1. **PURPOSE:** Genomic sequences generated by PulseNet participating laboratories for PulseNet surveillance are processed using the BioNumerics software and a two-phase analysis approach. In the first phase, the raw sequence data is subjected to assembly, quality check and species identification in the RefID database. In the second phase, the sequence data is subjected to allele calling and genotyping in the organism-specific databases. The analyzed results and the metadata associated with each strain are uploaded to the national database and the fastq files are uploaded to NCBI. The purpose of this document is to describe a standardized procedure for the workflow in the organism-specific surveillance databases. The cluster detection procedures in the local organism-specific databases and matching against the national database are covered in the SOP PND06 (PulseNet standard operating procedure for local cluster detection and matching against the national database).

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

3. **TABLE OF CONTENTS – HYPERLINKS TO THE PROCEDURE**

4. **DEFINITIONS/ACRONYMS:**
   - **4.1. 7-gene MLST:** Multi-Locus Sequence Typing scheme based on seven conserved housekeeping genes.
   - **4.2. 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.
   - **4.3. Allele code**: A downloadable sequence type name, assigned in the national database, that is based on comparing cgMLST profiles using single linkage clustering and hierarchical similarity thresholds. Composed of multiple digits, e.g., LMO1.0 – 5.1.1.2.5.1 where

each digit corresponds to a specific similarity threshold (allele difference). The more digits two sequences have in common from left to right, the more closely related they are.

4.4. 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.

4.5. ANI: Average Nucleotide Identity. A computational method that compares an unknown query sequence to a reference sequence to determine the genetic similarity of shared sequences and proportion of bases aligned.

4.6. BaseSpace: Illumina cloud-based computing environment for next generation sequencing data analysis, management and storage, including data sharing.

4.7. BioNumerics: Analysis software used by PulseNet, developed by Applied Maths (Sint-Martens-Latem, Belgium).

4.8. Bioproject: A collection of biological data related to a single initiative, originating from a single organization or from a consortium.

4.9. Biosample: Submitter-supplied descriptive information (metadata) about the biological materials from which the NCBI stored data are derived.

4.10. CC: Clonal Complex; comprises of groups of related STs that are likely to derive from a common ancestor.

4.11. CDC: Centers for Disease Control and Prevention.

4.12. 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.

4.13. CGE: Center for Genomic Epidemiology (http://genomicepidemiology.org).

4.14. cgMLST: Core Genome Multi-Locus Sequence Typing.

4.15. Cluster (outbreak) code: A downloadable name assigned to a detected cluster of cases in the national database that populates the "Outbreak" field. Follows the format *year-month-location-organism-cluster number*. For example: 2005GAEXH-1 was the first STEC O157 cluster detected in Georgia in May 2020; 2005MLJPX-2 was the second multi-state *Salmonella* Typhimurium cluster detected in May 2020.

4.16. Core genome: Genes shared by a vast majority (e.g., >99% of strains *Listeria monocytogenes*) of strains of the same species.

4.17. Coverage: The average number of reads that include a given nucleotide in the reconstructed sequence. In BioNumerics, coverage is expressed as average denovo coverage which is the calculated after quality-based trimming.

4.18. Critical Quality Metrics: Average denovo coverage, average quality, 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.

4.19. De Novo Assembly: A sequence assembly generated from the short raw reads without the use of a reference genome.

4.20. FASTQ: A text-based file format for storing both sequence and its corresponding quality scores.

4.21. In silico PCR: Computational tools used to calculate theoretical PCR results using a given set of primer sequences to amplify DNA sequences from a sequenced genome.

4.22. 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 sequence reads.

4.23. Organism-Specific Database: A BioNumerics database, v 7.6 or higher, used for comparing isolate sequences for surveillance. Part of the standard PulseNet workflow.

4.24. PCR: Polymerase Chain Reaction.

4.25. PN: PulseNet.

4.26. PubMLST: A public database hosted by the University of Oxford (https://pubmlst.org/) that holds the 7-gene MLST reference sequences and lists of known STs/CCs for different organisms.

4.27. PulseNet Central: PulseNet team at CDC comprising of the Database Unit (PulseNet@cdc.gov) and the Next Generation Subtyping Methods Unit (PulseNetNGSlab@cdc.gov).

4.28. QC: Quality Control.

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

4.30. 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.

4.31. REP code: an eight-character, alpha-numeric code (ex.: REPEXH01) given to a strain that is identified being as Reoccurring, Emerging or Persisting

4.32. SAMN Number: The unique NCBI identifier (accession number) for the sequence biosample (metadata).

4.33. SeqSero: A pipeline for *Salmonella* serotype determination from raw sequencing reads or genome assemblies. A web app is available at www.denglab.info/SeqSero.

4.34. SOP: Standard Operating Procedure.

4.35. SPUID: Submitter Provided Unique Id. Each NCBI SRA sequence upload has to be accompanied by a unique ID. In the case a sequence has to be retracted from NCBI due to quality issues, the same SPUID cannot be used again upon re-upload.

4.36. SRR Number: The unique NCBI identifier (accession number) for the uploaded raw sequence reads.

4.37. ST: Sequence Type; an allelic profile typically based on 7-gene MLST. The profile can be matched against a public nomenclature database, such as PubMLST (https://pubmlst.org/), to assign a single numerical value.

4.38. WGS: Whole Genome Sequencing.

4.39. WGS_id: A downloadable anonymized identification number assigned to all PulseNet sequences as part of the national database upload process. Designed to anonymize the origin (patient, submitting laboratory) of the sequence in the public repository (NCBI SRA). Follows the format *PNUSA-organism acronym-6 digit number*. For example, PNUSAL000325 was the 325th *Listeria* strain uploaded to the national database.

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

5. RESPONSIBILITIES:

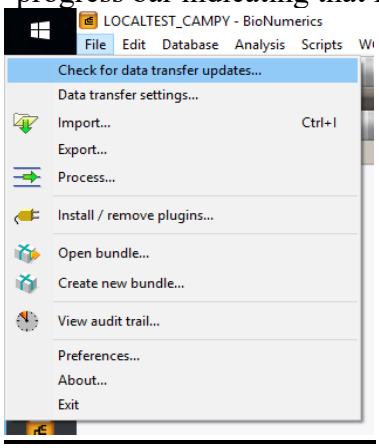
5.1. Analysis certified PulseNet public health laboratory personnel perform quality assessment, allele calling, genotyping, and upload of analyzed data and raw sequences to the PulseNet National Databases and NCBI using the BioNumerics 7.6 or higher organism-specific surveillance database workflow.

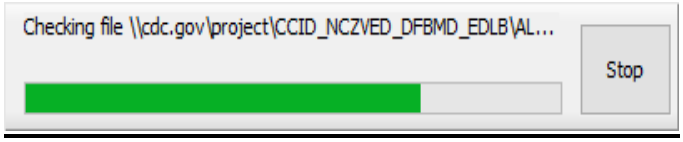
5.2. PulseNet Central Database Unit personnel perform additional quality control checks of the data submitted by the participating laboratories and request re-sequencing if necessary.

6. PROCEDURE:

6.1. Import Entries and Data from the RefID Database. NOTE: for the initial set-up and settings, follow the instructions in the Appendix PND05-1.

6.1.1. From the "File" drop-down menu, choose "Check for data transfer updates". A progress bar indicating that files are being checked will appear.





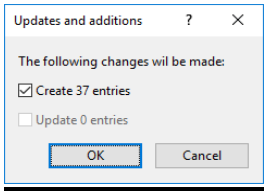
6.1.2. The "Updates and additions" window will appear.

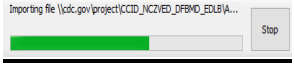
6.1.2.1. The number(s) in the window will indicate how many new entries will be created and/or how many existing entries will be updated. Click "OK".

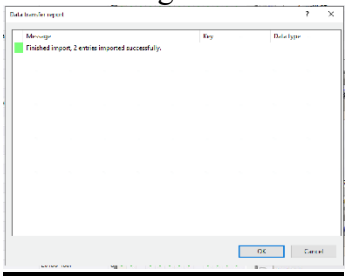
6.1.2.1.1. When importing new entries to the database, you should only see new entries created following this workflow. **NOTE:** you will see an entry updated if importing a repeated sequence to the database.

6.1.3. A progress bar indicating that files are being imported will appear.

6.1.4. The "Data transfer report" window will appear indicating success. Click "OK".







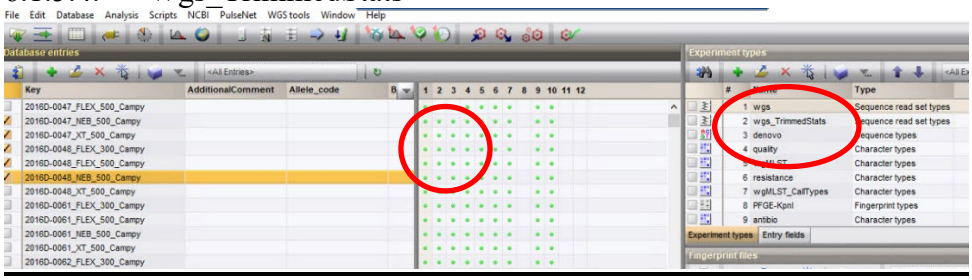
6.1.5. You should now see green dots in the columns associated with the following experiments:

6.1.5.1. Wgs

6.1.5.2. Denovo

6.1.5.3. Quality

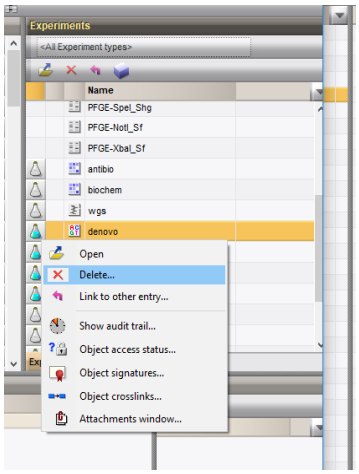
6.1.5.4. Wgs_TrimmedStats



6.1.6. An error may occur if you are importing sequence data to an entry that already has sequence data linked. Remove the existing links prior to the import:

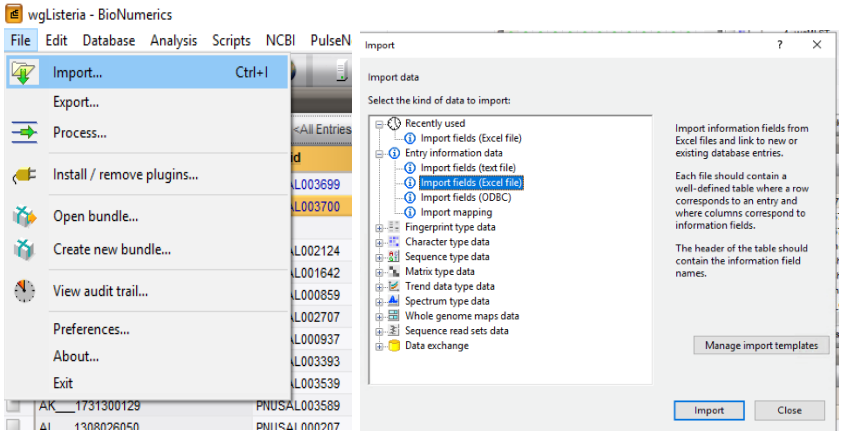
6.1.6.1. Double-click on the entry to open the "Entry edit" window.

