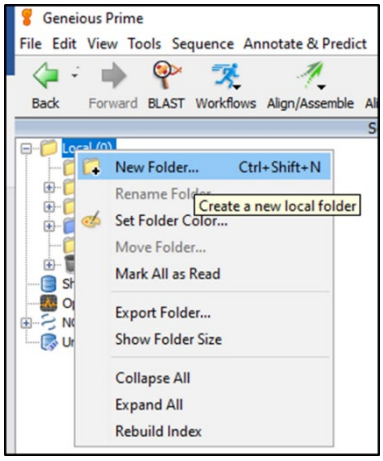
a. "Identities": this should read "632/632 (*100%*)" for a 100% allelic type match. Unique AT sequences will have less than a 100% AT match.
b. "Query": is the *rpoB* allelic type sequence you entered.
"Sbjct": is the database *rpoB* allelic type sequence and it should start at 1 and end at 632 (a few *rpoB* ATs have 829 bp or 635 bp).
c. If there are no *rpoB* ATs with a 100% match to the sequence queried (Identities do not equal "632/632 (100%)") you may have a new *rpoB* allelic

type.
Check in Sequencher if the differences between your sequence and the sequence in the database are legitimate, i.e., they are not artifacts introduced during editing.

If you have confirmed a 100% identity match between an rpoB allelic typing sequence currently listed in FMT and the query sequence, contact the user to inform her/him that this sequence is not new rpoB AT.
It is strongly advisable that you talk to the user to find out the basis on which they considered this sequence to be a potential new AT to ensure that the correct analysis is being carried out.

3.1.3 rpoB Allelic Type Assignment with Geneious (multiple isolates analysis)

1. Copy the rpoB current database and "rpoB_AT_Workflow" files from: \\cornell.edu\ag\FOOD\FOOD-MQIP\rpoB database\Current database file to a local folder on your computer. This file is available upon request.
   When starting a new analysis, always copy a new database file from the MQIP server to guarantee that you are using the current database version.
2. Open Geneious and create new folder.

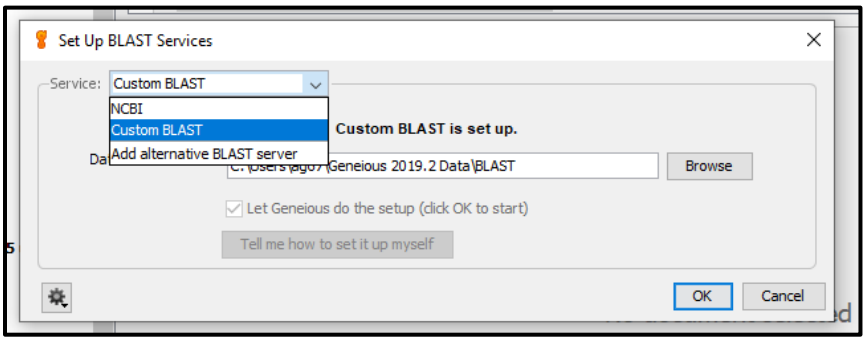


3. Drag and drop or import the *rpoB* current database file (File → import → from file: keyboard shortcut: Ctrl+I).
4. Create local Blast database in Geneious:

a. Set Up Blast Services: This step must be done only once on the first time of using local Blast in Geneious.

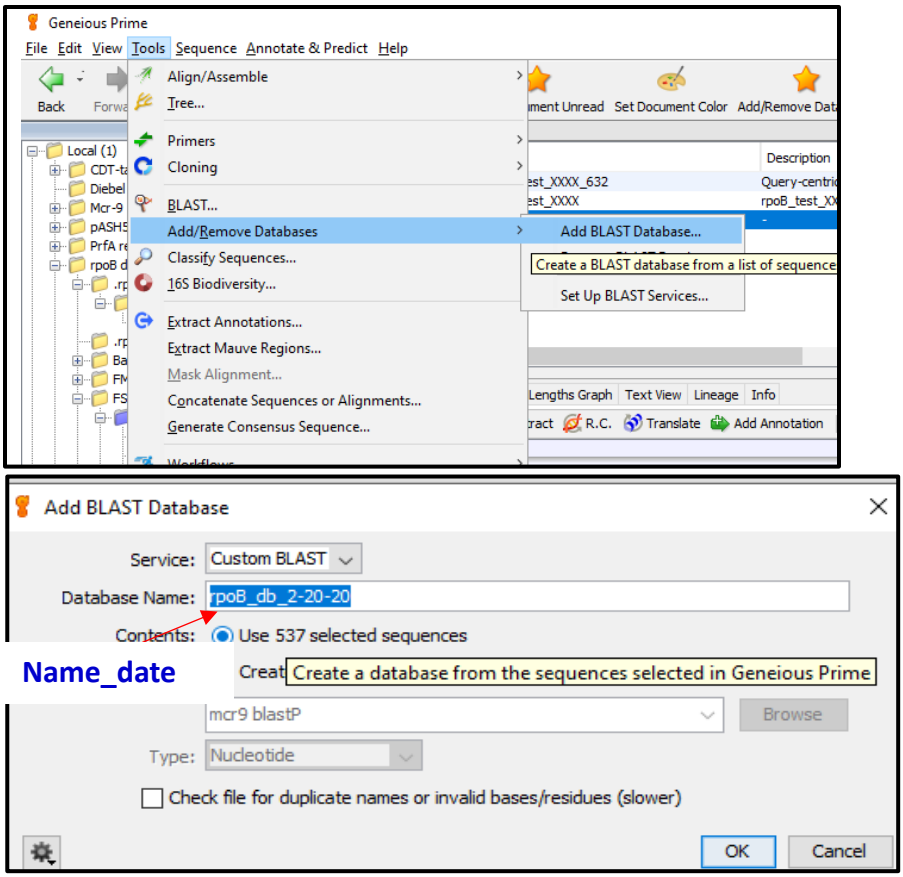
   i. In Geneious menu go to Tools→ Set Up Blast Services.

   ii. In the popup window:
   Service: select “Custom Blast”. You can leave Folder location as suggested or browse to change where you want to save your files. Hit OK and wait until the setup is complete.

