1. **PURPOSE:** All genomic sequences generated by PulseNet participating laboratories are analyzed using the BioNumerics software. In the first phase of analysis, the raw sequence data is checked for quality, a de novo assembly is generated and the genus/species is identified. Only those sequences that pass critical quality metrics and belong to PulseNet organisms are transferred to organism-specific surveillance databases. The purpose of this document is to describe a standardized procedure for the workflow in the Reference Identification (RefID) database.

2. **SCOPE:** This procedure applies to all whole genome sequence data generated by PulseNet participating laboratories.

3. **DEFINITIONS/ACRONYMS:**
   - **3.1. Allele:** One of two or more alternative forms of a gene that arise by mutation and are found at the same place on a chromosome.
   - **3.2. Analysis Certified:** An individual who is certified for checking the quality, performing analysis and uploading WGS data to the PulseNet National Database and NCBI using BioNumerics.
   - **3.3. ANI:** Average Nucleotide Identity.
   - **3.4. BaseSpace:** Illumina cloud-based computing environment for next generation sequencing data analysis, management and storage, including data sharing.
   - **3.5. BioNumerics:** Analysis software used by PulseNet, developed by Applied Maths (Sint-Martens-Latem, Belgium).
   - **3.6. CDC:** Centers for Disease Control and Prevention.
   - **3.7. CE:** Calculation Engine. A server application on a high performance computing cluster at CDC that is used by the BioNumerics client applications for doing calculation-intensive tasks, including de novo assembly, reference mapping and whole genome multi-locus sequence typing (wgMLST) allele calling.
   - **3.8. CE Store:** A temporary storage location for the fastq-files on the CDC server that is accessed by the calculation engine.
   - **3.9. Core genome:** genes shared by all strains of the same species.
   - **3.10. Coverage:** The average number of reads that include a given nucleotide in the reconstructed sequence.
   - **3.11. Critical Quality Metrics:** Average denovo coverage, average quality (Q score), assembly length, secondary species abundance (contamination) and percent core present. Failing to meet the minimum thresholds/acceptable range for any one of these metrics will result in rejection of the sequence.
   - **3.12. De Novo Assembly:** A sequence assembly generated from the short raw reads without the use of a reference genome.
   - **3.13. FASTQ**: A text-based file format for storing both sequence and its corresponding quality scores.
   - **3.14. NCBI:** National Center for Biotechnology Information, part of the National Institutes of Health (NIH). NCBI houses several databases relevant to biotechnology, including GenBank for DNA sequence assemblies and Sequence Read Archive (SRA) for raw reads.

3.15. Organism-specific Database: A BioNumerics database, v 7.6 or higher, used for comparing isolates for surveillance. Part of the standard PulseNet workflow.

3.16. PN: PulseNet.

3.17. QC: Quality Control.

3.18. Q score: The sequence quality score for each individual base position in a sequence. Phred scores are used, where Q = -10log (Error Probability). The higher the quality score, the more reliable the base call. A Q30 means a 1 in 1000 likelihood of an incorrect base call at that position.

3.19. Read: A unit of continuous DNA sequence derived by sequencing a part of the insert in the fragmented target DNA.

3.20. RefID Database: A BioNumerics database, v 7.6 or higher, used for quality control of raw sequence data, assembly of sequences, contamination detection and species identification. Part of the standard PulseNet workflow.

3.21. SOP: Standard Operating Procedure.

3.22. WGS: Whole Genome Sequencing.

3.23. wgMLST: Whole Genome Multi-Locus Sequence Typing.

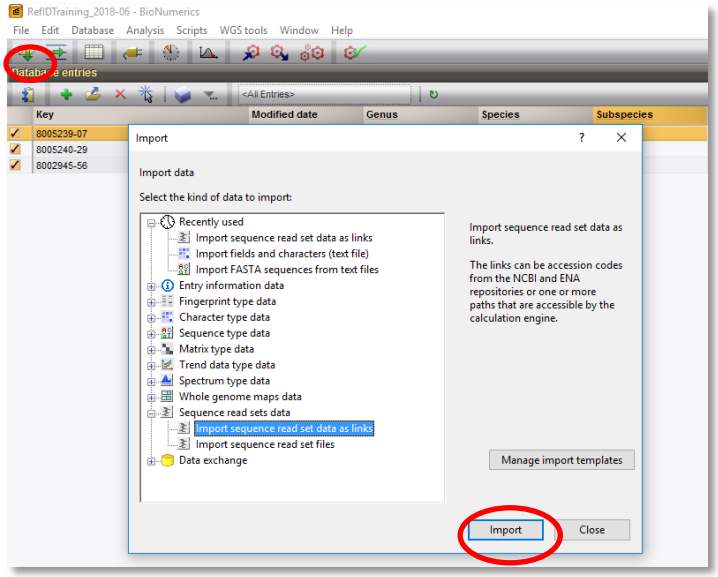
4. RESPONSIBILITIES:

4.1. Analysis certified PulseNet public health laboratory personnel perform quality checks, de novo assembly and genus/species identification of the sequences generated in their laboratory by using the BioNumerics 7.6 or higher RefID database workflow.

5. PROCEDURE:

5.1. Import Sequence Data as Links:

5.1.1. From the File menu in the main BioNumerics window, choose “Import”, expand the options for “Sequence read sets data”, select the “Import sequence read set data as links” option in the window that pops up, and then click the “Import” button.



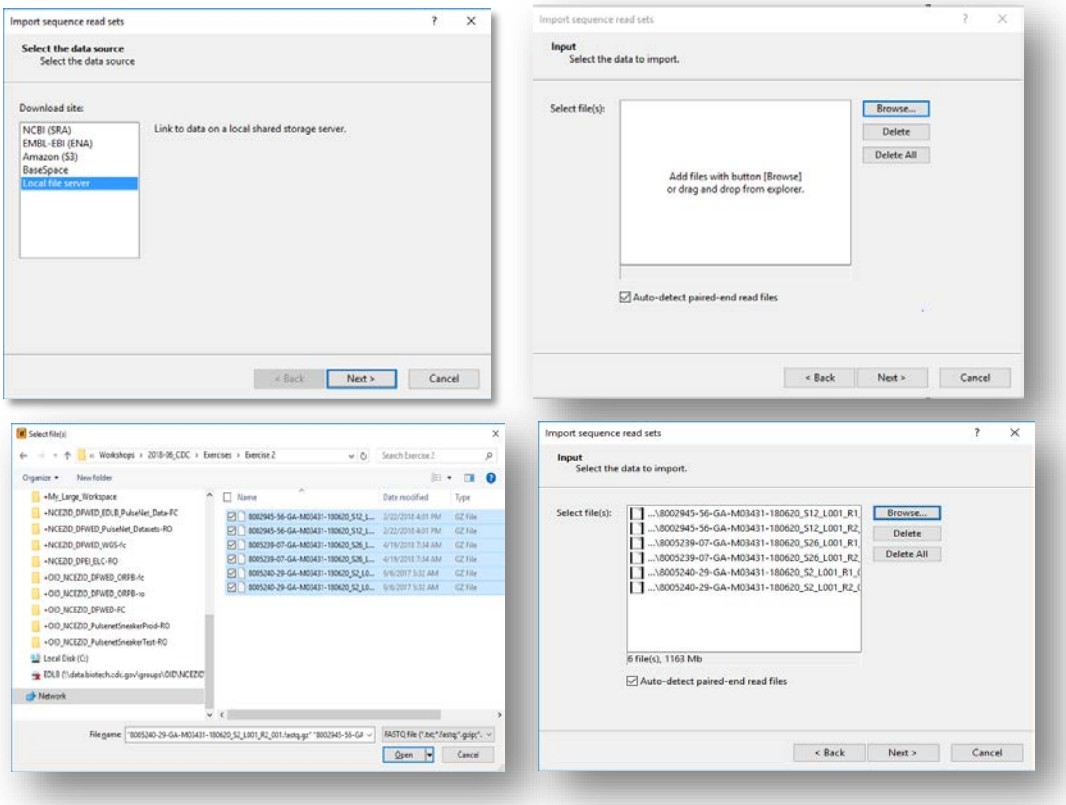
5.1.2. If fastq files are stored *locally* (desktop, local or network file share):

5.1.2.1. Select the “Local file server” option and click “Next”.

5.1.2.2. Click “Browse” in the next window and navigate to where your raw sequence data are stored.

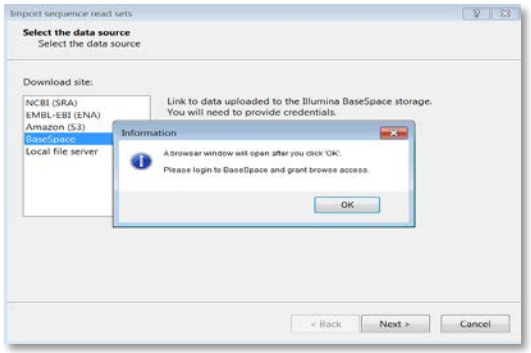
5.1.2.3. Select the sequence files (two files per sample) and click “Open”.