6.1.6.2. Right-click the flask on "wgs" and "denovo" and select "Delete" from the drop-down menu.



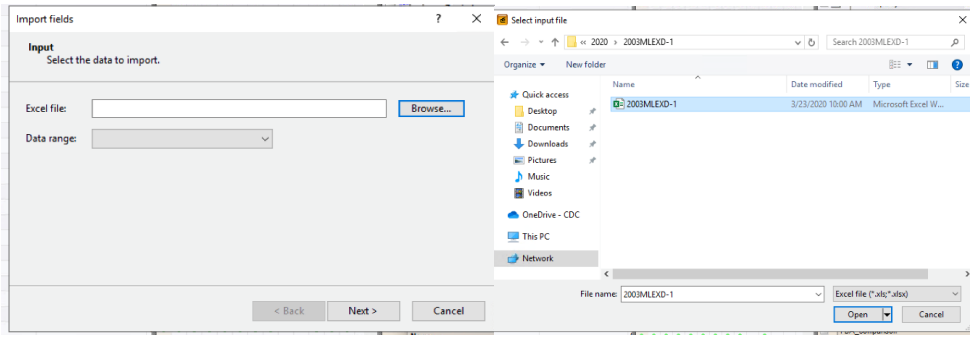
6.2. Import Metadata Associated with Each Strain Using an Excel File.

6.2.1. From the "File" drop-down menu, select "Import". An "Import" window will appear.

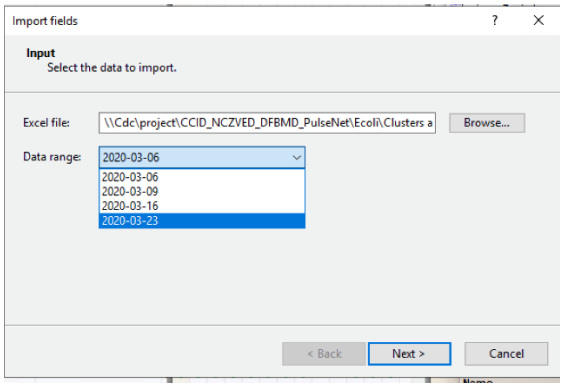


6.2.2. Expand "Entry information data", select "Import fields (Excel file)", and click "Import".

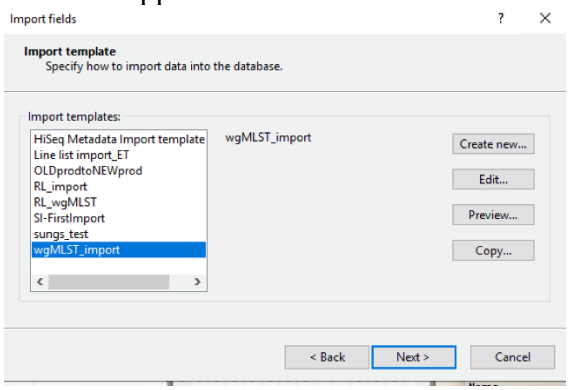
6.2.3. "Import fields" window will open. Click "Browse" to navigate to the location where the Excel file is saved, select the file and click "Open".



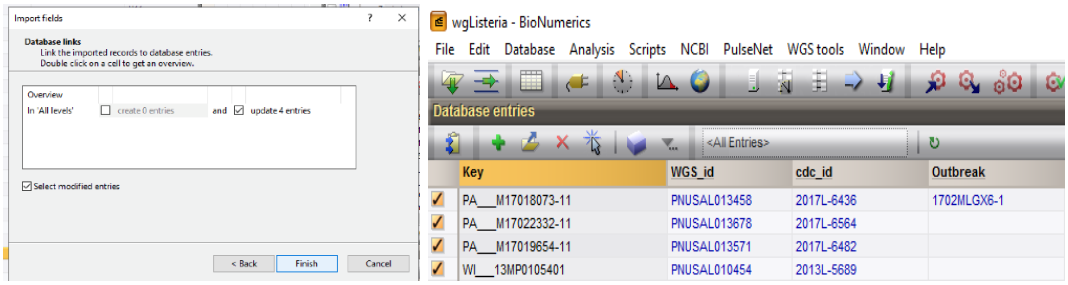
6.2.4. If the Excel file has multiple tabs, select the correct tab from the "Data range" drop-down menu and click "Next".



6.2.5. In the "Import fields/Import template" window, select the appropriate import template. Click "Next". **NOTE:** If you need to create a new template, follow the instructions in the Appendix PND05-2.



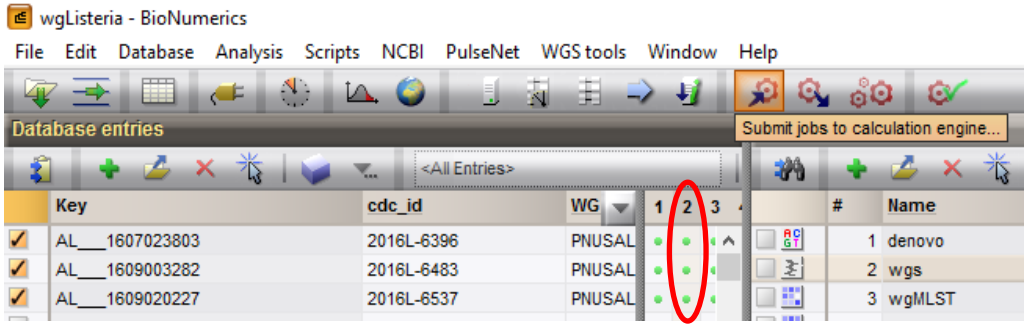
6.2.6. "Import fields/Database links" window will appear. Verify that the correct number of entries are being updated. Click "Finish". New updates are shown in blue in the database. **NOTE:** Entries should not be created as the metadata should be linked by key number to the sequence data already imported from the RefID database. If entries are created, then there is a mismatch of key numbers of sequences and metadata imported into the database.



6.3. Submit Jobs to the Calculation Engine (CE). NOTE: for the initial set-up and/or to verify the correct settings, follow the instructions in the Appendix PND05-3.

6.3.1. 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 (please see SOP PNQ08 for details).

6.3.2. 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 (green dot) present in the database.



6.3.3. Click the "Submit jobs" icon. A "Submit jobs" window will appear. **NOTE1: do not submit more than 40 sequences at a time**. **NOTE2:** the correct settings for each job are outlined in Appendix PND05-3.

6.3.3.1. The following boxes should be checked:

6.3.3.1.1. Assembly-based calls

6.3.3.1.2. Assembly-free calls

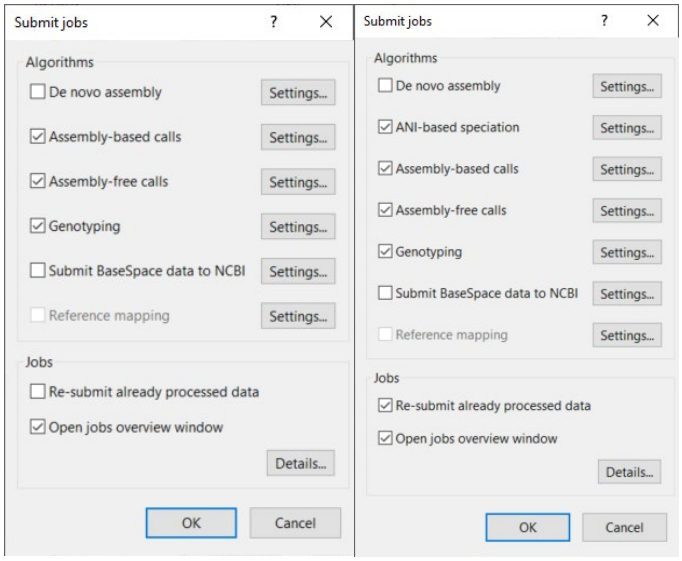
6.3.3.1.3. Genotyping

6.3.3.1.4. Open jobs overview window

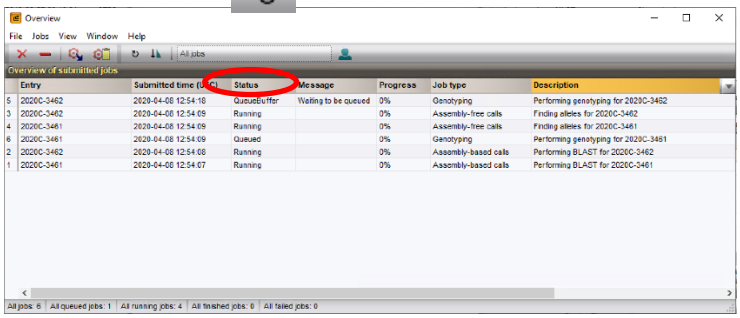
6.3.3.1.5. If you are submitting jobs for entries that have already gone through the CE before, check the box for "Re-submit already processed data".

6.3.3.1.6. If you want to determine the lineage for *Listeria*, check the box for "ANI-based speciation". In this case, you need to check the box for "Re-submit already processed data" as ANI had already been run in the RefID database.

6.3.3.2. Click "OK".

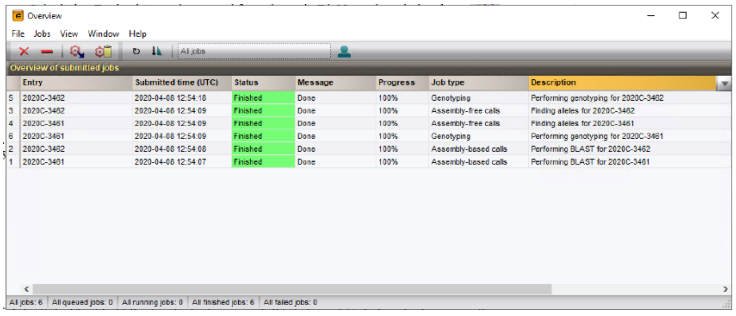


6.3.4. Check the "Status" column in the "Overview" window that appears and confirm that the assembly-based and assembly-free allele calling, and genotyping jobs are in the queue or running. The BioNumerics program can be closed at this point. **NOTE**: the "Overview" window automatically pops up when 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.