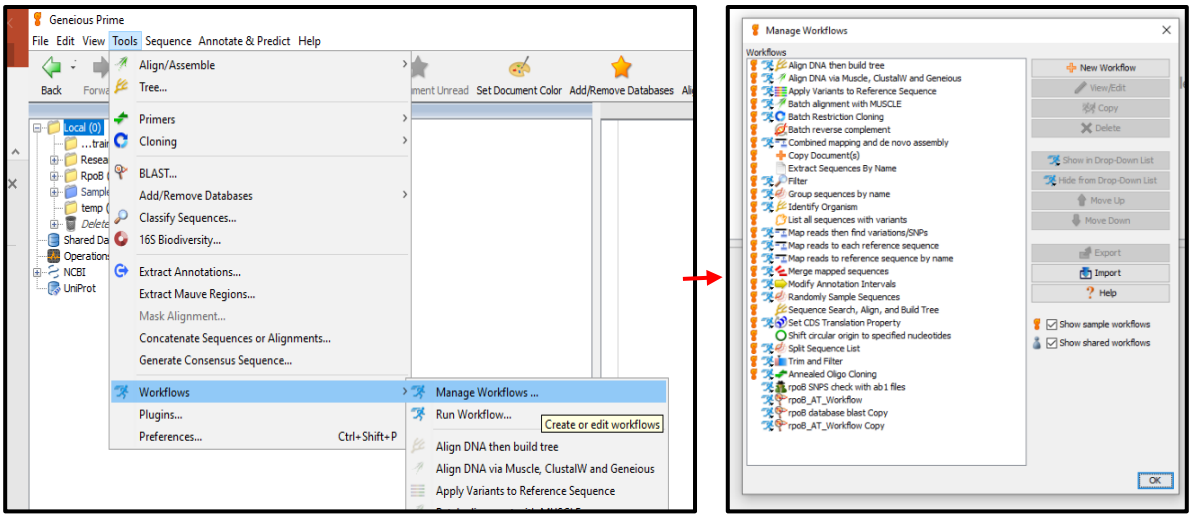


b. In Geneious, select the *rpoB* current database file (single left click) →Tools→ Add/remove databases → Add BLAST database.

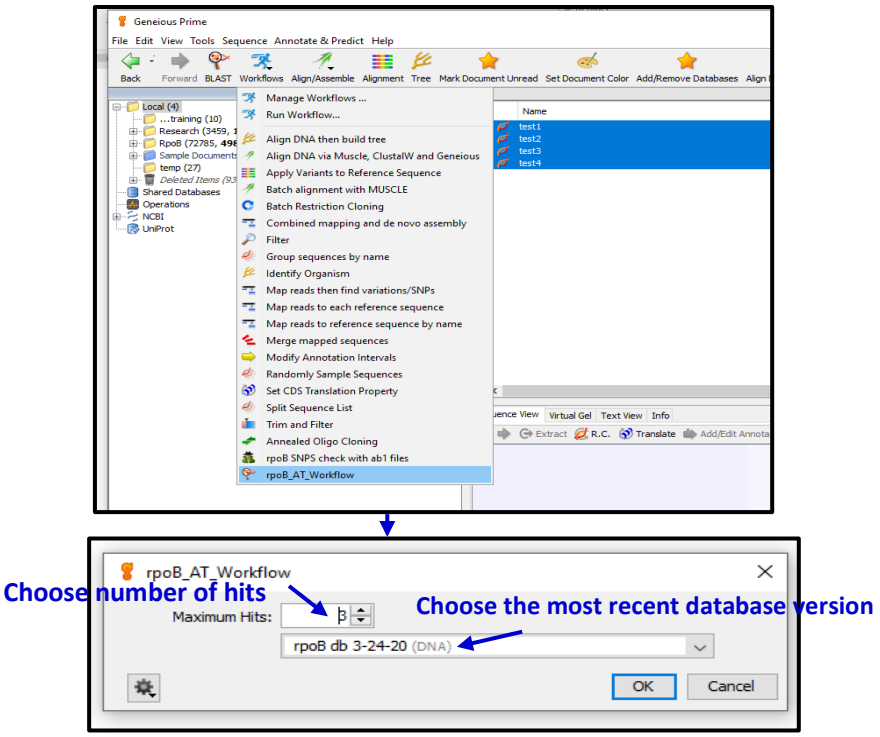
   In the popup window:
   Service: select “Custom BLAST”
   Database Name: “*rpoB* database DATE”
   Click on: Use “number” selected sequences
   Click: OK



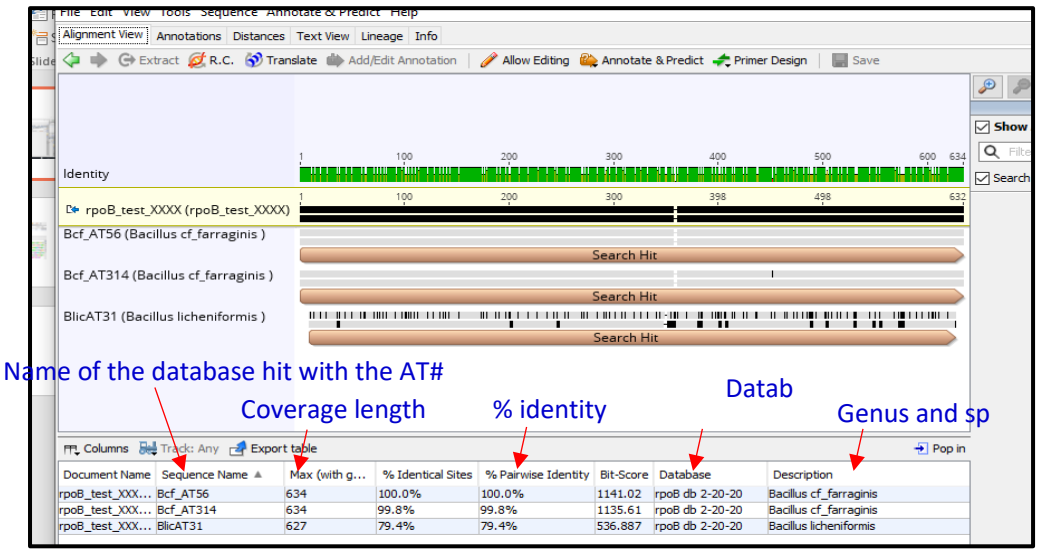
5. In Geneious, import the consensus *rpoB* sequences:
   Create a new folder; drag and drop all consensus *rpoB* sequences (in fasta format) into the folder. Geneious might ask to either keep the sequences separate or in a list: choose separate.
6. Import “rpoB_AT_Workflow”:
   In Geneious: go to Tools → Workflows → Manage workflows
   in the popup window: click Import (on the right) and upload the “rpoB_AT_Workflow”.



7. Select all *rpoB* sequences to be analyzed (depending on your computer's memory, you might have to analyze 50 to 100 sequences at a time).
8. Run the workflow:
   Tools → Workflows → click on "rpoB_AT_Workflow".
   Select the number of hits that you want to see for each sequence and the *rpoB* database (make sure to use the database with most recent date). Hit OK to run workflow.



9. The workflow will analyze each sequence separately and generate an alignment file for each sequence.



10. Click on the file and the alignment will appear in the right-bottom window. Click on "Annotations" in Geneious right-bottom window toolbar to show details of the

database hit, percent identity, description of the database hit, which includes genus, species, AT number etc.

Hint: if you do not see all the columns in the toolbar of Geneious right-bottom window, click on columns, select “show all”

11. If you have confirmed a 100% identity match between an rpoB allelic typing sequence currently listed in FMT and the query sequence, contact the user to inform her/him that this sequence is not new rpoB AT.
It is strongly advisable that you talk to the user to find out the basis on which she/he considered this sequence to be a potential new AT to ensure that the correct analysis is being carried out.