5.1.2.4. Confirm the selected files are listed in the “Select files” window, make sure that the box for “Auto-detect paired-end read files” is checked, then click “Next”.

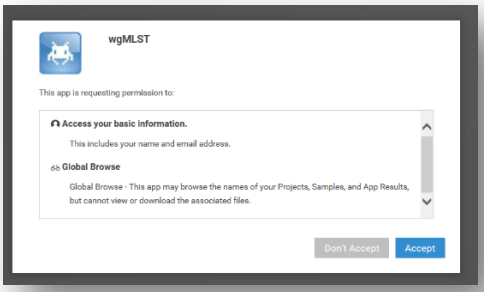


5.1.3. If fastq files are stored on *BaseSpace:*

5.1.3.1. Select the “BaseSpace” option, and click “Next” and choose “OK” to accept the message that your internet browser will open.



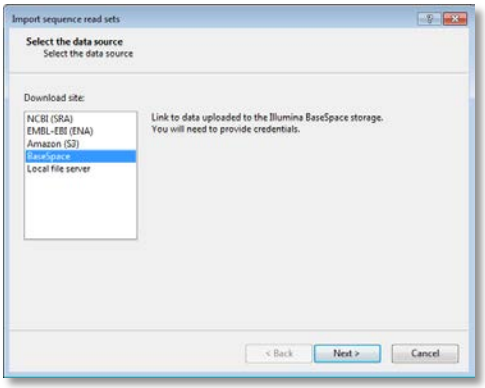
5.1.3.2. Sign in using your Illumina BaseSpace account credentials on the sign-in page. Accept the message that the wgMLST app is accessing your basic information. The webpage will then prompt you to close, choose "OK".

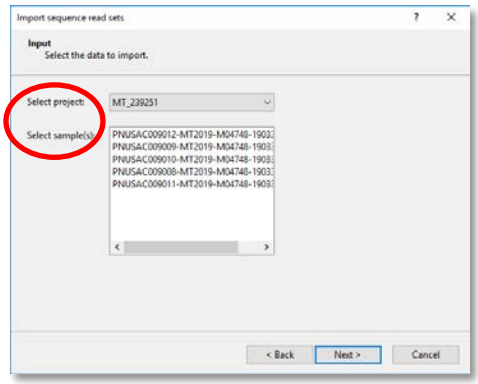


5.1.3.3. Once internet browser closes, choose "Next" in the "Import sequencing read sets" window.

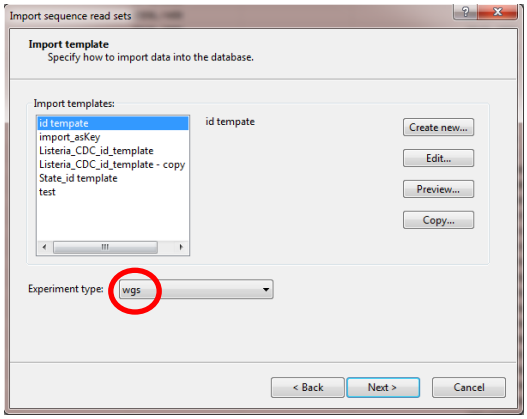
5.1.3.4. Select the project which contains the sequences for analysis.

5.1.3.5. Select the sequences (one file per sample) and choose "Next".





5.1.4. Select the appropriate import template on the next screen, confirm the Experiment type is "wgs", and click "Next >". **NOTE:** If an appropriate import template has not been created, click "Create new" on the right, and refer to Appendix PND20-1 to create one.

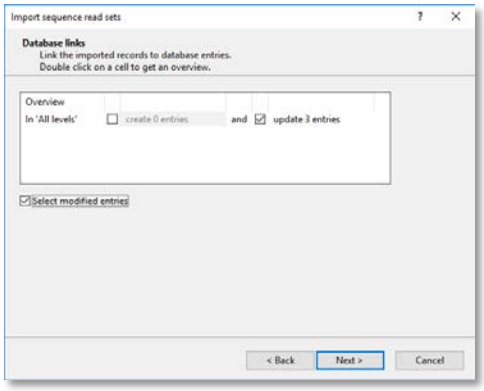


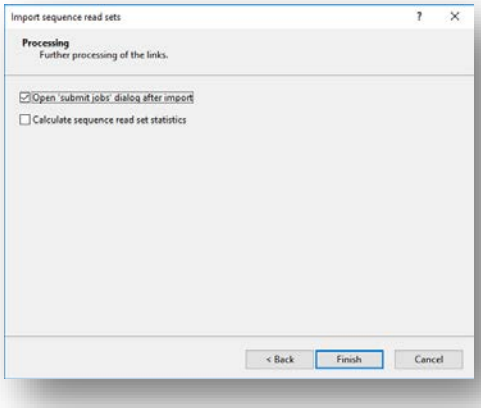
5.1.5. Review the actions that will be performed in the "Database links" window:

5.1.5.1. If the entry Key defined in the selected template already exists in the database, it will be updated; otherwise, a new entry will be created for each unique identifier not previously imported.

5.1.5.2. After choosing "Next", de-select "Calculate read set statistics" in the "Processing" window. Running this may take a while on your computer and isn't necessary.

5.1.5.3. Click "Finish".



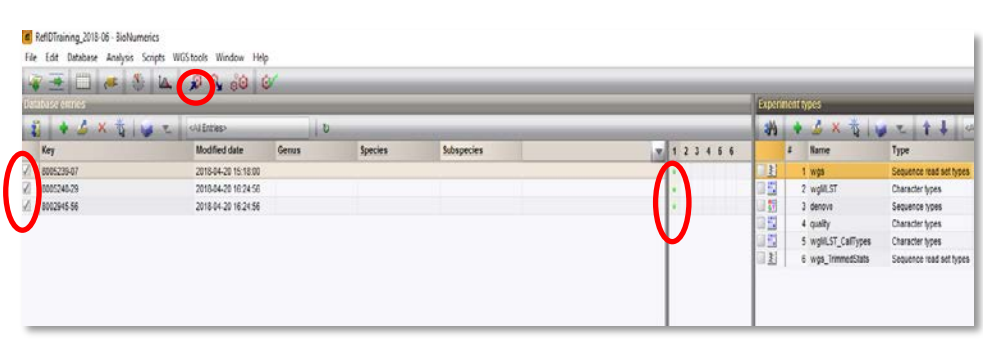


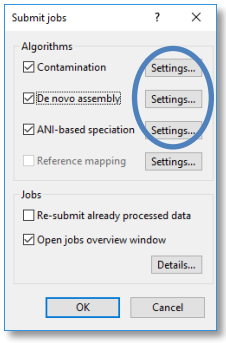
5.2. Submit jobs to the calculation engine (CE):

5.2.1. In order to access CDC resources you must authenticate to the PulseNet CE Firewall at https://PulseNetCE-USA.cdc.gov using the login credentials you received after passing the analysis certification.

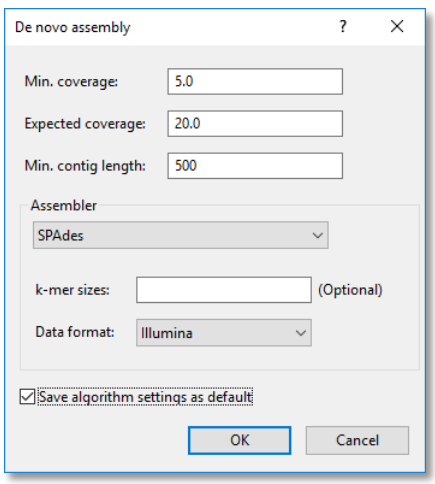
5.2.2. Select the database entries to be submitted for analysis by clicking the boxes on the left-hand side of the database window for each sample. Verify that each entry has a wgs link present in the database (green dots visible for wgs experiment indicate linked data).

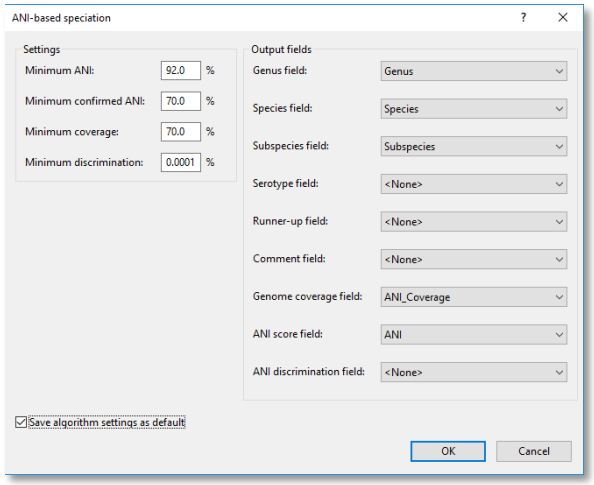
5.2.3. Click the "Submit jobs" icon,  make sure the boxes for "de novo assembly", "contamination check", and "ANI" are checked in the appearing "Submit jobs" window and click "OK".