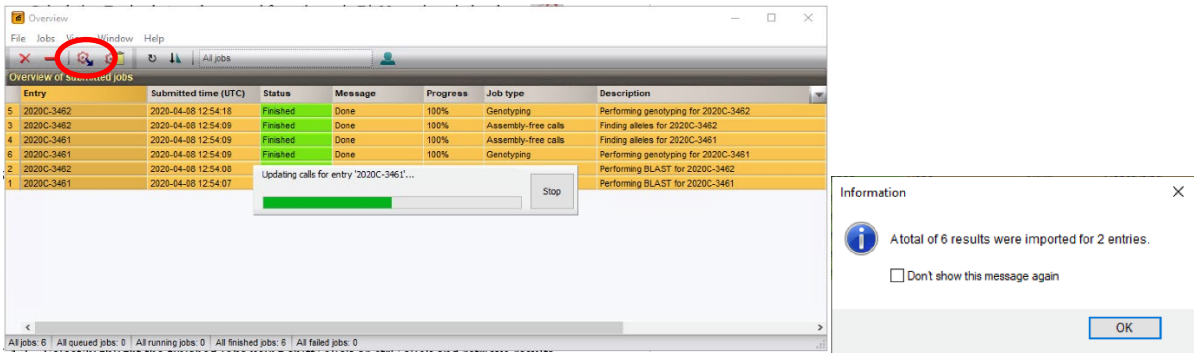


6.4. Retrieve the Finished Jobs from the Calculation Engine.

6.4.1. When the allele calling and genotyping jobs finish, status column will turn green in the "Overview" window.



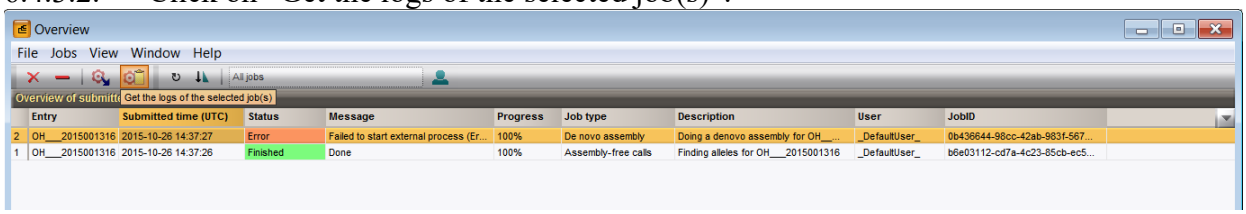
6.4.2. Select/highlight the finished jobs using shift+click or ctrl+click and retrieve results by clicking the "Get the results for the selected job(s)" icon. A progress bar will appear. When the download is finished, an information window will appear. Click "OK". **NOTE:** do not close the "Overview" window while retrieving the jobs.



6.4.3. For any jobs that fail, capture the error message.

6.4.3.1. Highlight jobs that failed.

6.4.3.2. Click on "Get the logs of the selected job(s)".

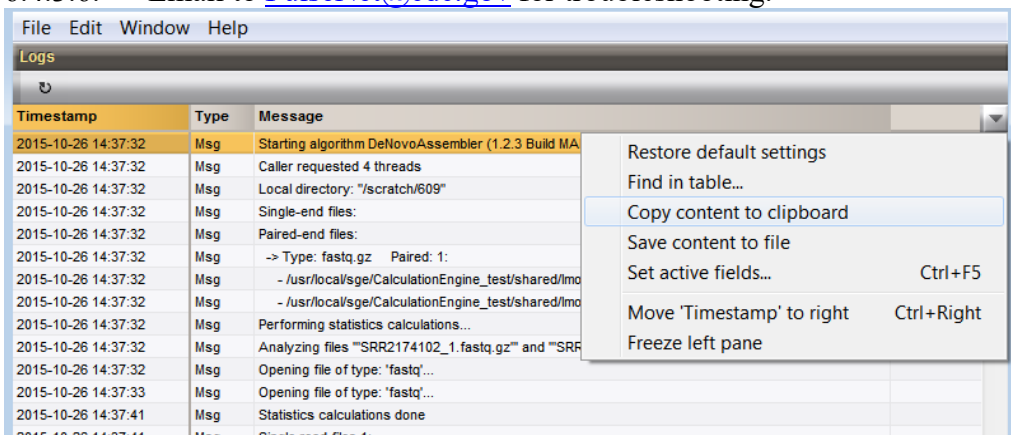


6.4.3.3. Click the pull-down arrow on the right.

6.4.3.4. Choose "Copy content to clipboard".

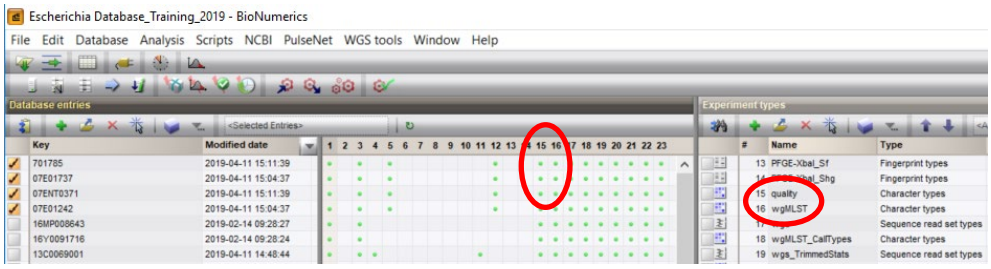
6.4.3.5. Paste into a new document (i.e., MS Word).

6.4.3.6. Email to PulseNet@cdc.gov for troubleshooting.



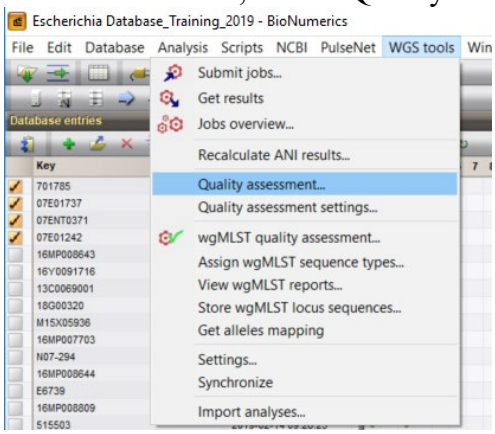
6.5. Review Quality Metrics. NOTE1: the configuration of the "Quality Assessment" window is outlined in the appendix PND05-4. **NOTE2:** quality thresholds and detailed interpretation of the data are outlined in the SOP PNQ07 (PulseNet standard operating procedure for Illumina sequence data quality control).

6.5.1. Select the database entries for which quality metrics are to be reviewed. Make sure there are green dots associated with quality and wgMLST experiment types.

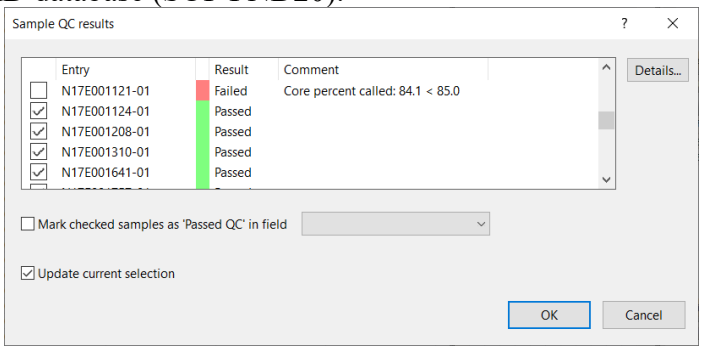


6.5.2. Viewing <u>critical quality metrics</u> one sequence at a time using the quality assessment window.

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



6.5.2.2. The appearing "Sample QC results" window lists the pass/fail results for core percent called and comments for each selected entry. Select an entry and click on "Details" to see specific metrics that pass/fail for each entry. **NOTE:** the other critical quality metrics (average quality, average denovo coverage, length and secondary species abundance) should have been reviewed in the RefID database (SOP PND20).

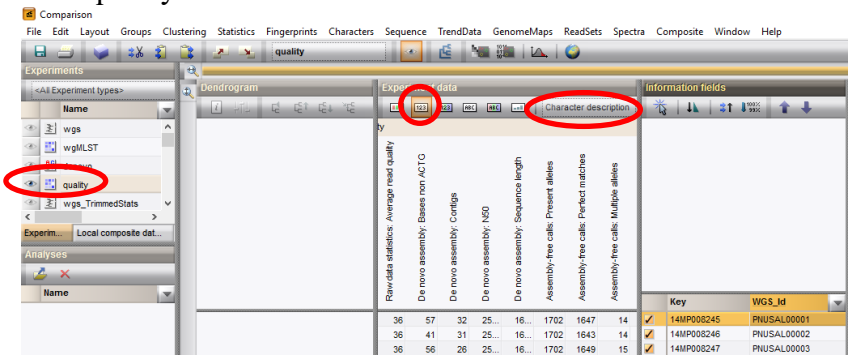


6.5.3. Viewing <u>all quality metrics</u> for multiple sequences using the comparison window.

6.5.3.1. Create a comparison by clicking on the green plus in the "Comparisons" panel in the main BioNumerics window or by pressing alt+c on the keyboard.

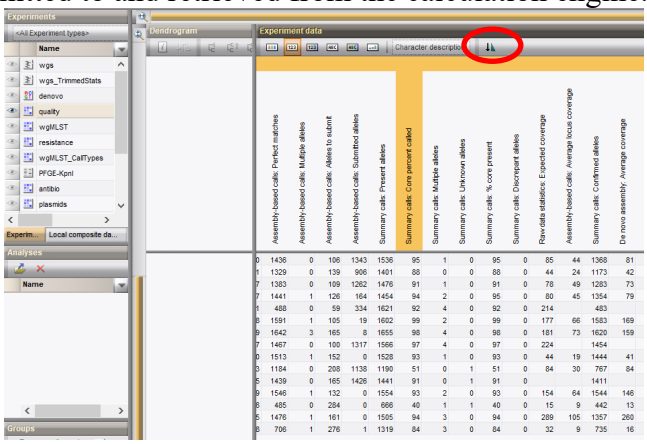
6.5.3.2. In the "Comparison" window, click on the eye next to the "quality" experiment in the "Experiments" panel. "Aspect" should remain "All characters".

6.5.3.3. Click on the first 123 icon ("Show values") and, if desired, change the "Character name" to "Character description" to view the full description of each quality metric.

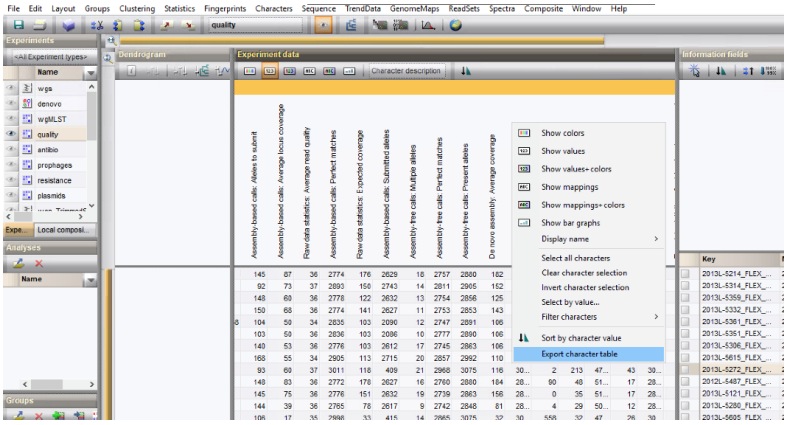


6.5.3.4. To quickly identify which sequences failed any particular quality metric, for example core percent called, highlight the quality metric in question and click on the "Sort by character value" button to sort the values from smallest to the largest.

NOTE: Also perform this step for an assembly-free metric (e.g., assembly-free calls: present alleles) and an assembly-based metric (e.g., assembly-based calls: present alleles) to ensure all sequences have values for these metrics indicating that the assembly-free and assembly-based allele calls have been submitted to and retrieved from the calculation engine.