3.2 Clade Clustering of New rpoB Allelic Types

Database manger must ensure that users confirmed all SNPs for all putative new *rpoB* AT sequences from the raw ab1 files and she/he must perform a second independent confirmation from the raw data.

3.2.1 Using Mesquite, RAxMLGUI, and Fig Tree to Determine Clade Clustering of New rpoB Allelic Types

After new rpoB allelic types have been confirmed using Food Microbe Tracker, further analysis is needed to determine clade clustering.

3.2.1.1. Assembling Alignments Using Mesquite

Mesquite is free software for PCs, Macs, and Linux systems. Use *Mesquite* to make an alignment of old and new rpoB sequences. If downloading Mesquite 3.0, use the 2 GB version. On some computers, an error prevents the user from running Mesquite 2GB. If this occurs, run the 1GB version. After downloading *Mesquite,* download the algorithm Muscle. Remember where the muscle file lives. *Muscle* contains the set of commands that dictates how the alignment is assembled.

A. After downloading *Mesquite*, open the program and then open the existing rpoB database file from within the program. The default file format for *Mesquite* is nexus. However, you can just as easily import other file types, including FASTA, and the program will prompt you to save in nexus format upon import.

B. Select *Show Matrix*. Taxa will be at the left of the *Character Matrix* and accompanying sequence to the right. Additional rpoB files can be added to the *Character Matrix* by drag and drop or by selecting *File Incorporation* → *Merge Taxa & Matrices* → select file type → *Fuse with Selected Taxa Block* → *Fuse with Selected Matrix*. If using drag and drop, drop new rpoB file(s) at the bottom of the Character Matrix. This keeps any new Allelic Types (ATs) in sequential order even though taxa can be moved around.
C. To create alignment select *Edit* → *Select all* → *Matrix* → *Align Multiple Sequences* → *Muscle Align* → do not run on separate thread → locate muscle file → *ok* . The process isn't instantaneous so there will be some waiting for the alignment to finish.
D. Scroll through the alignment and check new sequences for deletions or insertions. If any of these are present, they will need to be confirmed in the chromatogram (can be viewed in *Sequencher*) before moving forward. Trim off excess sequence from each end of the alignment by holding down shift and clicking the outer blocks of the area to be deleted (Numbers at the top of the alignment can be selected to trim multiple sequences at one time). Once selected, select *Edit* → *Cut* to selected entries.
E. Now save the alignment as an aligned FASTA file by going to *File* → *Menu*, selecting "FASTA (DNA/RNA)", and then select the "include gaps" checkbox, ensuring that no other checkboxes are selected. Click "Export," and name the file "rpoBdatabaseYYYYmmdd.fafa". Place this file in \\cornell.edu\ag\FOOD\FOOD-MQIP\rpoB database\Current database file, and move the old FASTA file to \\cornell.edu\ag\FOOD\FOOD-MQIP\rpoB database\Archived database files. It is the new master database file. For non-Cornell users, please contact the database manger.

3.2.1.2 Check for duplicates

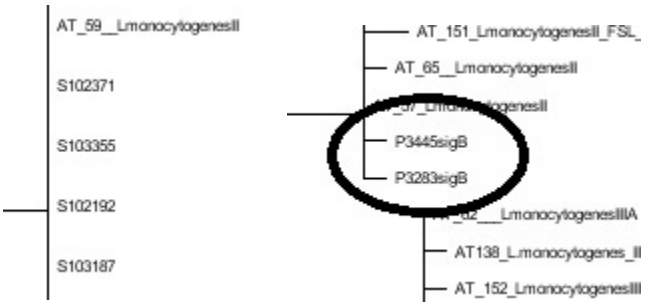
It is advisable at this stage to confirm that the database has no duplicates (entries with identical sequence, i.e., single unique sequence per AT number). This can be done by uploading the FASTA file to:

https://www.hiv.lanl.gov/content/sequence/elimdupesv2/elimdupes.html

3.2.1.3 Constructing Phylogenetic Trees Using RAxMLGUI

A. Export the Mesquite alignment into Phylip format for use in RAxML
   i. File->Export->Phylip (DNA/RNA)
   ii. Change maximum length of taxon names to 100
   iii. Click export.
B. Download *raxmlGUI* to construct a phylogenetic tree. For users more familiar with the command prompt, *RAxML* can be installed and used instead. *raxmlGUI* also requires that *Python 2.5-2.7* be installed. *raxmlGUI* is not compatible with Python 3.0 (as of September 2014). Open *raxmlGUI* and *load alignment* that was exported from Mesquite in Phylip format. RaxMLGUI will display a message 'RAxML found at least 1 sequence that is exactly identical to other sequences and/or gap-only characters in the alignment. Do you want to exclude it/them from the analysis?', select 'No'. Make sure *ML + rapid bootstrap* is selected and set *reps.* to a minimum of 100. All other parameters can be left at the default. Select *Run RAxML*. The run can take a while so it is best to work on something else during this time.
C. *RaxmlGUI* will generate multiple output files in the folder/directory the alignment was loaded from: (1) the best-scoring ML tree 'RAxML_bestTree.YOUR_FILE_NAME.tre, (2) Best-scoring ML tree with bootstrap support values 'RAxML_bipartitions.YOUR_FILE_NAME.tre, (3) Best-scoring ML tree with bootstrap support values as branch labels 'RAxML_bipartitionsBranchLabels. YOUR_FILE_NAME.tre', (4) Program execution info 'RAxML_info. YOUR_FILE_NAME.tre', (5) All 100 bootstrapped trees 'RAxML_bootstrap. YOUR_FILE_NAME.tre, (6) a phylip formatted file with all unique sequences 'YOUR_FILE_NAME.reduced', and (7) a file listing which sequences are identical to other sequences in the original file 'RAxML_info". The last two files are important for further analyses. The ".reduced" file will contain only unique rpoB sequences. Therefore, this file will only contain existing representative ATs and any representative new ATs. The "reduced" and "info" files can be used to identify a representative new allelic type and determine identical sequences.
D. To view the tree created by *RAxML*, open the file named *RAxML_bipartitions.filename* in *Fig Tree* or another tree viewing program. Type *bootstrap* in the text field when prompted to select a name for the node/branches.

E. Zoom in to find new isolates in the tree and observe species clustering. It can be determined whether any of your isolates are duplicates as in the "RAxML_info" file. These isolates will be next to one another on the same branch. The two isolates in the left picture below are right next to each other but are different ATs because each isolate has its own branch. The picture on the right shows an example of multiple identical isolates. Be careful when viewing the tree. If a group of isolates share the same vertical branch as depicted in the right picture, but are separated by nodes, they are still identical to each other.