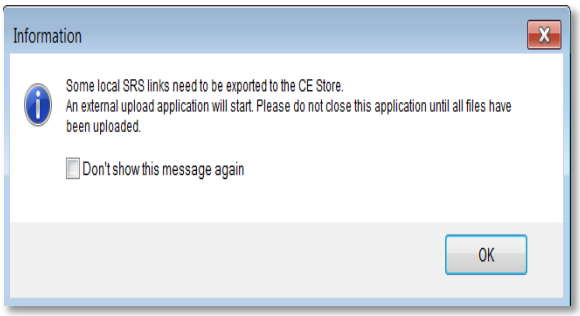
NOTE: Settings shown below for de novo assembly and ANI are correct and should not be modified once set. Contamination settings cannot be modified. Always check "Save algorithm settings as default" for any setting change.

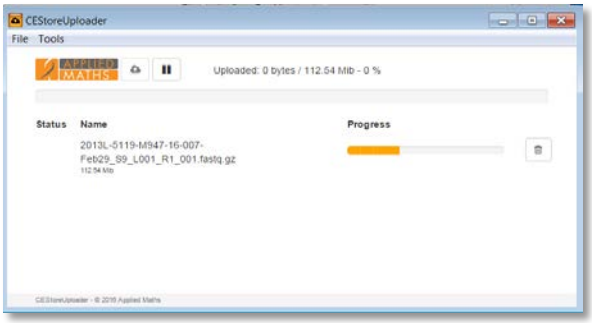




5.2.3.1. If sequence data is locally linked, an "Information" box for sequence data linked locally appears telling you the data will be exported to the CE Store and not to close the application until complete. Click "OK".

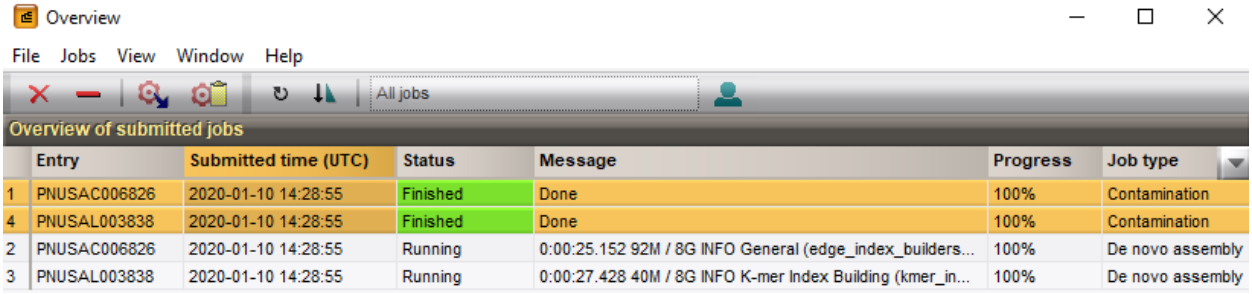
5.2.3.2. "CEStoreUploader" window appears showing progress of uploads. Keep this window open until the progress completes at 100%, at that point, you can close this window.





5.3. Check the "Status" column in the "Overview" window that appears and confirm that the de novo assembly and contamination check jobs are in the queue. The BioNumerics program can be closed at this point. **NOTE**: the "Overview" window automatically pops up when samples are submitted to the Calculation Engine but can be opened from the main BioNumerics window by clicking the triple-gear icon on the main toolbar and can be refreshed by clicking the "Refresh" icon in the "Overview" window.

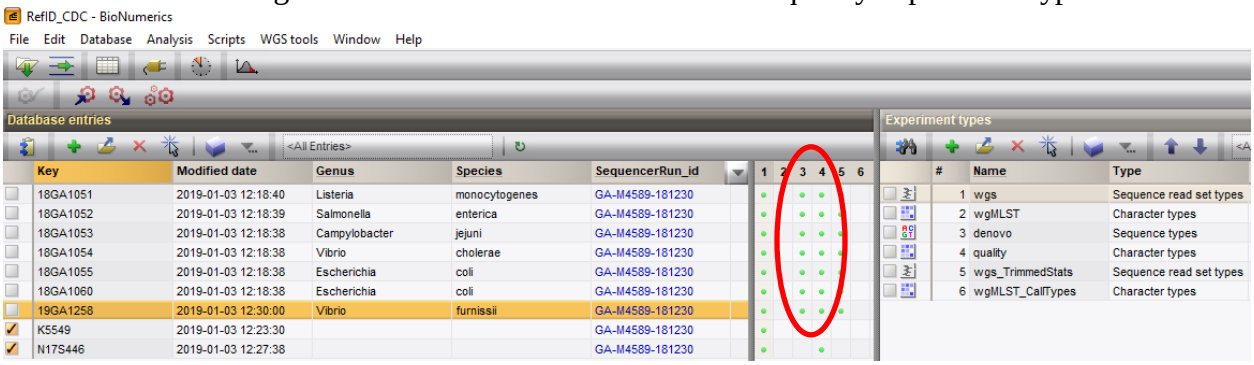
5.4. When the de novo and contamination jobs finish (status column will turn green) in the calculation engine, select/highlight the finished jobs using shift+click or ctrl+click and retrieve results by clicking the "Get the results for the selected job(s)" icon.



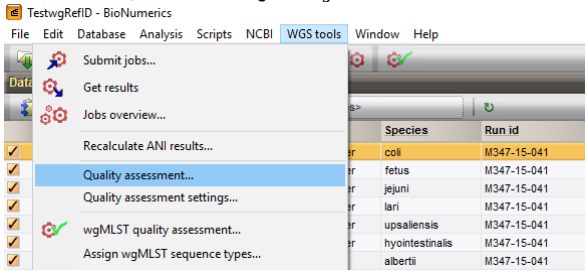
5.5. Now that you have an assembly for the genome, BioNumerics will automatically post your ANI job to the calculation engine. Refresh your "Overview" window to check the progress of that job and retrieve finished results.

5.6. Review sequence quality. **NOTE1:** the configuration of the "Quality Assessment" window is outlined in the appendix PND20-2. **NOTE2:** quality thresholds and detailed interpretation of the data are outlined in the SOP PNQ07 (PulseNet Standard operating procedure for Illumina sequence data quality control).

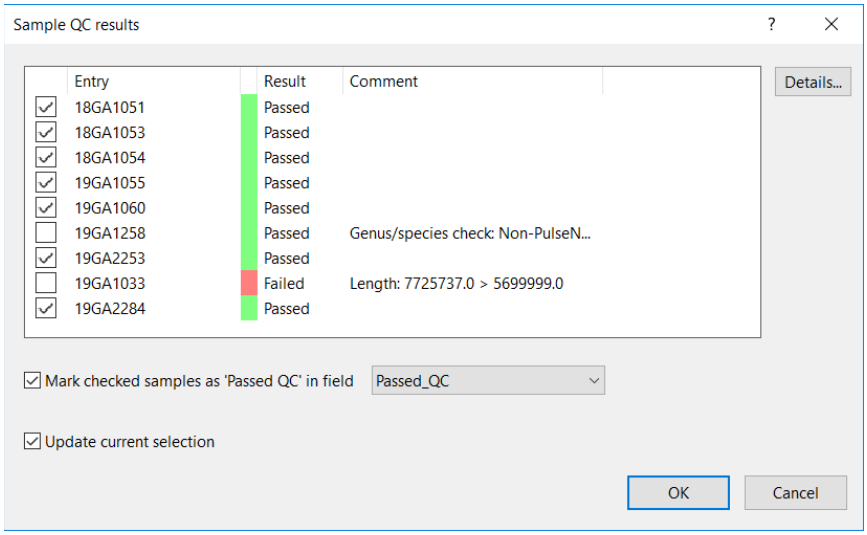
5.6.1. Select the database entries for which sequence quality is to be reviewed. Make sure there are green dots associated with de novo and quality experiment types.