6.5.3.5. If preferred, you can export the character table into Excel by right-clicking anywhere on the character table and choosing "Export character table" from the appearing drop-down menu. The character table will automatically open in Excel.



6.6. Review Genotyping Results.

6.6.1. *Salmonella* serotyping – the Genotyper tool populates the following two fields in the main window:

6.6.1.1. "AntigenForm_wgs": antigenic formula detected by SeqSero.

6.6.1.2. "Serotype_wgs": serotype name determined by comparing the antigenic formula to a look-up table containing the top 100 most common *Salmonella* serotypes from the Kauffmann-White-Le Minor scheme. **NOTE:** the "Serotype" field refers to the traditional serotype determined using conventional methods.

cdc_id	AntigenForm_wgs	Serotype_wgs
2015K-1205	I 4:d:1,7	Schwarzengrund
2015K-1206	I 4:d:1,7	Schwarzengrund
2015K-1207	I 4:d:1,7	Schwarzengrund
2015K-1208	I 4:d:1,7	Schwarzengrund
2015K-1209	I 4:d:1,7	Schwarzengrund
2015K-1210	I 4:d:1,7	Schwarzengrund
2015K-1211	I 9:g,m:-	Enteritidis
2015K-1212	I 4:d:1,7	Schwarzengrund

6.6.1.2.1. The field will be populated with "Needs further review" if the serotype is rare or requires further review by the CDC *Salmonella* reference lab.

6.6.1.2.1.1. If the sequence passes all quality metrics, you can look up the serotype name from the Kauffmann-White-Le Minor scheme at https://www.pasteur.fr/sites/default/files/veng_0.pdf. Manually replace "Needs further review" with the serotype name.

6.6.1.2.1.2. Alternatively, the sample can be uploaded to the national database with "Needs further review". The CDC *Salmonella* Reference Laboratory will perform the review, the "Serotype_wgs" field will be updated in the national database, and your laboratory will be notified via email by PulseNet Central that the "serotype_wgs" field has been updated in the national database.

Serotype_wgs	AntigenForm_wgs
Needs further review	I 3,10:z10:1,5
Needs further review	I 3,10:z10:1,5

6.6.2. *Escherichia* serotyping – the Genotyper tool populates the "Serotype_wgs" field with O and/or H antigens based on the CGE SerotypeFinder tool

(http://genomicepidemiology.org/). **NOTE1:** the "Serotype" field refers to the traditional serotype determined using conventional methods. **NOTE2:** O and H antigens are also detected for *Shigella* but do not correlate with traditional *Shigella* serotypes (e.g. 2a, 2b, 3a, etc).

WGS_id	Serotype	Serotype_wgs
PNUSAE017621	Shigella sonnei (Subgroup D)	:H16
PNUSAE017336	Shigella sonnei (Subgroup D)	:H16
PNUSAE021915	EAEC	O92:H33
06-3612	E. coli O118:H16	O118/O151:H16
06-3601	E. coli O177:NM	O177:H45
06-3555	E. coli O55:H7	O55:H7
PNUSAE011795	E. coli O118:H2	O118/O151:H2
PNUSAE015996	ETEC	O88:H38
PNUSAE020283	ETEC	:H34
PNUSAE021843	EAEC	:H33
PNUSAE021848	EAEC	O99:H4
PNUSAE018415	E. coli O145:HNT	O145:H25
2011C-3907	E. coli O104:Undetermined	O104:H4
PNUSAE016265	Shigella flexneri 1 [l:group anti...	O13/O135:H14

6.6.2.1. O antigen may not always be identified – the region encoding O antigen often sequences poorly even with overall high coverage, particularly with the Nextera XT library prep kit.

6.6.2.2. Multiple O antigens may be identified – some O antigen sequences are nearly identical and cannot be distinguished from each other. Some examples for STEC below:

6.6.2.2.1. O106/O17/O77:H45
6.6.2.2.2. O17/O44/O77:H18
6.6.2.2.3. O169/O183:H41
6.6.2.2.4. O123/O186:H2
6.6.2.2.5. O123/O186:H11
6.6.2.2.6. O118/O151:H16

6.6.2.3. Sequences with missing O antigens or multiple O antigens can be uploaded to the national database.

6.6.3. *Escherichia* pathotype – the Genotyper tool populates the "Pathotype" field in the main window based on the combination of genes detected by the VirulenceFinder tool (http://genomicepidemiology.org/). **NOTE:** two pathotypes can be reported out as a hybrid pathotype (e.g., STEC/EAEC) if genes from both pathotypes are present.

Pathotype abbreviation	***E. coli* Pathotype**	**Genetic markers**	**Gene combination**	**Gene function**
Shigella/EIEC	Shigella/Enteroinvasive	ipaH, ipaD, or icsA (virF)	One or more	Invasion plasmids
STEC (EHEC)	Shiga toxin-producing	stx1 or stx2	Any	Shiga toxin
EAEC	Enteroaggregative	aaiC, aggR, aatA, aafA	aaiC, aggR or aafA; aatA must also have one of the other three present	Secreted protein, regulator, mediated adherence, aggregative adherence fimbriae
ETEC	Enterotoxigenic	cfa, est, eltA	cfa plus one or more toxin	colonization factors, heat stable toxin, heat-labile toxin
EPEC	Enteropathogenic	eae, ehyl, pEAF	eae plus one of the other two	Intimin, hemolysin, plasmid encoded EPEC Adherence factor
DAEC	Diffusively adherent	afaC		Afimbrial adhesin

Table 1. Genes defining E. coli pathotypes

6.6.4. *Escherichia* stx subtype – the Genotyper tool populates the "Toxin_wgs" field in the main window based on the stx1 and stx2 subtypes detected by the in silico PCR tool. **NOTE:** There are known cross-reactivity problems with the detection of stx2d. If the stx2d subtype is detected together with other stx2 subtypes, the result is likely false positive.

Pathotype	Toxin_wgs
STEC	stx2b
STEC	stx1a
STEC	stx1a
STEC	stx1a
EIEC/Shigella	
EIEC/Shigella	
EIEC/Shigella	
EIEC/Shigella	
STEC	stx1a; stx2a; stx2c; stx2d
EIEC/Shigella	
EIEC/Shigella	
EIEC/Shigella	
STEC	stx1a; stx2c
STEC	stx1a
STEC	stx1a

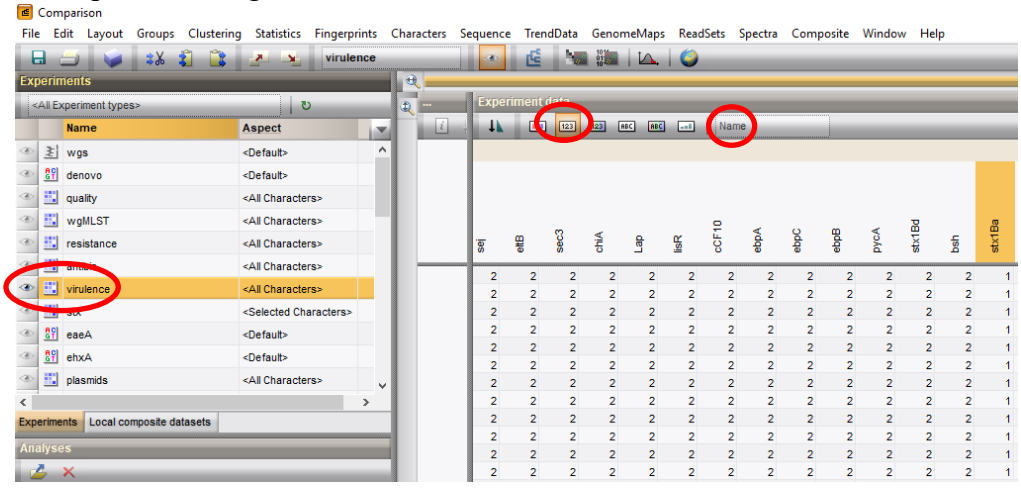
6.6.5. The experiments for resistance (all databases), plasmids (all databases), virulence (*Escherichia, Salmonella*) and stx (*Escherichia*; displays the 4 most important virulence factors for STEC) can be viewed in two ways:

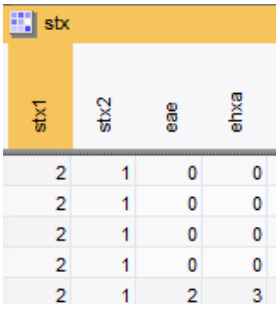
6.6.5.1. Multiple samples at the same time in a comparison:

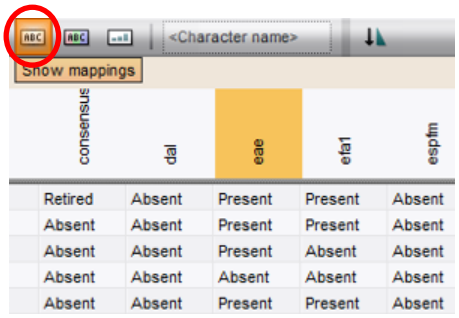
6.6.5.1.1. Select the entries and click on the green plus in the "Comparisons" panel in the main BioNumerics window or press alt+c on the keyboard.

6.6.5.1.2. In the comparison window, click on the eye next to the appropriate genotyping experiment in the "Experiments" panel. "Aspect" should remain "All characters".

6.6.5.1.3. Click on the first 123 icon ("Show values") to view the results in numeric format (1= gene present with >90% identity to the reference, 2= gene absent or <90% identity to the reference) or click on the ABC icon ("Show mappings") to view the results in Present/Absent format. For the resistance and virulence experiments, you can use the "Character name" drop-down menu to switch between the actual gene name and the description of the gene function.



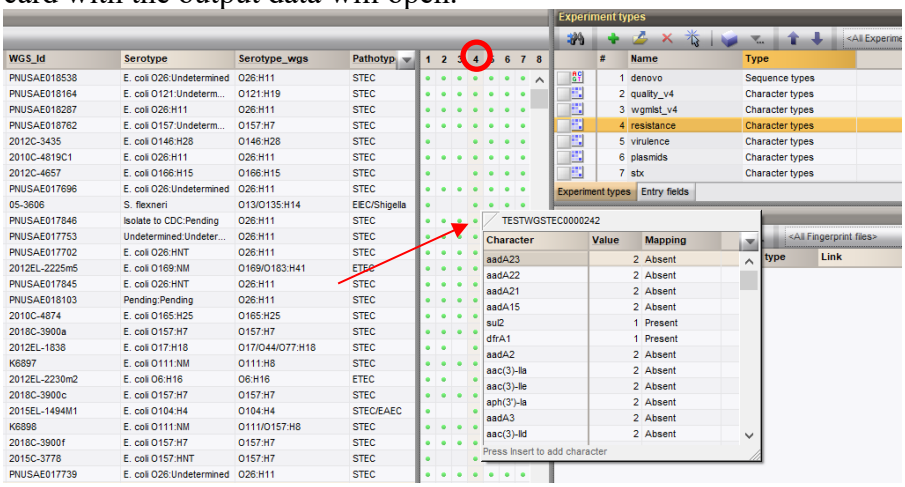




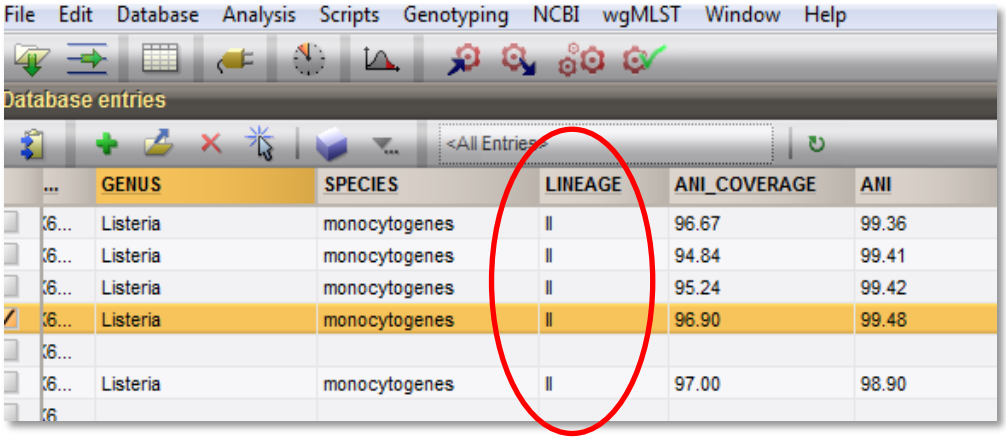
6.6.5.1.4. If desired, you can export the character table into Excel by right-clicking anywhere on the character table and choosing "Export character table" from the appearing drop-down menu.