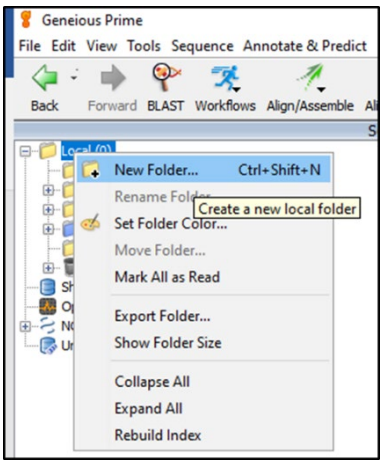


3.2.2 Using Geneious to Determine Clade Clustering of New rpoB Allelic Types

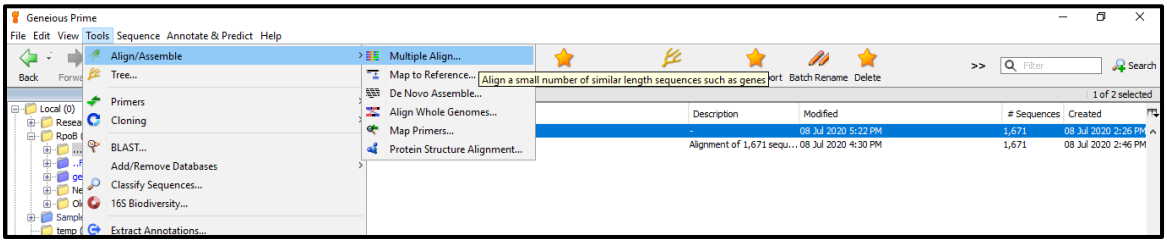
After new rpoB allelic types have been confirmed, further analysis is needed to determine clade clustering.

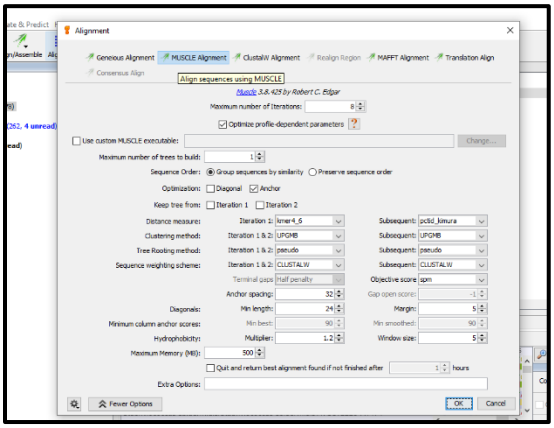
3.2.2.1 – Assembling Alignments

A. Open Geneious and create new folder. Right click on Local in the left window and from the submenu, select New Folder.

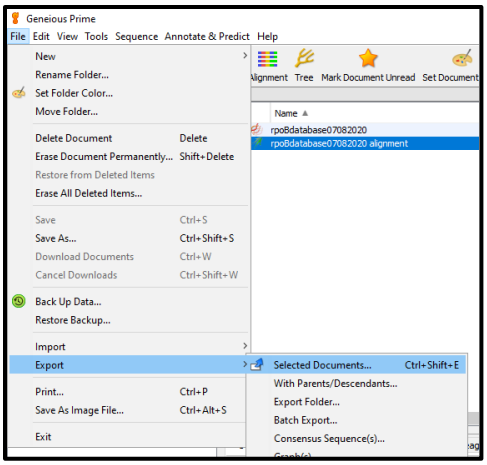


B. Import the database FASTA files by dragging and dropping the files into Geneious folder. Alternatively, import the file from file menu → import → "From file" (or Ctrl+I), browse to the files location, select the files, and click "import".
C. Construct multiple alignment by selecting all the FASTA files and select from the menu → Tools → Align/Assemble → Multiple Align.





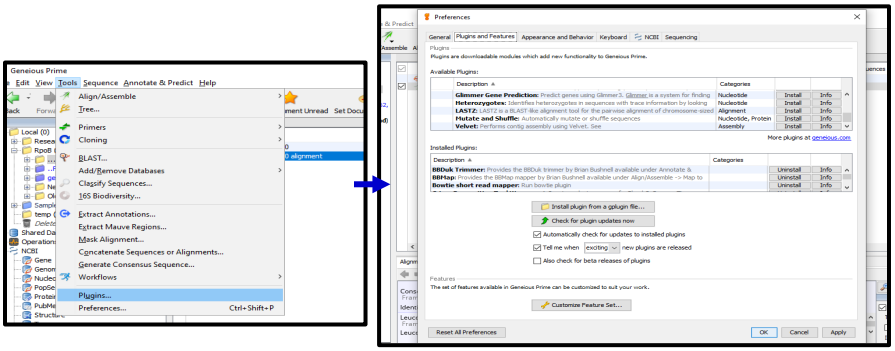
D. From the popup window select MUSCLE Alignment tab, keep default and hit OK. Wait until alignment is done.
E. You may need to rename the alignment file: select the alignment file, hit F2 and rename the file in the format: rpoBdatabaseYYYYmmdd.
F. Export both the alignment and the non-aligned files as FASTA files from File → Export → Selected Documents (or Ctrl+Shift+E), selecting "FASTA Sequences/Alignment" from the drop menu and click "Export". Place these files in \\cornell.edu\ag\FOOD\FOOD-MQIP\rpoB database\Current database file, and move the old FASTA file to \\cornell.edu\ag\FOOD\FOOD-MQIP\rpoB database\Archived database files. For non-Cornell users, please contact the database manger.



The phylogenetic tree can be constructed using RaxML within Geneious (see below), however, it is slower than using the standalone RaxMLGUI. If you will use RaxMLGUI, export the alignment file in phylip format by going to File → Export → Selected Documents (or Ctrl+Shift+E), selecting "phylip Alignment" from the drop menu and click "Export". Select strict file format from the popup window.

3.2.2.2 Constructing Phylogenetic Trees Using RAxMLGUI within Geneious (overnight run is strongly advised).

A. Install RaxML and Figtree plugins in Geneious. This is done only once before using the plugin for the first time. From the menu go to Tools → Plugins. From the popup window, scroll to RaxML from "Available Plugins" section and click "install". Select Figtree plugin and click install.

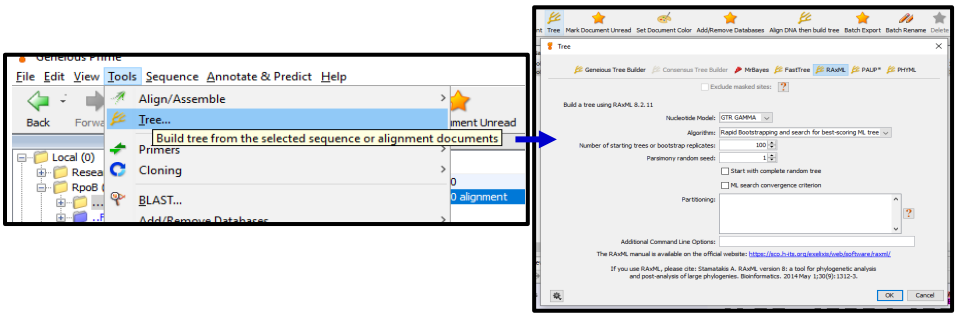


B. Select the alignment file and build phylogenetic tree. From the menu go to Tools→Tree.
C. From the popup menu, select RaxML tab.