5.6.2. In the "WGS tools" menu, select "Quality assessment".



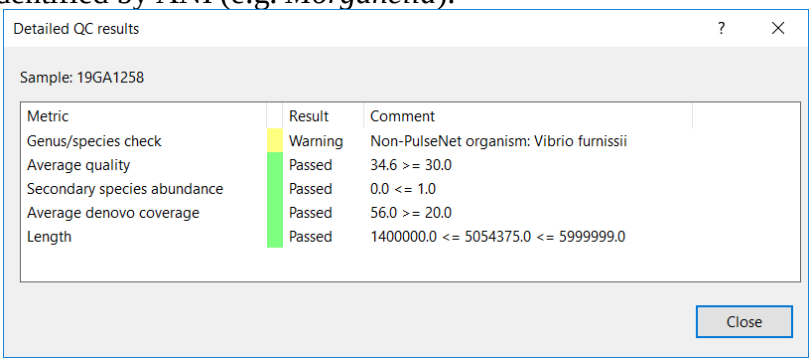
5.6.3. The appearing "Sample QC results" window lists the pass/fail results for critical quality metrics and comments for each selected entry. Select an entry and click on "Details" to see specific metrics that pass/fail for each entry.



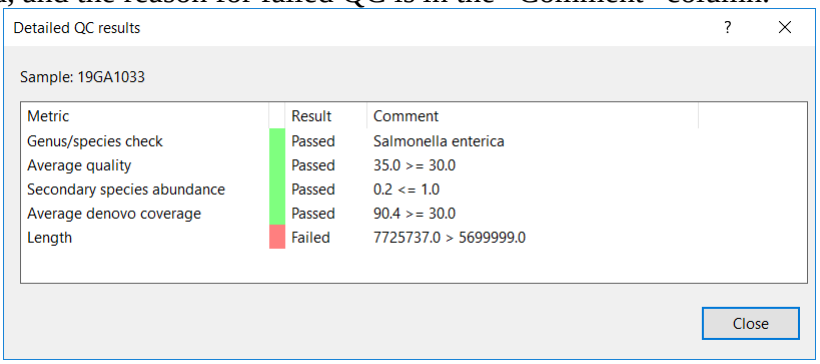
5.6.3.1. If the sequence is flagged green and is a PulseNet organism, the sequence passes quality, remains checked, and the comment is blank.

5.6.3.2. If the sequence is flagged green and is not a PulseNet organism, the sequence passes quality, becomes unchecked, and the comment indicates "Genus/species check: Non-PulseNet organism".

5.6.3.2.1. Click on "Details" to pull up a "Detailed QC results" window for all sequences that have comments listed. In the "Detailed QC results" window, the Genus/species check is flagged yellow if it was identified by ANI (e.g. non-PulseNet *Vibrio* or *Campylobacter* species) or red if it was not identified by ANI (e.g. *Morganella*).



5.6.3.3. If the sequence is flagged red, the sequence fails quality, becomes unchecked, and the reason for failed QC is in the "Comment" column.

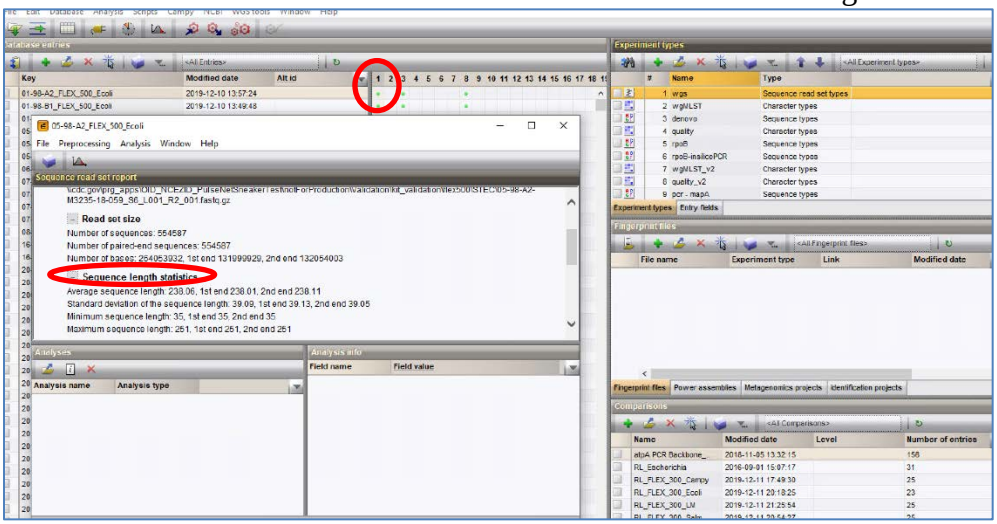


.

5.6.4. Review average read length (a non-critical quality metric):

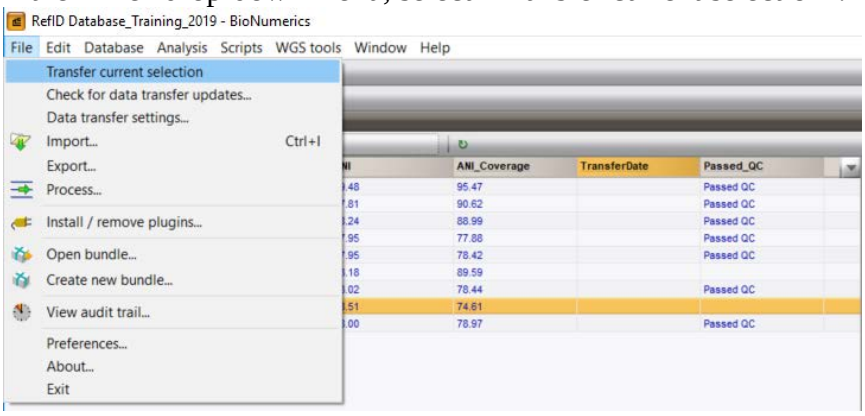
5.6.4.1. For each isolate, click on the green dot corresponding to the WGS experiment.

5.6.4.2. "Sequence read set information" window will open. Scroll down to "Sequence length statistics" where you can find average read lengths for the first and the second read and minimum and maximum read lengths.

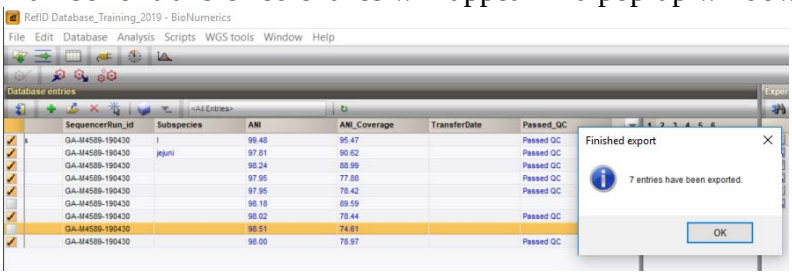


5.7. Transfer sequences that belong to PulseNet organisms and pass the critical quality metrics to organism-specific databases. **NOTE:** To configure data transfer settings, follow the instructions in PND20-3.

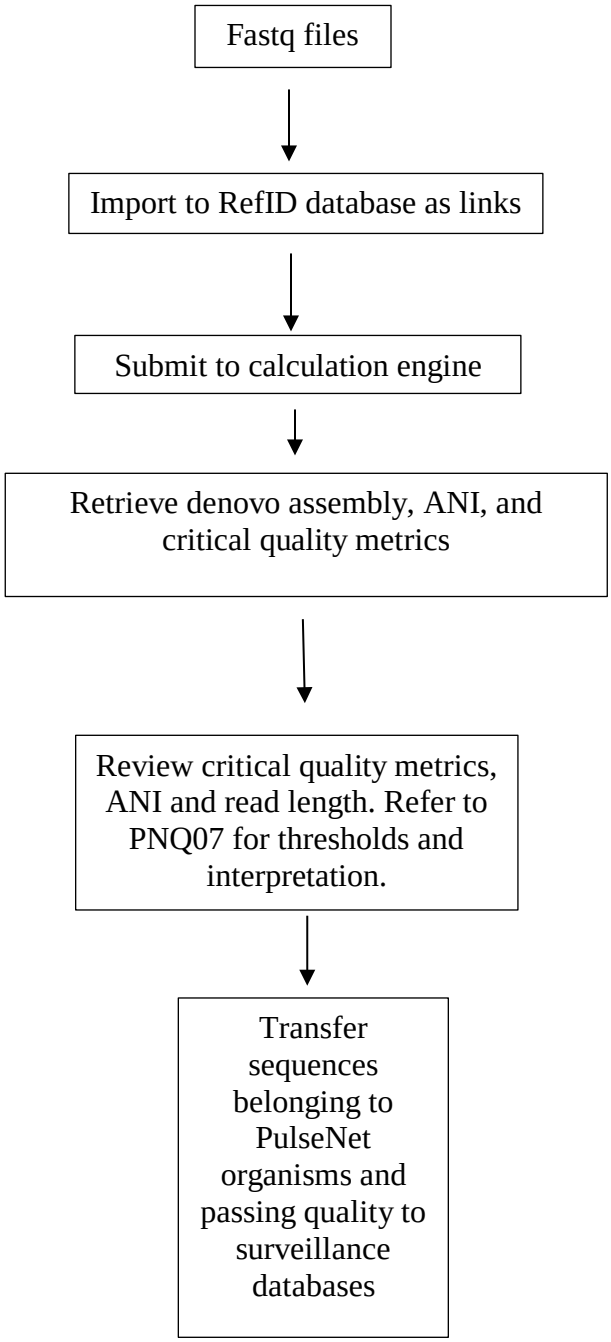
5.7.1. From the "File" drop-down menu, select "Transfer current selection".