6.6.5.2. <u>One sample at a time</u> in an experiment card:

6.6.5.2.1. For a given strain, click on the green dot in the main window associated with the appropriate genotyping experiment. An experiment card with the output data will open.



6.6.6. *Listeria* lineage – the Genotyper tool will populate the "Lineage" field in the main window if you submitted the job for ANI-based speciation.



6.7. Review 7-Gene MLST Results.

6.7.1. 7-gene MLST schemes available in different organism-specific databases:

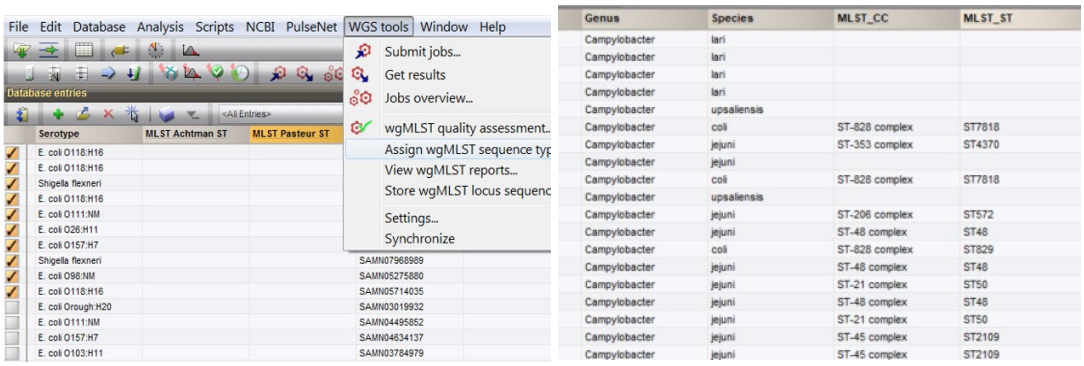
6.7.1.1. *Campylobacter:* Oxford schemes for *jejuni* (works also for *coli*), *lari, fetus*, *upsaliensis*

6.7.1.2. *Escherichia:* Achtman (used to populate the "MLST_ST" field), Pasteur, Whittam

6.7.1.3. *Listeria:* Pasteur (called "MLST" in the database)

6.7.1.4. *Salmonella:* Achtman

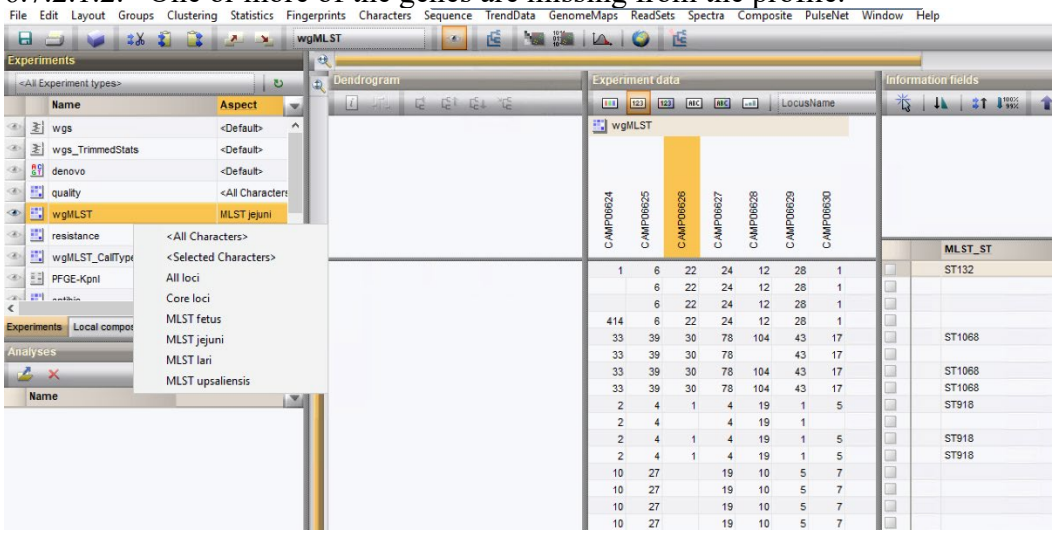
6.7.2. "MLST_ST" and/or "MLST_CC" fields will be automatically populated with the public ST nomenclature from the PubMLST database when allele call results are retrieved from the calculation engine. **NOTE:** you can manually assign the ST for historical sequences in the database by selecting the appropriate entries and then choosing "Assign wgMLST sequence types" from the "WGS tools" drop-down menu.



6.7.2.1. Unmatched/undetected sequence types are left blank. You can further investigate the reason for this by creating a comparison for the isolates with missing STs, highlighting the eye next to the "wgMLST" experiment in the "Experiments" panel and choosing the appropriate 7-gene MLST scheme (listed in 6.7.1.) from the "Aspect" drop-down menu. Now you see only the 7 housekeeping genes. The ST is usually left blank for the following reasons:

6.7.2.1.1. All genes are present but the profile is new to the public database.

6.7.2.1.2. One or more of the genes are missing from the profile.

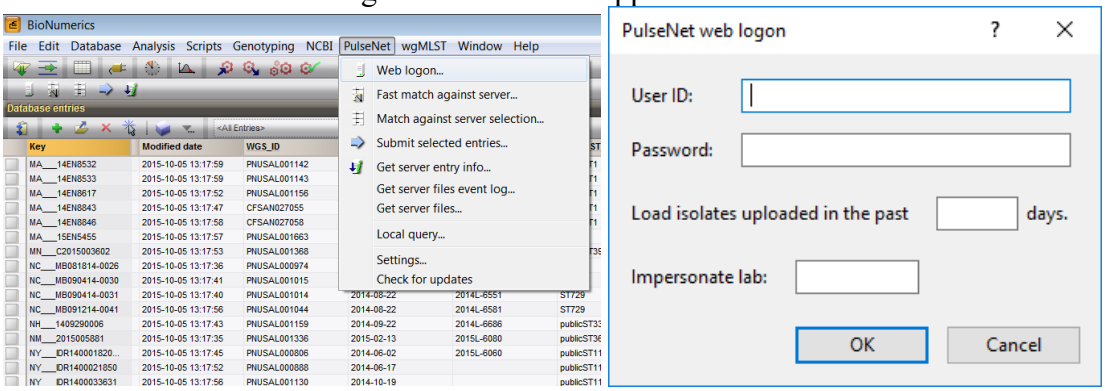


6.8. Upload to the National Database.

6.8.1. Authenticate to the PulseNetWGS firewall at https://PulseNetWGS-USA.cdc.gov using the login credentials you received after passing the analysis certification.

6.8.2. Connect to the national database.

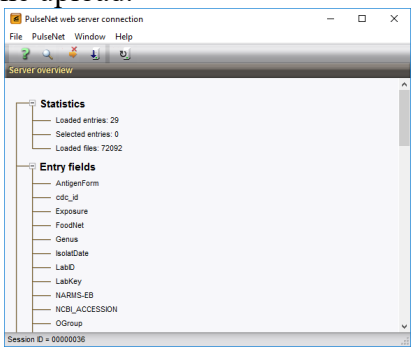
6.8.2.1. From the "PulseNet" drop-down menu, choose "Web logon". The "PulseNet web logon" window will appear.



6.8.2.2. Enter the user ID and password provided by CDC.

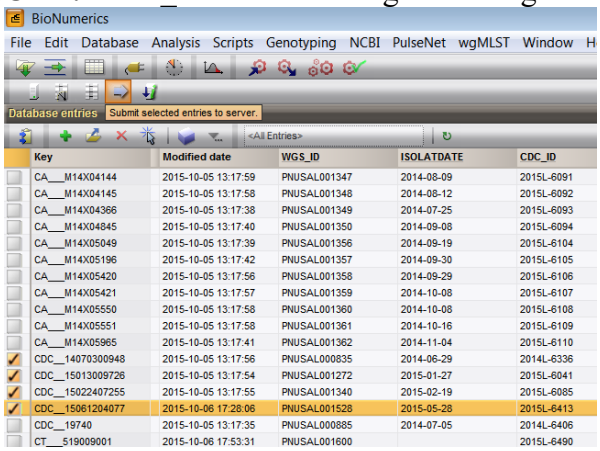
6.8.2.3. If you are only uploading (not searching), enter 0 for "Load isolates uploaded in the past".

6.8.2.4. Click "OK". The "PulseNet webserver connection" window will appear. **NOTE:** you can minimize the window, but do not close it until you have completed the upload.



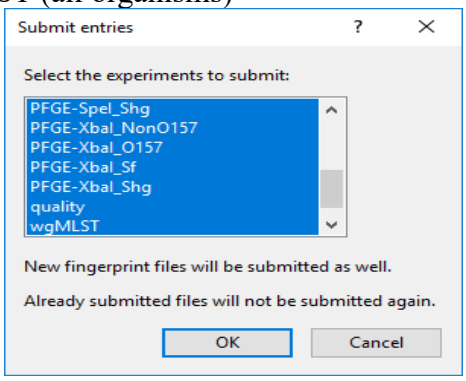
6.8.3. Select the entries to be uploaded to the national database.

6.8.4. Click "Submit selected entries to server". The "Submit entries" window will appear. **NOTE:** WGS_ids are auto-assigned during the upload process.