D. Change the following settings:

   i. Algorithm: Rapid Bootstrapping and search for best scoring ML-tree.
   ii. Number of starting trees or bootstrap replicates :100

E. Click OK. This step might take more than 12 hours.



F. The best scoring tree will have the name “file name RAxML Tree”.
G. If you select the tree file, it should directly open with Figtree plugin within Geneious. However, you can export the file from File → Export → Selected Documents (or Ctrl+Shift+E) as newick and open it with the standalone FigTree.

3.3 Assigning Genus/species identifications to new rpoB allelic types

Background: Identification of the genus/species associated with a new rpoB allelic type will first be attempted by rpoB sequence analysis (section 3.3.1 below); if these data do not allow for unambiguous assignment, 16S rDNA sequence data will be generated for the isolate with a given “new” rpoB AT and these 16S data will be used to assign the genus/species of the new rpoB AT, using the RDP database (see section 3.3.1 below).

3.3.1 Genus and species assignment based on rpoB ML tree

If a new rpoB allelic type (see section 3.2 on how to produce a rpoB ML tree with RAxML)

i. clusters in the rpoB ML tree with previously identified cluster of isolates or bacillales type strain with a bootstrap support ≥ 70.0%, and shows ≥97.0% nt BLAST identity to a type strain or an existing sequence (over the whole 632 nt) in the database, the genus/species name of existing isolate is given to the new AT. Numbers will not be rounded to determine whether they meet cut-off (e.g., if

bootstrap support is 69.999%, it will have not passed the threshold even though it could be rounded to 70.0%).

ii. clusters in the rpoB ML tree with 2 or more type strains, use the name format : "Genus name" sp. (clade #)_"species 1 name "_"species 2 name". For example: ***Paenibacillus sp._massiliensis_panacisoli***.

iii. If phylogenetic relationship to existing rpoB allelic types is unclear (< 97% rpoB sequence BLAST identity to an existing sequence in the database and/or it does **not** cluster with a group of previously identified isolates with a bootstrap value of ≥ 70%), a partial 16s rDNA sequence will be needed for species assignment. 16S rDNA sequencing will be performed using primers 16s-PEU7 and 16s-DG74 for the PCR and 16s-PEU7 and 16s-P3SH for the sequencing as outlined in the FSL 16s protocol.

3.3.2. Overview of 16S rDNA analysis

Once 16s rDNA data is received (from person doing original rpoB sequencing), the 16S rDNA data will be used for similarity based analyses in RDP (http://rdp.cme.msu.edu/), and phylogenetic analysis of a master 16s rDNA matrix of isolates with rpoB allelic types (\\cornell.edu\ag\FOOD\FOOD-MQIP\rpoB database\16S sequences for representative ATs). For non-Cornell users, please contact the database manger.

3.3.2.1. RDP analysis:

Go to the RDP website (http://rdp.cme.msu.edu/) and select 'Sequence Match (http://rdp.cme.msu.edu/seqmatch/seqmatch_intro.jsp)'. Paste your 16S sequence in the window and select the following option in the radio button menu: Strain; Type, Source; Both, Size; Both, Quality; Good, Taxonomy; Nomenclatural. These settings will allow for a search against a database of type strains only. Press 'submit'. Within seconds you should get a list showing the taxonomic ranking of your organism. Click on 'view selectable matches' to see the match to type strains in RDP. The 'S_ab score' will give you a similarity score of your sequence to the matching sequence in the database. Based on the similarity score, you can assign genus and species directly (see section 3.3.3.1) or you will have to construct a phylogenetic tree based on 16S sequences (see section 3.3.2.2 and 3.3.3.2).

3.3.2.2. Phylogenetic analysis based on 16S sequences using RAxML GUI:

12. Read section 3.2 on how to use RAxML and Mesquite
13. Download a file with the 'current' 16S sequences of *Bacillales* type strains from RDP:
14. Go to the RDP website (http://rdp.cme.msu.edu/) and click 'hierarchy browser'.
15. In the radio button menu choose: Strain; Type, Source; Both, Size; Both, Quality; Good, Taxonomy; Nomenclatural. Click 'browse'
16. Click 'phylum Firmicutes
17. On the next page click on the '+' symbol next to 'order Bacillales' this selects all type strain sequences for download
18. click 'download' in the upper right corner
19. on the next page click 'Download XXX sequence(-s) for alignment model RDPX-2 Bacteria' to download a FASTA file with the sequences.
20. Import the RDP FASTA file, a FASTA file with unique 16S sequences of previous searches (\\cornell.edu\ag\FOOD\FOOD-MQIP\rpoB database\16S sequences for representative ATs), and the FASTA file(-s) of the isolates to be analysed into Mesquite and align using the Muscle option. After alignment, trim the matrix in such a way that the start and the end of the matrix do not contain gaps. The minimum length of the matrix should be 650 characters (including gaps). For non-Cornell users, please contact the database manger.
21. Export the aligned and trimmed file in phylip format and import in RAxML GUI.
22. Run RAxML using at least 100 pseudoreplicates
23. Open the tree in Figtree; display bootstrap values on branches.
24. Check if phylogenetic placement is congruent with RDP based results.

3.3.3 Genus and species assignment based on 16S sequence similarity or 16S ML tree

3.3.3.1 Genus and species assignment based on 16S sequence similarity

Based on 16S sequence similarity/' S_ab score' in RDP we apply the following naming conventions (if a situation does not fall under any of these scenarios below or if a situation does not pass the common sense test, consult with the Martin.

A. ≥99% similarity to one type strain: adopt name of the type strain.

B. ≥99% similarity to ≥ 2 type strains: Usually an indication that the isolate belongs to a species complex (e.g., *Bacillus subtilis* complex). In this case, the "sensu lato" notation is used to indicate a high level of similarity with multiple closely related species (for instance, *B. cereus* sensu lato); currently this is only used for
   i. The *B. subtilis* complex (see Rooney PMID 20048064 for species in this complex).
   ii. The *B. cereus* complex: This group is highly complex and species assignment based on rpoB and 16S sequences can be unclear. As of July 2020, nomenclature is based on the proposed nomenclatural framework for the *B. cereus* group as described by Carroll et al., (mBIO 2020, PMID: 32098810). This complex includes *B. cereus*, *B. anthracis*, *B. mycoides*, *B. pseudomycoides*, *B. thuringiensis* and *B. weihenstephanensis.*
   iii. If the 16S tree along with a literature search indicates a possible species complex that has not been previously described, designate as species complex, add to Appendix A, and name after species that was first described (see Bergey's) in the complex (e.g., *B. licheniformis* complex, which also includes *B. aerius* and *B. sonorensis*). In some cases, it may also be appropriate to designate an AT to a single species (for example, when there is no evidence of a species complex); these instances should be indicated and explained in Appendix A. In these cases where rDNA based trees are needed for identification, you should also contact the resident taxonomy expert (indicated below).