5.7.2. The number of transferred entries will appear in a pop-up window. Click "OK".



6. FLOW CHART:


Fastq files
Import to RefID database as links
Submit to calculation engine
Retrieve denovo assembly, ANI, and critical quality metrics
Review critical quality metrics, ANI and read length. Refer to PNQ07 for thresholds and interpretation.
Transfer sequences belonging to PulseNet organisms and passing quality to surveillance databases

7. REFERENCES:

8. CONTACTS:

8.1. CDC PulseNet Database Team Inbox: PulseNet@cdc.gov

9. AMENDMENTS:

9.1. 02/06/2020 – New Document

10. APPROVAL SIGNATURES:

Approved By: Patricia C. Lafon -S (Affiliate) Digitally signed by Patricia C. Lafon -S (Affiliate) Date: 2020.02.09 10:23:33 -05'00' Date: 02/09/2020
PulseNet QA/QC Personnel

Approved By: Kelley B. Hise -S Digitally signed by Kelley B. Hise -S Date: 2020.02.07 11:51:08 -05'00' Date: 02/07/2020
PulseNet Outbreak Detection and Surveillance Unit Chief

Approved By: N/A Date: N/A
PulseNet PFGE Reference Unit Chief

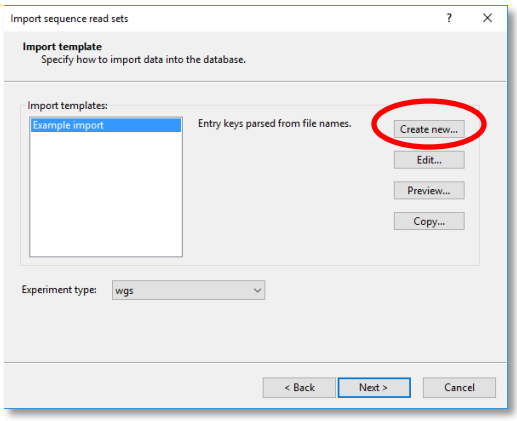
Approved By: Rebecca L. Lindsey -S Digitally signed by Rebecca L. Lindsey -S Date: 2020.02.07 14:10:21 -05'00' Date: 02/07/2020
PulseNet Next Generation Subtyping Methods Unit Chief

Approved By: Efrain M. Ribot -S Digitally signed by Efrain M. Ribot -S Date: 2020.02.07 13:57:56 -05'00' Date: 02/07/2020
PulseNet Reference Outbreak Surveillance Team Lead

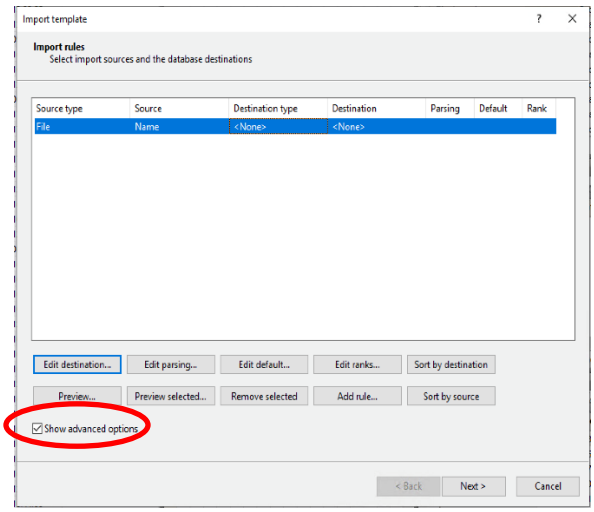
Appendix PND20-1. Create an Import Template in BioNumerics

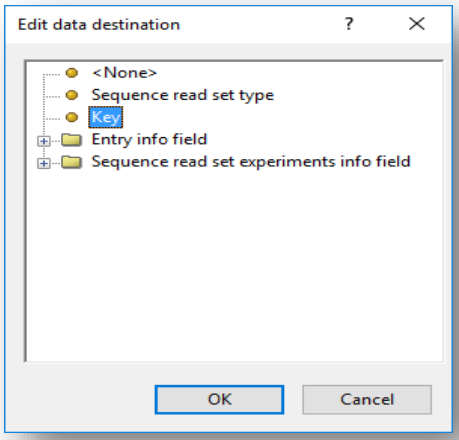
NOTE: these instructions assume that your sequence file name contains the PulseNet Key and that the "Key" in the RefID database is used to link the sequence data.

1. From the "Import Template" window, click the "Create new" button.

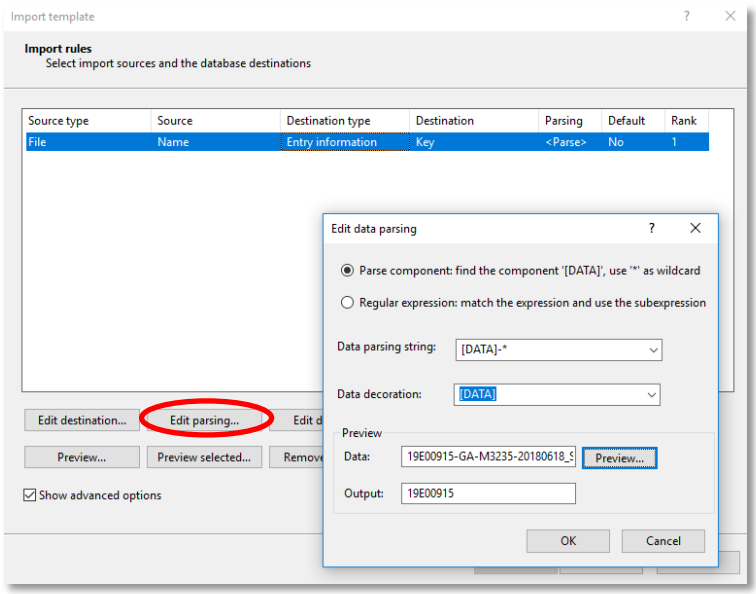


2. In the appearing "Import template/Import rules" window, check the "Show advanced options" box.
3. Highlight the File/Name line and click the "Edit destination" button. In the appearing "Edit data destination" pop-up, select the field for which the unique isolate identifier will be placed, in this case the "Key" field, then click "OK".



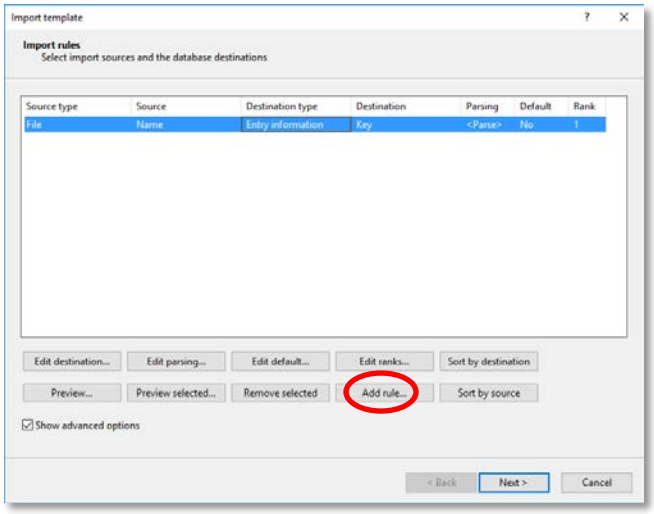


4. Click the "Edit parsing" button in order to parse out the PulseNet Key from the file name. The "Edit data parsing" window will appear.
    a. Change the content in the "Data parsing string" box such that any text in the file name of the sequences is cut off, leaving only the PulseNet Key where "[DATA]" is located in the string. In the case of sequence file name 19E00915-GA-M03431-180620_S12_L001_R1_001.fastq, adding **[DATA]-*** parses everything but the linking key name **19E00915** as shown below. **NOTE:** The pull-down menus for "Data parsing string" and "Data decoration" retain values previously used with other templates. You may select one that is applicable for your template or use a new one if desired.