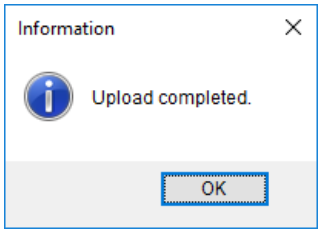


6.8.5. By default, all uploadable experiments are selected. De-select the experiments you are not uploading. **NOTE:** it is especially important to de-select the PFGE experiments if the strain already has confirmed PFGE patterns in the national database. Click "OK" after you are done. The experiments to be uploaded are:

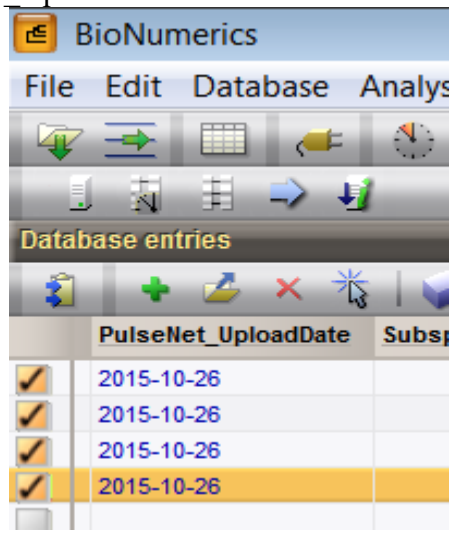
6.8.5.1. Quality (all organisms)
6.8.5.2. Antibio (all organisms)
6.8.5.3. Resistance (all organisms)
6.8.5.4. Virulence (*Escherichia*, *Salmonella*)
6.8.5.5. Stx (*Escherichia*)
6.8.5.6. Plasmids (all organisms)
6.8.5.7. wgMLST (all organisms)



6.8.6. Information window confirming that the upload was completed appears. Click "OK".



6.8.7. The "PulseNet_UploadDate" field in the main window should now be populated.



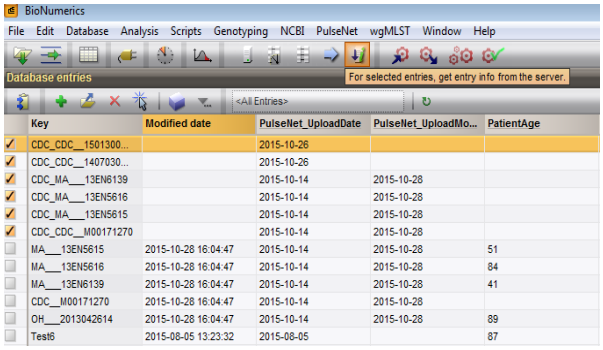
6.8.8. Close the "PulseNet web server connection" window. Click "Yes" on the appearing confirmation window.

6.9. Download Content for Information Fields (Outbreak Codes, Allele Codes, REP codes, WGS_ids) from the National Databases.

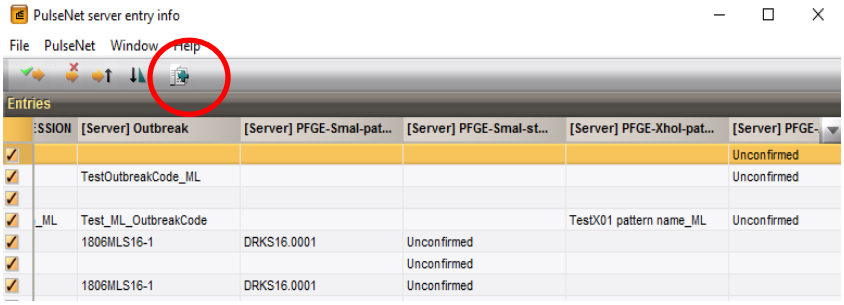
6.9.1. Authenticate to the PulseNetWGS firewall and connect to the national database. **NOTE:** the "Get entry info" functionality will download information from the entire database regardless of the number of days you choose to connect to.

6.9.2. Select the entries in the local database to get information from the national database.

6.9.3. Click on "For selected entries, get entry info from the server".

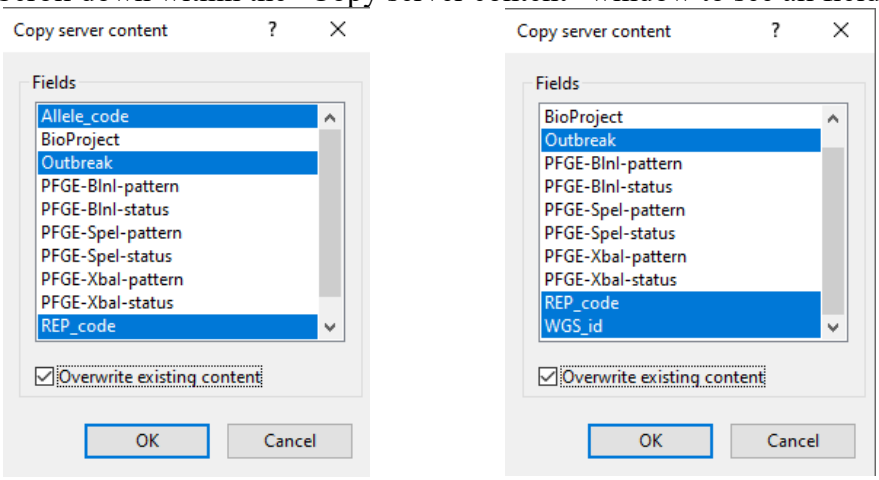


6.9.4. The "PulseNet server entry info" window will open. Click on "Copy server content for the selected entries".

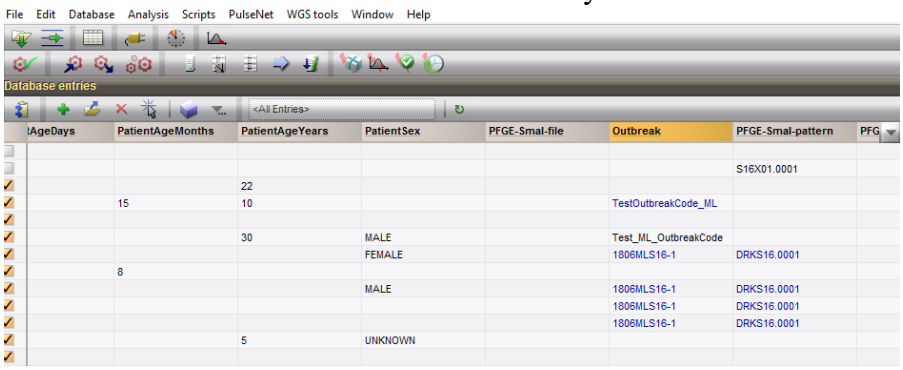


6.9.5. The "Copy server content" window will open. Choose the information fields you want to download, make sure the box for "Overwrite existing content" is checked and click "OK". **NOTE: this will overwrite any existing data you have in those fields in your local database.**

Scroll down within the "Copy server content" window to see all field options.



6.9.6. New information that was downloaded to your local database will be in blue.

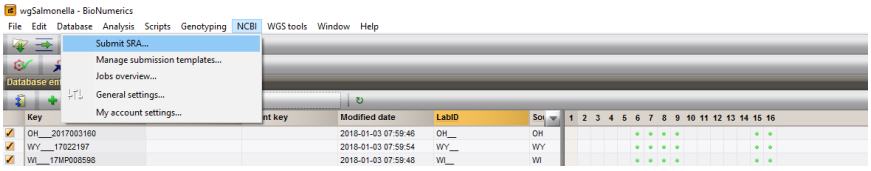


6.10. Upload to NCBI.

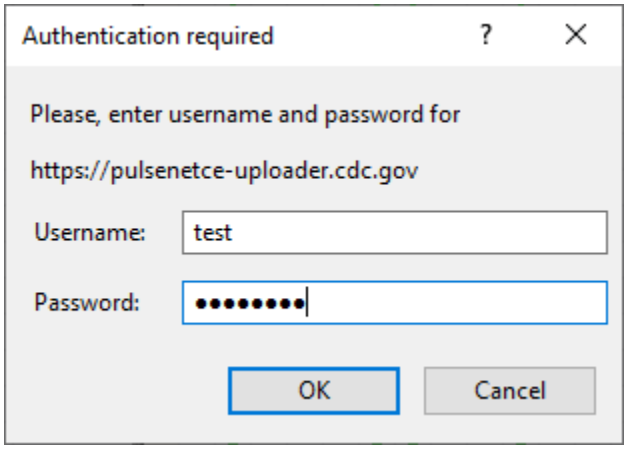
6.10.1. In the BioNumerics main window, select the entries to be uploaded. **NOTE1:** verify that the entries have the "WGS_id" field populated. Refer to step 6.9. on how to download WGS_ids from the national database. **NOTE2:** clinical and non-clinical isolates, different species, different library preparation kits and different instrument types require separate submission templates and cannot therefore be uploaded at the same time.

6.10.2. Locally linked data.

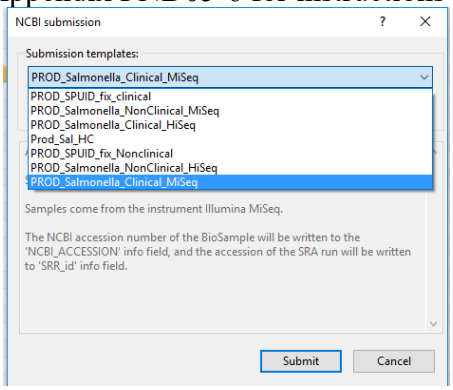
6.10.2.1. Authenticate to the CDC Firewall and from the "NCBI" drop-down menu, choose "Submit SRA".



6.10.2.2. "Authentication required" window appears. Enter your username and password for NCBI provided to you by CDC after you passed the analysis certification. Click "OK".

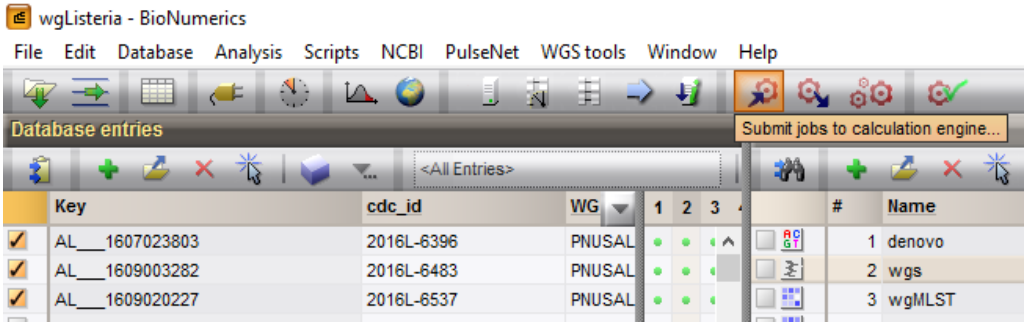


6.10.2.3. "NCBI submission" window appears. Choose the appropriate submission template and click "Submit". **NOTE1:** instructions for creating submission templates are outlined in Appendix PND05-5. **NOTE2**: If sequence data for the sample in question has been retracted from NCBI due to quality issues, please see Appendix PND05-6 for instructions to resubmit data.