C. <99% similarity to a type strain, create a 16S tree (see section 3.3.2.2 on building the tree):
   i. If the isolate clusters with one type strain in the 16S rRNA tree (meaning shows a common ancestor), use the name of that type strain, but insert cf. (confer) notation to denote taxonomic uncertainty (for instance, *Paenibacillus* cf. *peoriae*).
   ii. If the isolate clearly clusters with 2 or more type strain in the 16S rRNA tree (meaning shows a common ancestor (ancestral node) with more than one type strain); use the genus name as confirmed by phylogenetic analysis of 16S, and call the isolate as *Genus sp.* clade # (e.g., *Brevibacillus* sp. clade 1) and in FMT indicate the species it clusters with under "Basis of species ID" (e.g., clusters with *Brevibacillus limonophilus and B. reuszeri)*. For rpoB database enter as *Brevibacillus*|sp_clade1_*limonophilus_reuszeri*. Note that this is similar to

(B) and the inclusion of closely related species names is optional (and a judgment call). If the 16S rDNA sequence clusters with 3 or more type strains (different species), do not include all species names and just name as *Genus sp.* clade # (e.g., *Brevibacillus* sp. clade 1).

iii. If the isolate does not group with one or more type strains in the 16S rRNA tree; use the genus name as confirmed by phylogenetic analysis of 16S, and a numbered clade (e.g., *Paenibacillus* sp. clade 1) to indicate taxonomic distinctiveness and taxonomic uncertainty. Give each clade a unique numerical identifier (e.g., clade 1, clade 2). Isolates that fall in this category may represent new species and should be discussed with a *"Bacillales* taxonomy expert" (as of 07/2020 this would be Martin Wiedmann or Ahmed Gaballa). If this 16S sequence does not clearly cluster with a group of type strain representing a given genus, name as *Bacillaceae* genus nov. (add the isolate in cue for genome sequencing).

Updating Database Files/Records

1. After new rpoB allelic types have been confirmed in *Fig Tree* and their genus/species determined (see section 3.3), taxa need to be updated in the master FASTA file using *Mesquite* or Geneious. This will include renaming taxa with AT #, genus/species, and FSL #. Double click on each "Taxa" name, and add "rpoB|ATxxxx|[FSL ID]|Genus|Species", adding new AT numbers in consecutive order, FSL ID should remain. New rpoB ATs numbers will be assigned chronologically (by date of discovery). New duplicate ATs should also be removed from the alignment by highlighting taxon/character → *Edit* → *Cut*.
2. The new FASTA file (e.g. rpoBdatabase20150908.afa) should be saved and put in \\cornell.edu\ag\FOOD\FOOD-MQIP\rpoB database\Current database file. The other file in the folder (which should now be older) must then be moved into \\cornell.edu\ag\FOOD\FOOD-MQIP\rpoB database\Archived database files.
3. Add new AT sequence(s) to Food Microbe Tracker (FMT) → open FMT → go to the given isolate's page → *Add* a *DNA Sequence* → select *rpoB allelic typing* from drop down → paste in trimmed sequence or upload FASTA file of final trimmed sequence → *submit*. *Add* AT # under *Additional Characteristics*.

4. Add new ATs to rpoB Access Database (maintained by database manager). Information to add includes the date, manager's net i.d., AT range added, and specific ATs with associated species.

3.5 Periodic database clean-up

Background: To ensure that Genus/species information remains consistent for all isolates of the same rpoB allelic type, the database will be queried periodically to find all entries with an rpoB allelic type entered and for which the Genus/species doesn't match the Genus/species of the corresponding representative rpoB allelic type strain.

3.5.1 Send an email to Qi Sun (qisun@cornell.edu) and request a copy of the rpoB species discrepancy table. If the table contains no records, no clean-up is necessary. Otherwise, save the table to MQIP→ rpoB database→Species discrepancy tables→[current date].xlsx. Open the species discrepancy table.

3.5.2 Update the FMT entry for each isolate listed in the species discrepancy table: (1) Open the FMT page for the isolate. (2) Next to "General isolate info," click "Edit." In the "Anecdotal Isolate History" field, add "Previous ID:" followed by the Genus and species that are presently entered for the isolate. (3) Copy the "rep genus" field from the species discrepancy table. Paste it into the "Genus" field on the FMT page. (3) Copy the "rep species" field from the species discrepancy table. Paste it into the "Species" field on the FMT page. (4) Scroll to the bottom of the FMT page and click "Submit." (5) Review the FMT page to ensure that the Genus and species fields now match the Genus and species of the representative rpoB allelic type strain, and that the previous Genus and species are recorded in "Anecdotal isolate history".

3.5.3 After completing 3.4.2 for each isolate in the table, send another email to Qi Sun to verify that the species discrepancy query now returns no records.

4. TROUBLESHOOTING

4.1 If "Identities" in your results read anything but out of 632 (e.g. 630/630 or 626/631), the BLAST algorithm has somehow trimmed your sequence. First check the second best match, if that one isn't out of 632 either, then it is best to pull out rpoB sequences from each isolate's FMT page and align them (ClustalW or Mesquite can do this) and see if the complete length matches for 632bp.

4.1.1 Matches showing 635/635 usually indicate a match with a *Staphylococcus sp*. Close attention should be paid to these isolates. Record percentage identities to the best rpoB AT match. Report to database manager, or Martin Wiedmann. These sequences may be added to the database so that these Genera can be identified, however if you find many of these in your project there may be a breakdown in laboratory methods that may need to be investigated.

5. REFERENCES

1. Branquinho R, Klein G, Kampfer P, Peixe LV. (2015). The status of the species *Bacillus aerophilus* and *Bacillus stratosphericus*. Request for an Opinion. *Int J Syst Evol Microbiol* 65(Pt 3):1101.
2. Carroll LM, Wiedmann M, Kovac J. (2020) Proposal of a Taxonomic Nomenclature for the Bacillus cereus Group Which Reconciles Genomic Definitions of Bacterial Species with Clinical and Industrial Phenotypes. mBio 11(1).
3. Dunlap CA. The status of the species *Bacillus aerius*. Request for an Opinion. (2015) Int J Syst Evol Microbiol 65(7):2341.
4. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput *Nucleic Acids Res.* 32(5):1792-1797.
5. Finore I, Gioiello A, Leone L, Orlando P, Romano I et al. (2017) Aeribacillus composti sp. nov., a thermophilic bacillus isolated from olive mill pomace compost. *Int J Syst Evol Microbiol* 67(11):4830-4835.
6. Gupta RS, Patel S. (2019) Robust Demarcation of the Family Caryophanaceae (Planococcaceae) and Its Different Genera Including Three Novel Genera Based on Phylogenomics and Highly Specific Molecular Signatures. *Front Microbiol* 10:2821.
7. Krishnamurthi S, Chakrabarti T, Stackebrandt E. (2009) Re-examination of the taxonomic position of *Bacillus silvestris* Rheims et al. 1999 and proposal to transfer it to *Solibacillus* gen. nov. as *Solibacillus silvestris* comb. nov. *Int J Syst Evol Microbiol* 59(Pt 5):1054-1058.
8. Liu Y, Lai Q, Shao Z. (2018) Gnome analysis-based reclassification of *Bacillus weihenstephanensis* as a later heterotypic synonym of *Bacillus mycoides*. Int J Syst Evol Microbiol 68(1):106-112.
9. Liu Y, Ramesh Kumar N, Lai Q, Du J, Dobritsa AP et al. (215). Identification of strains Bacillus aerophilus MTCC 7304T as Bacillus altitudinis and Bacillus

stratosphericus MTCC 7305T as a Proteus sp. and the status of the species Bacillus aerius Shivaji et al. 2006. Request for an Opinion. *Int J Syst Evol Microbiol* 65(9):3228-3231.