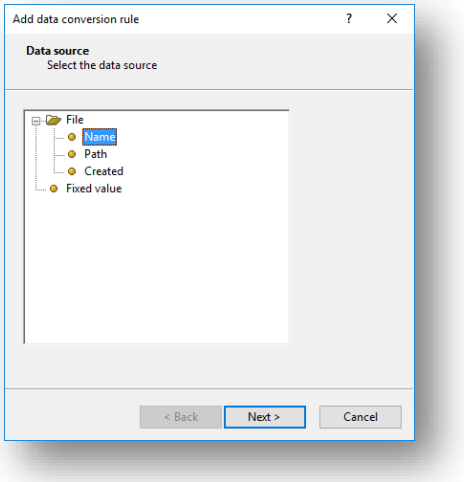


b. Click the "Preview" button to confirm the expected conversion from full file name to correct strain identifier.
c. Click "OK" when completed with parsing to be taken back to the main "Import templates/Import rules" window.

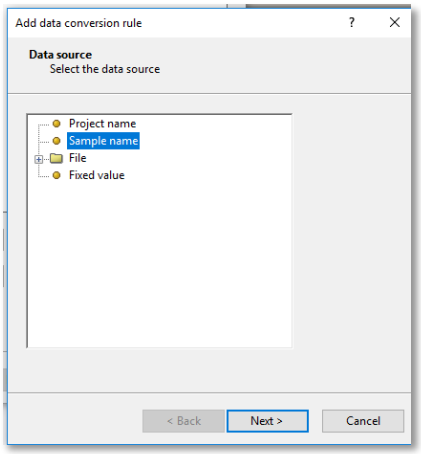
5. Once you have linked/parsed your Key out of the sequence file name, also parse the RunID.
   a. Select "Add rule".



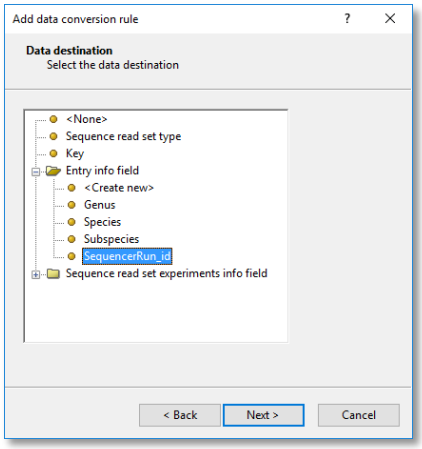
   b. "Add data conversion rule/Data source" window appears:
      i. If the sequence data is stored *locally*: expand "File", select "Name" and click "Next".



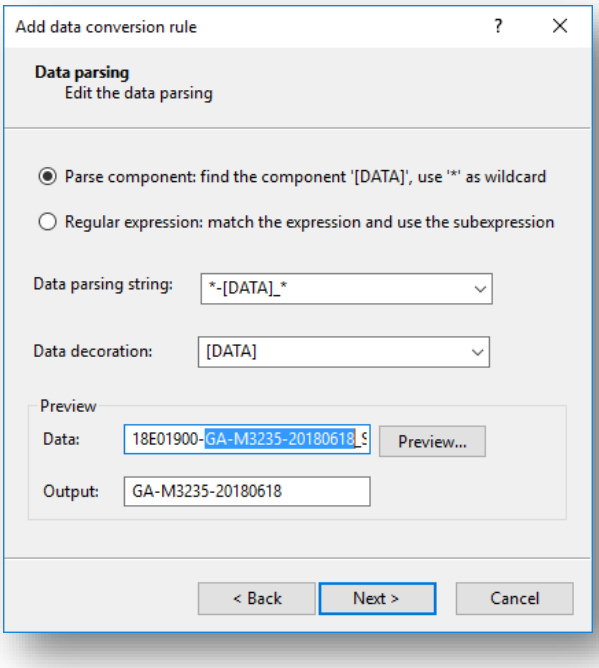
ii. If the sequence data is stored in *BaseSpace*: select "Sample name" and click "Next".



c. In the appearing "Add data conversion rule/Data destination" window, expand "Entry info field", select database field to parse data into: "SequencerRun_id" and click "Next".

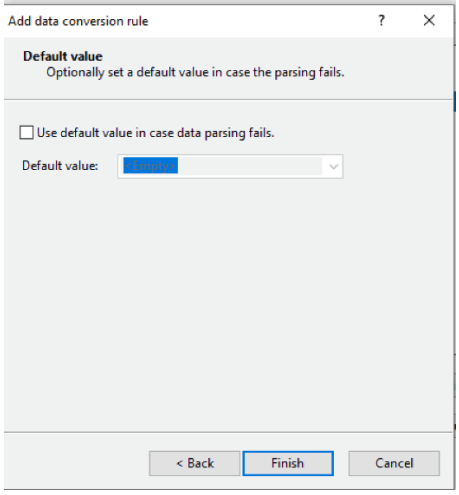


d. In the appearing "Add data conversion rule/Data parsing" window, set data parsing to remove the Key and additional MiSeq tagged data:
    i. If the sequence data is stored *locally*, ***-[DATA]_*** will remove anything before your first dash and everything after your underscore leaving you with the run id: **GA-M23235-20180618**

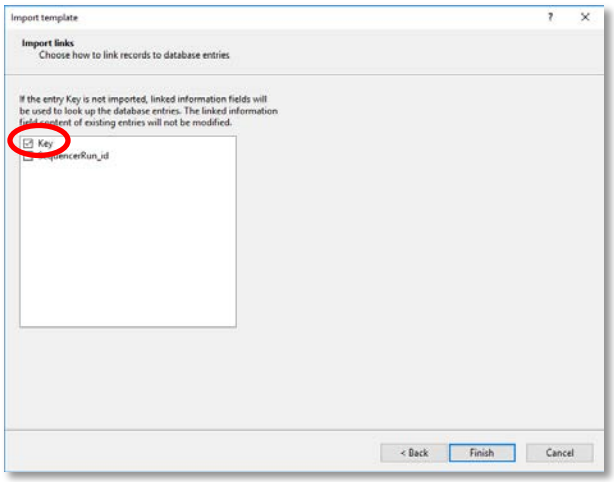


    ii. If the sequence data is stored in *BaseSpace*: an example of a parsing string for a file name like 12345-**GA-M70521-190524** is ***-[DATA]**

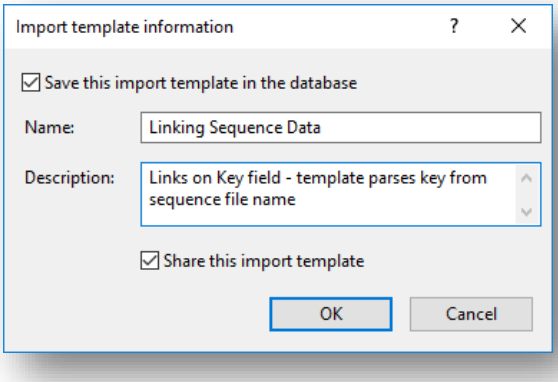
e. Click the "Preview" button to confirm the expected conversion from full file name to correct run identifier.
f. Click "Next" to be taken to "Add data conversion rule/Default value" window.
g. Leave "Use default value in case data parsing fails" unchecked and click "Finish". This will take you back to the "Import templates/Import rules" window.



6. Click “Next” to be taken to the “Import templates/Import links” window. Select the field, in this case the “Key” field, that will be used as a look-up for associating the parsed data above to a unique entry and click “Finish”.



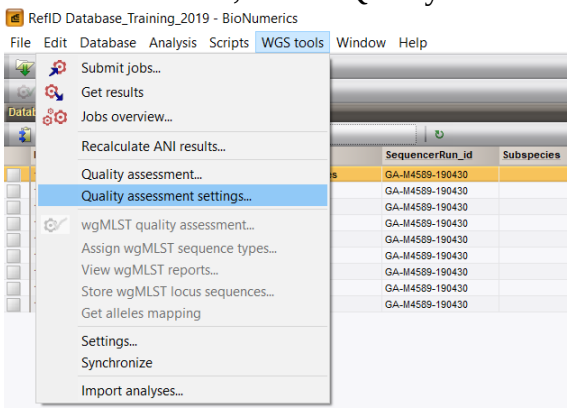
7. In the “Import template information” pop-up window, you can choose to save your template.
    a. Enter a name and description for this import template and keep “Save this import template in the database” and “Share this import template” checked.
    NOTE1: If you use a standardized naming scheme for sequence files, you can continue to use this saved template for all your sequence imports and parsing.
    NOTE2: Adding a description will help you remember what the template does, especially if you have multiple templates saved.



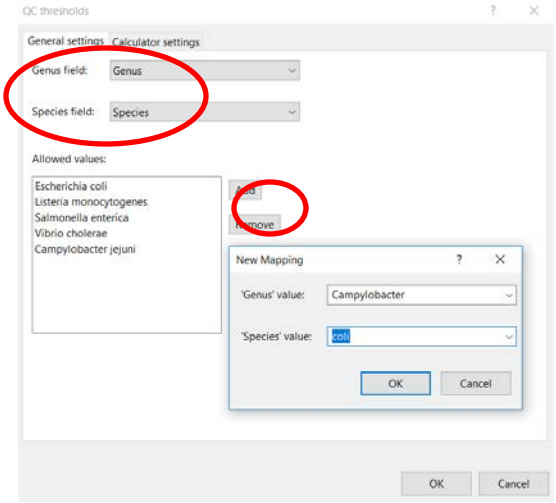
Appendix PND20-2. Configuration of the Quality Assessment Window in BioNumerics

NOTE: the quality assessment settings only need to be configured once.