6.10.3. BaseSpace linked data.

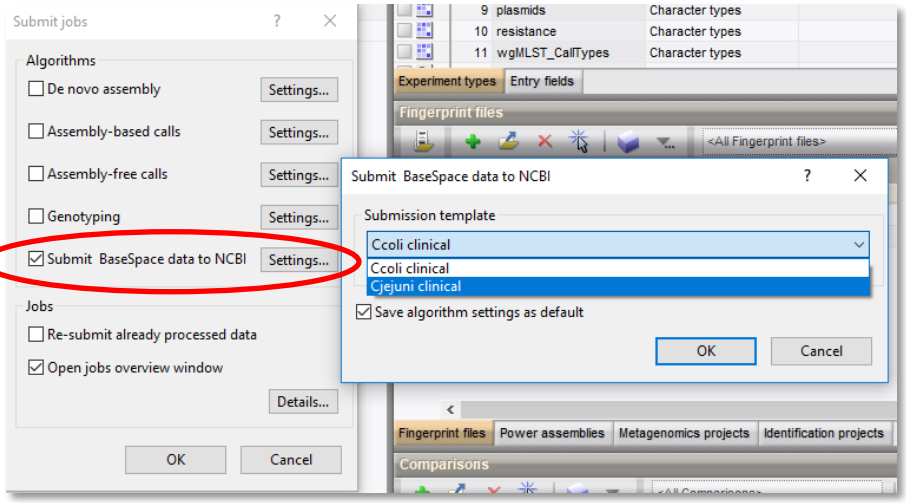
6.10.3.1. Click "Submit jobs to the calculation engine".



6.10.3.2. In the "Submit jobs" window, check the box for "Submit BaseSpace data to NCBI" and click on "Settings".

6.10.3.3. In the appearing "Submit BaseSpace data to NCBI" window, select the appropriate submission template and click "OK". **NOTE:** instructions for creating submission templates are outlined in Appendix PND05-5.

6.10.3.4. Click "OK" in the "Submit jobs" window.

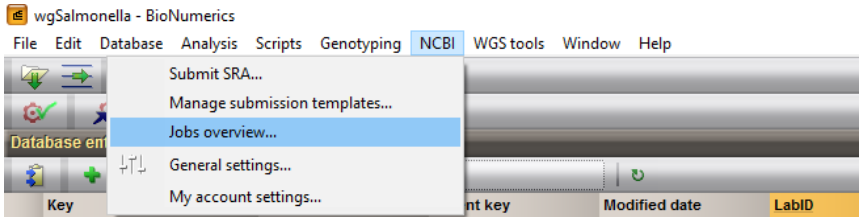


6.10.3.5. Monitor the job in the calculation engine "Jobs overview" window. When the status changes to "Finished" retrieve the job(s).

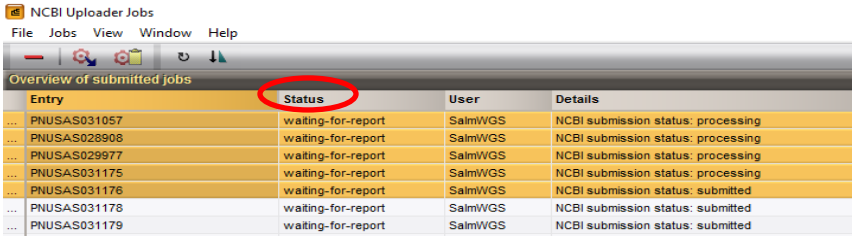
6.10.3.5.1. If you are not already authenticated to NCBI, you will be prompted to provide your credentials. Enter your username and password provided by CDC after you passed the analysis certification.

6.10.4. Monitor the status of NCBI submissions and retrieve the accession numbers.

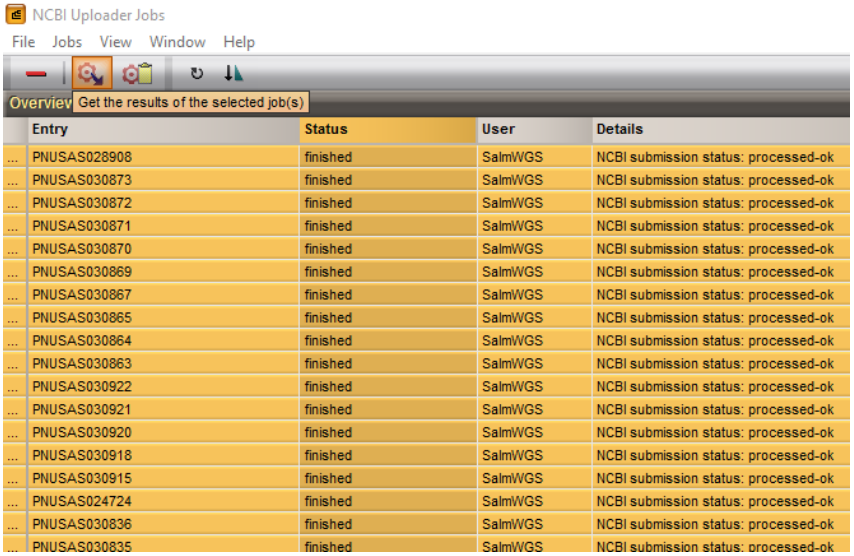
6.10.4.1. In the BioNumerics main window, choose "Jobs overview" from the "NCBI" drop-down menu. The "Status" field indicates the progress of the submission.



6.10.4.1.1. If the status is "Submitting" or "Waiting-for-report", you can close the "Jobs overview" window in BioNumerics and check back later.

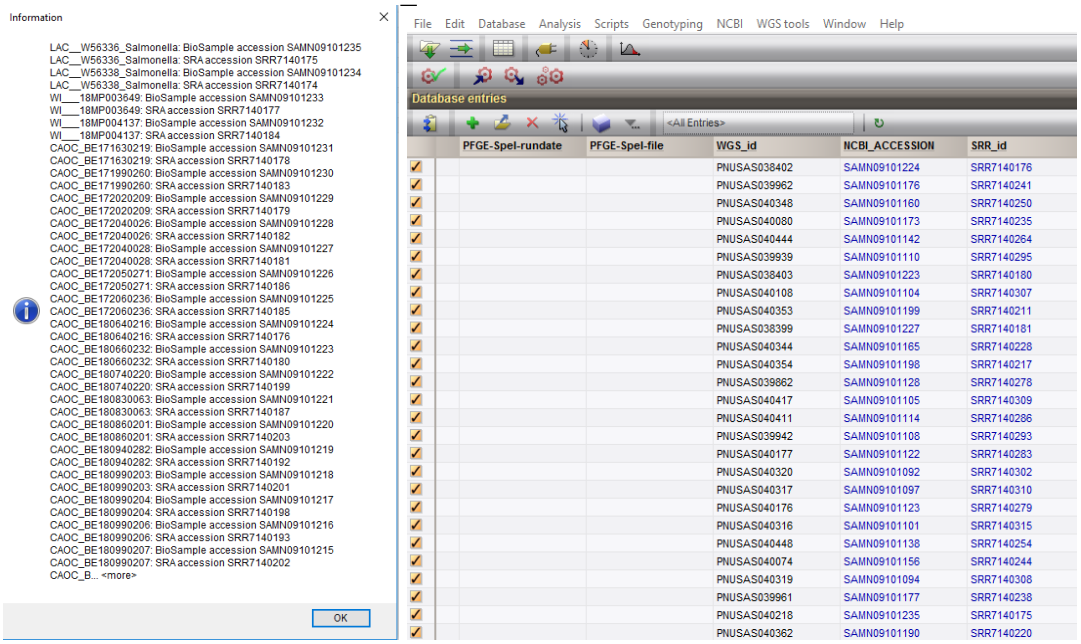


6.10.4.1.2. If the status is "Finished", highlight the finished jobs and click on "Get the results of the selected job(s)".



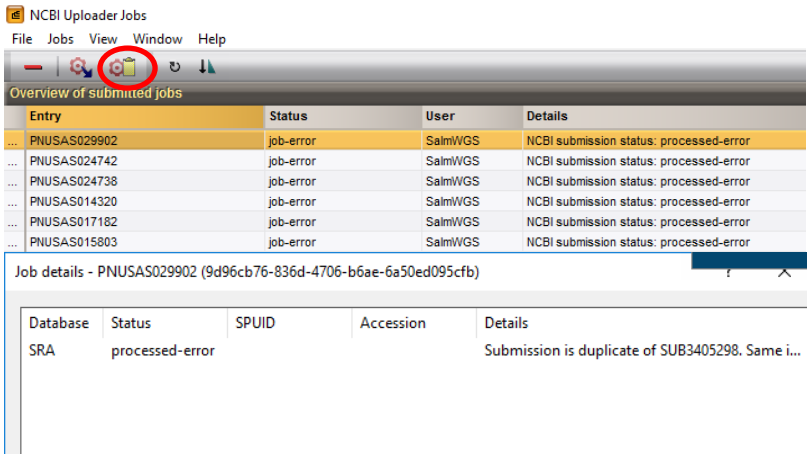
6.10.4.1.2.1. An information window with a summary of the SAMN and SRR numbers will appear. Click “OK”.

6.10.4.1.2.2. Updated entries are selected in the database with the imported accession numbers highlighted in blue in the “NCBI_ACCESSION” and “SRR_id” fields.



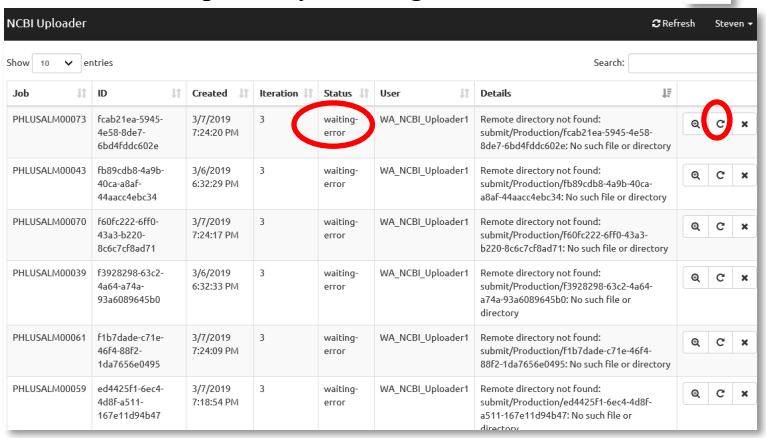
6.10.4.1.3. If the status is “Job-error”, first check NCBI (https://www.ncbi.nlm.nih.gov/) to determine if the submission went through despite erroring out in BioNumerics. If not present on NCBI, proceed to troubleshooting.

6.10.4.1.3.1. Highlight the errored job and click on “Get the logs of the selected job(s)” icon to open the “Job details” window.



6.10.4.1.3.2. If you need more details than what the "Jobs overview" window provides, authenticate to the PulseNetWGS firewall, login to NCBI viewer at https://pulsenetce-uploader.cdc.gov and enter your NCBI credentials.