10. Liu GH, Narsing Rao MP, Dong ZY, Wang JP, Che JM et al. (2019) Genome-based reclassification of *Bacillus plakortidi*s Borchert et al. 2007 and *Bacillus lehensis* Ghosh et al. 2007 as a later heterotypic synonym of *Bacillus oshimensis* Yumoto et al. 2005; *Bacillus rhizosphaerae* Madhaiyan et al. 2011 as a later heterotypic synonym of *Bacillus clausii* Nielsen et al. 1995. *Antonie Van Leeuwenhoek* 112(12):1725-1730.
11. Madhaiyan M, Poonguzhali S, Lee JS, Lee KC, Hari K. (2011) Bacillus rhizosphaerae sp. nov., an novel diazotrophic bacterium isolated from sugarcane rhizosphere soil. *Antonie Van Leeuwenhoek* 2011;100(3):437-444.
12. Reddy GS, Uttam A, Shivaji S. (2008) *Bacillus cecembensis* sp. nov., isolated from the Pindari glacier of the Indian Himalayas. *Int J Syst Evol Microbiol* 58(Pt 10):2330-2335.
13. Shivaji S, Chaturvedi P, Suresh K, Reddy GSN, Dutt CBS et al. (2006) Bacillus aerius sp. nov., Bacillus aerophilus sp. nov., Bacillus stratosphericus sp. nov. and Bacillus altitudinis sp. nov., isolated from cryogenic tubes used for collecting air samples from high altitudes. *Int J Syst Evol Microbiol* 56(Pt 7):1465-1473.
14. Stamatakis, F. Blagojevic, C.D. Antonopoulos, D.S. Nikolopoulos: (2007) Exploring new Search Algorithms and Hardware for Phylogenetics: RAxML meets the IBM Cell *Journal ofVLSI Signal Processing Systems* 48(3):271-286.
15. Yoon JH, Lee KC, Weiss N, Kho YH, Kang KH et al. (2001) Sporosarcina aquimarina sp. nov., a bacterium isolated from seawater in Korea, and transfer of Bacillus globisporus (Larkin and Stokes 1967), Bacillus psychrophilus (Nakamura 1984) and Bacillus pasteurii (Chester 1898) to the genus Sporosarcina as Sporosarcina globispora comb. nov., Sporosarcina psychrophila comb. nov. and Sporosarcina pasteurii comb. nov., and emended description of th. *Int J Syst Evol Microbiol* 51(Pt 3):1079-1086.

APPENDIX A: SPECIES COMPLEXES AND EXCEPTIONS

Complex name/exception	Members	Support
Bacillus cereus	➢ *B. cereus* ➢ ***B. anthracis* → *B. mosaicus subsp. anthracis*** ➢ *B. mycoides* ➢ *B. pseudomycoides* ➢ ***B. thuringiensis*→ *B. cereus sensu stricto* serovar Berliner biovar Thuringiensis strain ATCC 10792** ➢ ***B. weihenstephanensis*→ *B. mycoides*** ➢ *B. toyonensis* ➢ *B. luti* ➢ ***B. mobilis*→ *B. mosaicus* strain 0711P9-1** ➢ ***B. pacificus* → *B. mosaicus* strain EB422** ➢ ***B. paranthracis*→ *B. mosaicus* strain MN5** ➢ ***B. wiedmannii* → *B. mosaicus* strain FSL W8-0169**	1. Nomenclature is based on the proposed nomenclatural framework for the B. cereus group as described by Carroll et al., (mBIO 2020, PMID: 32098810). 2. *Bacillus weihenstephanensis* is a heterotypic synonym of *Bacillus mycoides* (PMID 29095136).
Bacillus subtilis s.l.	*B. mojavensis* *B. atrophaeus* *B. amyloliquefaciens*	http://www.ncbi.nlm.nih.gov/pubmed/20048064

Bacillus licheniformis s.l.	*B. licheniformis* *B. aerius* *B. sonorensis* *B. aerophilus* *B. stratosphericus*	See Bacillus aerius entry in Bergey's second edition http://www.ncbi.nlm.nih.gov/pubmed/16825614
Paenibacillus amylolyticus s.l.	*P. amylolyticus* *P. xylanexedens* *P. tundrae*	http://www.ncbi.nlm.nih.gov/pubmed/19542122
Bacillus altitudinis group	*Bacillus aerius* *Bacillus aerophilus* *Bacillus stratosphericus* *Bacillus altitudinis*	*B. aerius , B. aerophilus, B. stratosphericus & B. altitudinis* were described in 2006 (PMID: 16825614). In 2015, IJSEM two publications proposed to the Judicial Commission of the International Committee of Systematics of Prokaryotes to place the names of *B. aerophilus, B. stratosphericus & B. aerius* on the list of rejected names (PMID: 25908707 & PMID: 25479956). Later, it was suggested that on the basis of 16S rRNA, *rpoB, gyrB* and *pycA* gene sequence analyses, characterization of biochemical features and other phenotypic traits and pulsed-field gel electrophoresis (PFGE) fingerprinting that B. aerius MTCC 7303 and B. aerophilus MTCC 7304 were indistinguishable from Bacillus altitudinis DSM 21631T (PMID: 26297145).

Bacillus rhizosphaera	*Bacillus rhizosphaera* *Bacillus clausii*	Strain was described as a new species *Bacillus rhizosphaera* (PMID: 21671194), later it was re-classified based on WGS and described as heterotypic synonym of *Bacillus clausii* (PMID: 31312953).
Solibacillus isronensis *Solibacillus silvestris*	*Solibacillus isronensis* *Solibacillus silvestris*	The rpoB and 16S partial sequences are identical in *Solibacillus isronensis* and *Solibacillus silvestris* (PMID: 19406792 & PMID: 32010063)
Aeribacillus composti	*Aeribacillus composti* i	*Aeribacillus composti* is closely related to *Aeribacillus pallidus* (16S 99.8%). Only one publication described *Aeribacillus composti* (PMID: 28984237). 16S sequence is not in RDP database as type strain. *Aeribacillus composti* more likely is an isolate of *Aeribacillus pallidus*.
Bacillus globisporus	*Bacillus globisporus*	Homotypic synonym: *Sporosarcina globispora* (PMID: 11411676)
Bacillus cecembensis	*Bacillus cecembensis*	The strain was identified as *Bacillus cecembensis* (PMID: 18842851) and it was renamed *Solibacillus cecembensis* (PMID: 32010063).