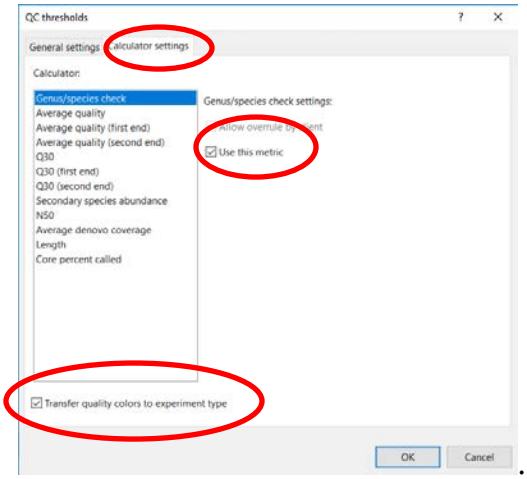
1. From the "WGS tools" menu, select" Quality assessment settings".



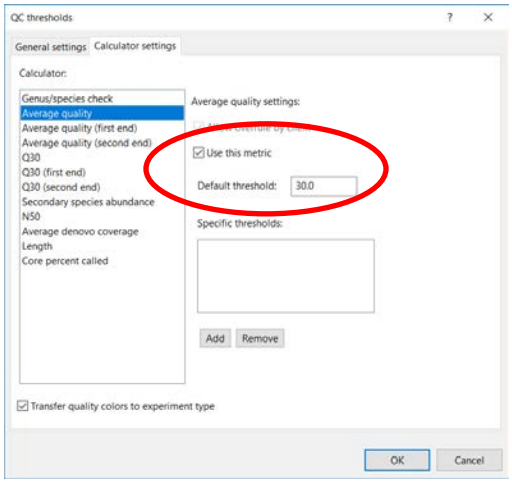
2. In the appearing "Quality thresholds" window & "General settings" tab, make sure the "Genus" and "Species" fields are set to "Genus" and "Species".
    a. Click "Add", enter the Genus values and Species values listed below one at a time and click "OK":
        i. *Campylobacter jejuni*
        ii. *Campylobacter coli*
        iii. *Campylobacter lari*
        iv. *Campylobacter upsaliensis*
        v. *Campylobacter fetus*
        vi. *Escherichia coli*
        vii. *Salmonella bongori*
        viii. *Salmonella enterica*
        ix. *Vibrio cholerae*
        x. *Vibrio parahaemolyticus*
        xi. *Vibrio vulnificus*
        xii. *Listeria monocytogenes*



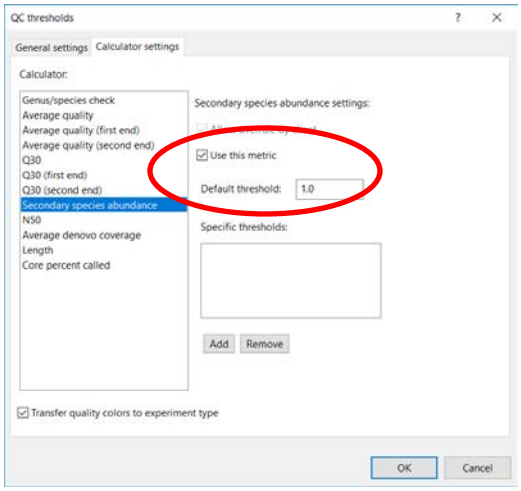
3. Select the "Calculator settings" tab and make sure "Transfer quality colors to experiment type" and "Use this metric" are checked for Genus/species check.



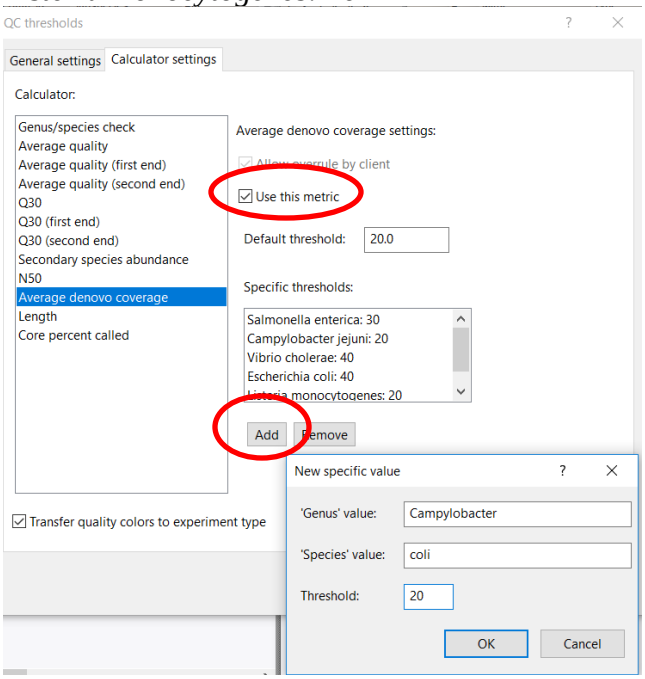
4. Select "Average quality" from the "Calculator" menu, make sure "Use this metric" is checked and enter "30.0" in the "Default threshold" field.



5. Select "Secondary species abundance" from the "Calculator" menu, make sure "Use this metric" is checked and enter "1.0" in the "Default threshold" field.

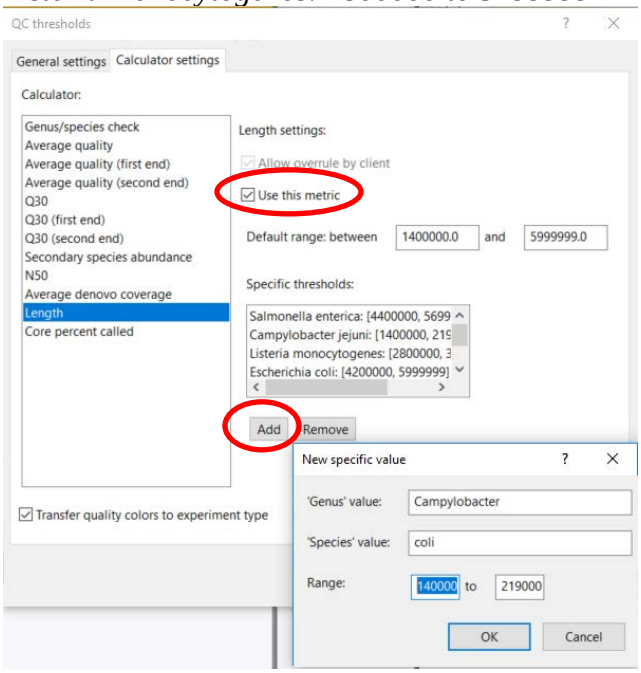


6. Select "Average denovo coverage" from the "Calculator" menu and make sure "Use this metric" is checked.
    a. Click "Add", enter Genus/Species values and corresponding coverage thresholds listed below one at a time and click "OK":
        i. *Campylobacter jejuni: 20*
        ii. *Campylobacter coli: 20*
        iii. *Campylobacter lari: 20*
        iv. *Campylobacter upsaliensis: 20*
        v. *Campylobacter fetus: 20*
        vi. *Escherichia coli: 40*
        vii. *Salmonella bongori: 30*
        viii. *Salmonella enterica: 30*
        ix. *Vibrio cholerae: 40*
        x. *Vibrio parahaemolyticus: 40*
        xi. *Vibrio vulnificus: 40*
        xii. *Listeria monocytogenes: 20*