APPENDIX B:ALTERNATIVE SEARCH TOOLS

B.1 Using BioEdit

Uploading a local nucleotide database file: The Food Safety Laboratory and Milk Quality Improvement Program have local rpoB allelic type databases. Every time one of these databases is updated, it also needs to be updated in Bioedit before blasting any sequences. To update database in BioEdit, go to Select Accessory Application→BLAST→Create a local nucleotide database file→find the most recent rpoB database file (make sure it is in FASTA format) → select the most current rpoBhaplotypes file (e.g. rpoBhaplotypes08062010.fas) →Open.

Blasting rpoB sequences that are in FASTA – concatenated format: Select *File→Open…*→find your FASTA – concatenated file containing your final sequences *→Open→Edit→Select All Sequences→Accessory Application→BLAST→Local BLAST*→select *Yes* when prompted to do a batch job→select the most current rpoB file from the *Nulceotide Database* drop-down menu→*1* for *Max number of hits to report→1* for *Max number of alignments to show→Do Search*. The window that appears contains all your sequences in the appropriate order. The relevant information under each isolate or query is as follows: the *Database* that was blasted against (should be the most current rpoB file), the best allelic type match with the corresponding genus/species/lineage (e.g. AT_25_*Paenibacillus odorifer*), *Identities* (if your isolate is a perfect match with an existing allelic type, this will read *632/632*), an alignment between your isolate and the best allelic type match (e.g. AT_25_*Paenibacillus odorifer*). Record the allelic type and genus/species/lineage of your isolate. Scroll down to view the next isolate. If *Identities* are not out of 632 (e.g. 630/630), there was an editing error (e.g. too much was trimmed off one end of the sequence) or the sequencing reaction was not sufficient to produce clean sequence of adequate length. If the sequence cannot be re-edited to produce a sequence out of 632, it will have to be re-sequenced.

Blasting sequences that are in FASTA form: This is identical to Blasting sequences that are in FASTA – concatenated format with a few exceptions. If you try to open multiple FASTA files at once within Bioedit, they will all open in their own window which is one reason to use FASTA – concatenated files for blasting multiple sequences at once. Additionally, since there is only one isolate in the FASTA file unlike the multiple isolates within the FASTA – concatenated file, the BLAST will only return one search within the BLAST window. Also, you will not be prompted to do a batch job.

New allelic types: If your isolate is not a perfect match with the closest rpoB allelic type match, *Identities* will read 631/632, 626/632, etc. This may indicate a new rpoB

allelic type. To legitimize a new rpoB allelic type, view the alignment between your isolate (*Query*) and the best match (*Sbjct*). Where there is a single nucleotide polymorphism (SNP), a line will be missing between the *Query* and *Sbjct* in the alignment. Copy the *Query* sequence around the SNP (copy roughly 25 bases) and remember where the SNP resides within the copied sequence. Open the contig file and chromatogram for that isolate using Sequencher. After opening select *Select→Find Bases…*→paste the copied sequence in the text field→*Exact Matches*→*Find*→observe the chromatogram at the location of the SNP to make sure the peaks are clean and that the SNP is not a result of an editing error. Repeat for every other SNP if there is more than one. If SNP(s) are legitimate, export the new allelic type from Sequencher as its own .fas file. Add to alignment in Mesquite (see below).

Do not save any BLAST searches upon closing Bioedit.

B2. Using USEARCH

USEARCH, as with BioEdit, is a free sequence analysis program. It does not have a graphical user interface. Therefore, you will need to use the command prompt. Sufficient instructions are provided here to use USEARCH with the command prompt.

1. Download USEARCH and move the program file to a permanent location.
2. Open the command prompt by typing *cmd* in the start menu search field
3. In the command prompt, the current directory will be showing (e.g. C:\Users\skw59>). To change directories type *cd* followed by a space and then the directory you want to go to (e.g. C:\Users\skw59>cd Desktop). Hit enter and you will be taken to that directory (e.g. C:\Users\skw59\Desktop>). Keep on going until you find the directory USEARCH is in.
4. The following is what you might enter next: "c:\Program Files (x86)\usearch.exe" -search_global LauraNewRpoB09092014forusearch.fas -id 1 -db rpoBhaplotypes02152011.fas -strand both -maxaccepts 1 -blast6out rpoB.out (don't add period at end of this). The program you are running and its location is in between the quotation marks. –search_global is the command which searches a file against a database.

The file you are searching, *LauraNewRpoB09092014forusearch.fas*, follows the *–search_global* command. *–id* refers to the identity threshold. It is set to 1 so that only 100% matches are shown in the output file. *-db* specifies the database being searched against. *–strand both* will check both the forward and its reverse complemented strand. *-maxaccepts* is the number of hits each sequence will receive. This is set to 1 so only the closest match is seen. *-blast6out* indicates the type of output file and *rpoB.out* is the name of the output file. It is also important to note that the query file (e.g. *LauraNewRpoB09092014forusearch.fas)* and database file (e.g. *rpoBhaplotypes02152011.fas)* need to be in the same folder (e.g. Desktop). The pathway all together is as follows:
C:\Users\skw59\Desktop>"c:\Program Files (x86)\usearch.exe" -search_global LauraNewRpoB09092014forusearch.fas -id 1 -db rpoBhaplotypes02152011.fas -strand both -maxaccepts 1 -blast6out rpoB.out
The information for the search criteria described above can be found on the USEARCH website.

5. *-blast6out* produces a tab-separated text file that can be opened in excel. All the output fields for *-blast6out* are below but can also be found on the USEARCH website:

Field	Description
1	Query label.
2	Target (database sequence or cluster centroid) label.
3	Percent identity.
4	Alignment length.
5	Number of mismatches.
6	Number of gap opens.
7	1-based position of start in query. For translated searches (nucleotide queries, protein targets), query start<end for +ve frame and start>end for -ve frame.
8	1-based position of end in query.
9	1-based position of start in target. For untranslated nucleotide searches, target start<end for plus strand, start>end for minus strand.
10	1-based position of end in target.
11	E-value calculated using Karlin-Altschul statistics.
12	Bit score calculated using Karlin-Altschul statistics.

6. To determine if there are any new allelic types, run the program with *–id* .97 (you can use .98 or .96). Give the output file a different name. This will report all queries that are a 97% match or better to an allelic type already in the database. Take the Queries reported from the .97 file and compare them to the original file with only 100% matches. Use sorting

and formatting options to determine which queries are missing from the 100% file that are in the 97%. These are the new allelic types.

Use BioEdit or Food Microbe Tracker to confirm the new ATs and determine if the SNPs are legitimate (see 3.1.3 above).