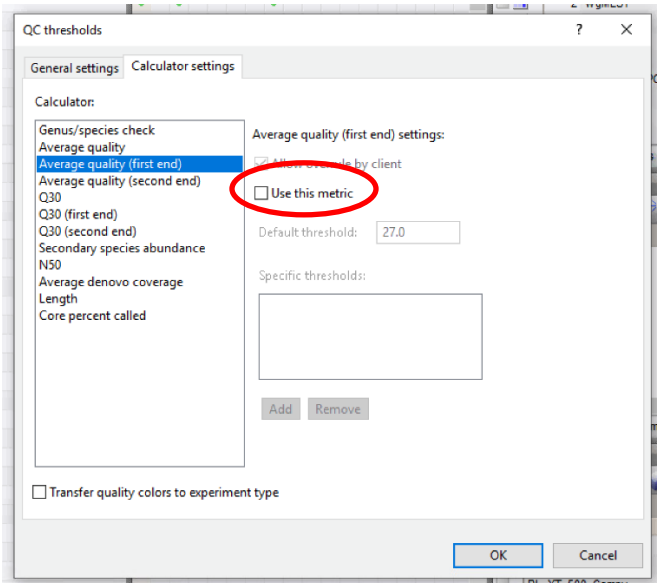


7. Select "Length" from the "Calculator" menu and make sure "Use this metric" is checked.
    a. Click "Add", enter Genus/Species values and corresponding length thresholds listed below one at a time and click "OK":
        i. *Campylobacter jejuni: 1400000 to 2199999*
        ii. *Campylobacter coli: 1400000 to 2199999*
        iii. *Campylobacter lari: 1400000 to 2199999*
        iv. *Campylobacter upsaliensis: 1400000 to 2199999*
        v. *Campylobacter fetus: 1400000 to 2199999*
        vi. *Escherichia coli: 4200000 to 5999999*
        vii. *Salmonella bongori: 4400000 to 5699999*

viii. *Salmonella enterica: 4400000 to 5699999*
ix. *Vibrio cholerae: 3800000 to 4299999*
x. *Vibrio parahaemolyticus: 4900000 to 5499999*
xi. *Vibrio vulnificus: 4700000 to 5299999*
xii. *Listeria monocytogenes: 2800000 to 3199999*



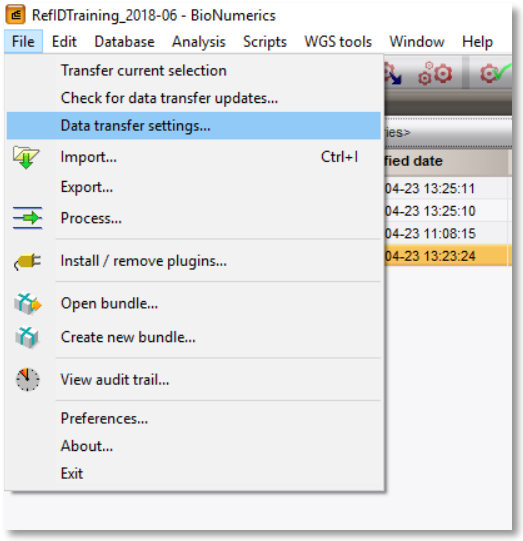
8. Manually uncheck "Use this metric" for "Average quality (first end)", "Average quality (second end)", "Q30", "Q30 (first end)", "Q30 (second end)", "N50", and "Core percent called".
9. Click "OK".



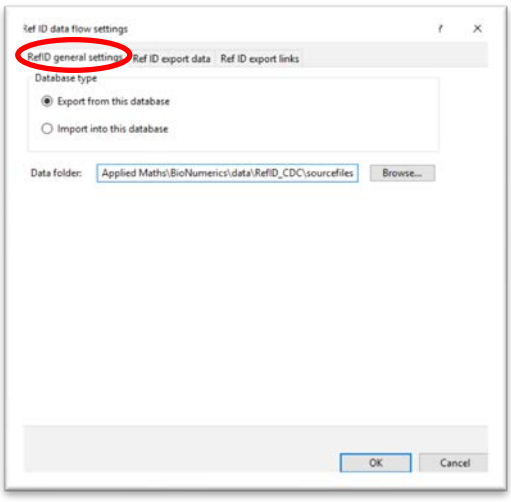
Appendix PND20-3. Configuration of the Data Transfer Settings.

NOTE: the data transfer settings only need to be configured once.

1. From the "File" drop-down menu, select "Data transfer settings". A "RefID data flow settings" window will open.



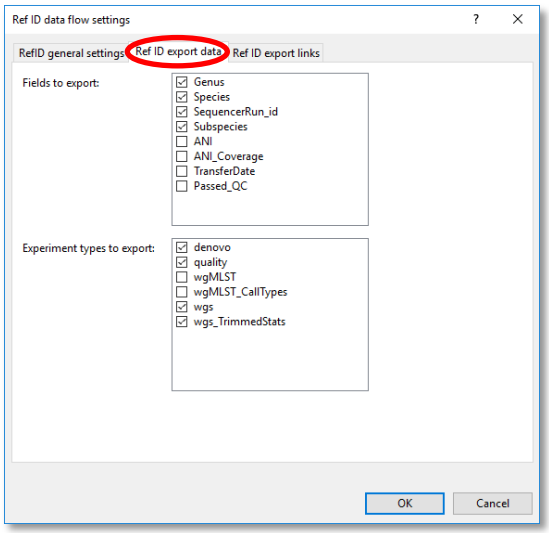
2. In the "RefID general settings" tab, choose "Export from this database". The data folder should be the source files location for your Reference ID database.



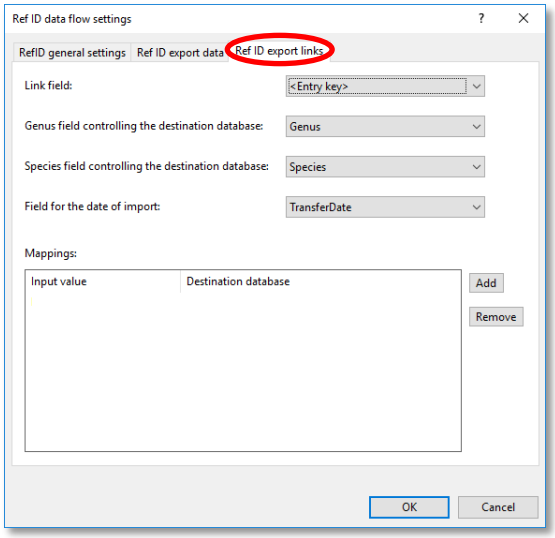
3. In the "RefID export data" tab, select the database fields and experiments to export.
    a. Fields:
        i. Genus
        ii. Species
        iii. SequencerRun_id
        iv. Subspecies

        NOTE: these instructions assume that metadata are added in the organism-specific surveillance database. If metadata are added in the RefID database then appropriate fields need to be checked in this step.
    b. Experiments:

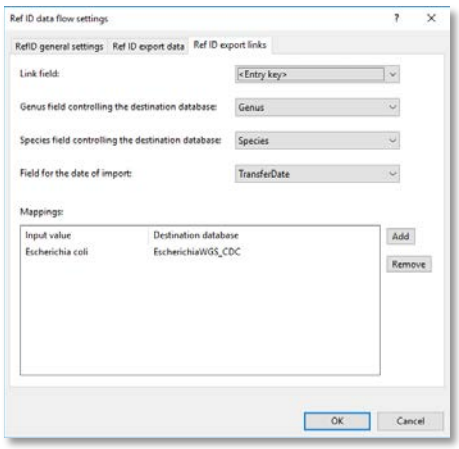
i. Denovo
ii. Quality
iii. Wgs
iv. Wgs_trimmedStats

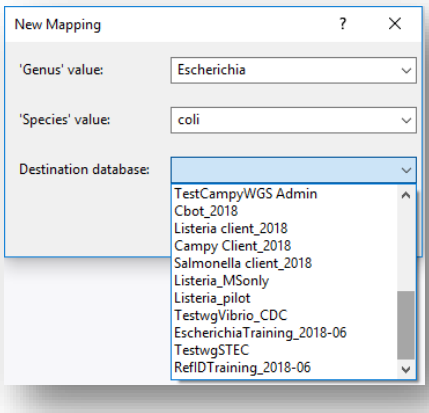


4. In the RefID export links" tab:
    a. Define the "Linking field" as "Entry Key".
    b. Define the "Genus field controlling the destination database" as "Genus".
    c. Define the "Species field controlling the destination database" as "Species".
    d. You can also choose to add the Transfer Date to your RefID database.



    e. Map the databases you will be importing to:
        i. Click "Add". A "New mapping" pop-up window will open.
        ii. Define values in the "Genus" and "Species" fields.
        iii. Select the destination database.
        iv. Click "OK".
        v. Repeat the procedure for each organism and database combination.


Ref ID data flow settings
RefID general settings
Ref ID export data
Ref ID export links
Link field:
<Entry key>
Genus field controlling the destination database:
Genus
Species field controlling the destination database:
Species
Field for the date of import:
TransferDate
Mappings:
Input value
Destination database
Escherichia coli
EscherichiaWGS_CDC
Add
Remove
OK
Cancel


New Mapping
'Genus' value:
Escherichia
'Species' value:
coli
Destination database:
TestCampyWGS Admin
Cbot_2018
Listeria client_2018
Campy Client_2018
Salmonella client_2018
Listeria_MSonly
Listeria_pilot
TestwgVibrio_CDC
EscherichiaTraining_2018-06
TestwgSTEC
RefIDTraining_2018-06