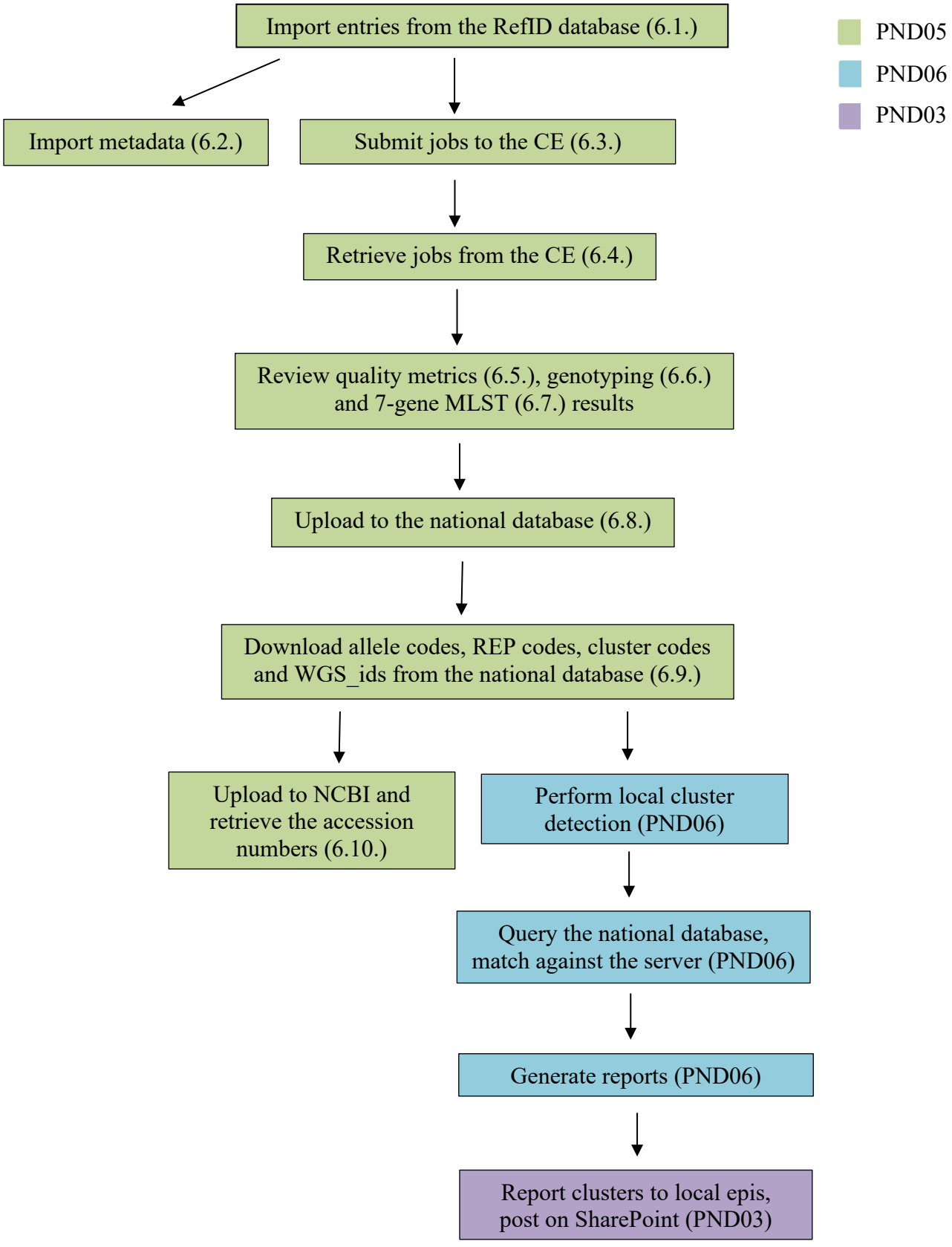
6.10.4.1.3.2.1. If the job failed due to a waiting-error, try refreshing the upload by clicking on the arrow icon.



6.10.4.1.3.2.2. Click on the magnifying glass to get a detailed report. If you are unable to determine the cause of the error, email the report to PulseNet@cdc.gov for troubleshooting. You may be instructed to send the report to NCBI SRA Helpdesk (genomes@ncbi.nlm.nih.gov) for further troubleshooting. **NOTE:** do not try to re-upload to NCBI until the cause of the error has been determined and fixed.


Job details: PHLUSALM00073
Report:
<Response status="processed-error">
<Message error_code="34" severity="warning" error_source="data">Submission processing may be delayed due to necessary curator review. Please check spelling of organism, current information could not be resolved automatically and will require a taxonomy consult.</Message>
<Object target_db="BioSample" object_id="" spuid="PHLUSALM00073" spuid_namespace="GenomeTrakr"/>
</Response>
<Response status="processed-error">
<Message error_code="20" severity="error-stop" error_source="data">These attribute values are not valid so please correct the data. Refer to https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/ for required format descriptions. If you do not have information for the required field(s), please provide the value as either 'not collected', '
Submission:
<?xml version="1.0" encoding="utf-8"?>
<Submission schema_version="2.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="https://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/submit/public-docs/common/submission.xsd?view=co">
<Description>
<Organization role="owner" type="consortium">
<Name>GenomeTrakr</Name>
<Contact email="micromolecularlab@doh.wa.gov">
<Name>
<First>Genome</First>
<Last>GenomeTrakr</Last>
Close

7. FLOW CHART:


Import entries from the RefID database (6.1.)
Import metadata (6.2.)
Submit jobs to the CE (6.3.)
Retrieve jobs from the CE (6.4.)
Review quality metrics (6.5.), genotyping (6.6.) and 7-gene MLST (6.7.) results
Upload to the national database (6.8.)
Download allele codes, REP codes, cluster codes and WGS_ids from the national database (6.9.)
Upload to NCBI and retrieve the accession numbers (6.10.)
Perform local cluster detection (PND06)
Query the national database, match against the server (PND06)
Generate reports (PND06)
Report clusters to local epis, post on SharePoint (PND03)
PND05
PND06
PND03

8. RELATED DOCUMENTS:

Document Number	Title
PNQ07	Illumina Sequence Data QC SOP
PNQ08	PulseNet WGS Certification SOP
PND03	PulseNet SharePoint Access and Use SOP
PND06	Cluster Detection SOP
PND20	RefID Database Workflow SOP

9. REFERENCES:

10. CONTACTS:

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

11. AMENDMENTS:

11.1. 06/15/2020 – New Document

11.2. 03/25/2022

– Added language about downloading REP codes

– Added related documents section #8

- Updated the instructions for creating an NCBI submission template for food and environmental sequences

- Added separate instructions for creating an NCBI submission template for animal sequences

– Added the Appendix PND05-6, instructions for resubmitting sequence data to NCBI

12. APPROVAL SIGNATURES:

Approved By: Jennifer K. Adams Digitally signed by Jennifer K. Adams Date: 2022.03.31 11:43:48 -04'00' Date: 3/31/2022

PulseNet QA/QC Personnel

Approved By: Kelley B Hise Digitally signed by Kelley B Hise Date: 2022.03.31 11:22:46 -04'00' Date: 3/31/2022

PulseNet Outbreak Detection and Surveillance Unit Chief

Approved By: Rebecca L. Lindsey -S Digitally signed by Rebecca L. Lindsey -S Date: 2022.03.31 10:00:27 -04'00' Date: 3/31/2022

PulseNet Next Generation Subtyping Methods Unit Chief

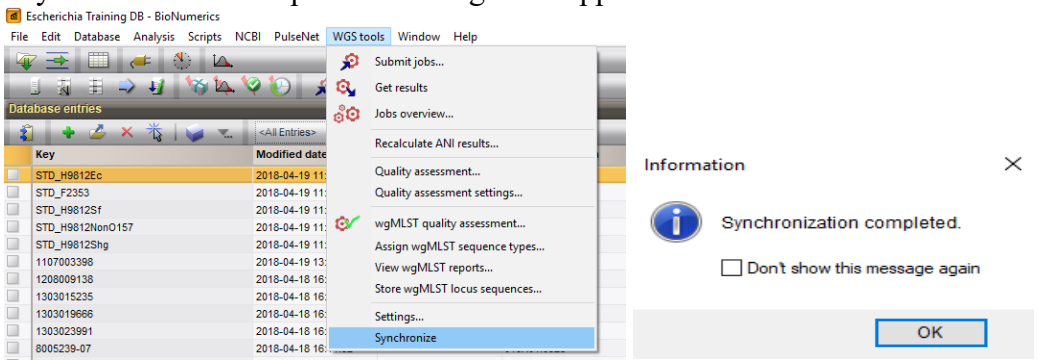
Approved By: Efrain M. Ribot -S Digitally signed by Efrain M. Ribot -S Date: 2022.03.31 09:57:22 -04'00' Date: 3/31/2022

PulseNet Reference Outbreak Surveillance Team Lead

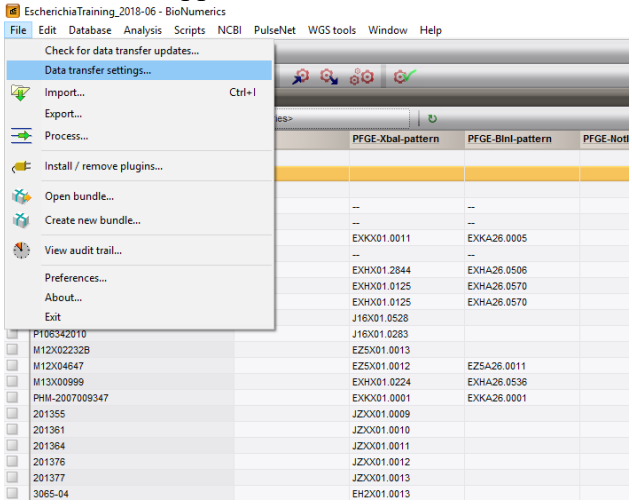
13. APPENDICES:

Appendix PND05-1. Initial Set-up and Settings in Organism-Specific Databases for Importing Data from the RefID Database

1. Verify updates from the calculation engine.
    a. From the "WGS tools" drop-down menu, choose "Synchronize". **NOTE:** this may be quick if updates have been processed before. If not, this can take a couple of hours to complete. Leave BioNumerics running while updates are occurring.
    b. "Synchronization completed" message will appear. Click "OK".

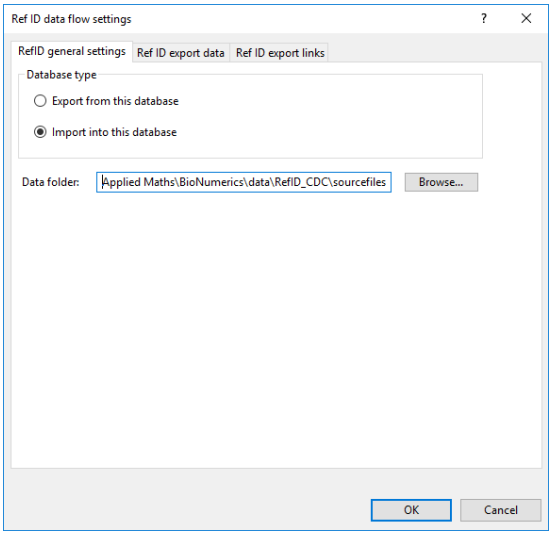


2. Before importing data from the RefID database for the first time, input/verify the correct data transfer settings.
    a. From the "File" drop-down menu, choose "Data transfer settings". "RefID data flow settings" window will appear.



    b. In the "RefID general settings" tab, make sure 'Import into this database" option is selected.
    c. Make sure the "Data folder" is linking to the correct location which is your RefID database source files folder. **This is the same path set up for export from the RefID database (PND20).**

d. The "RefID export data" and the "RefID export links" tabs should remain untouched. Click "OK".

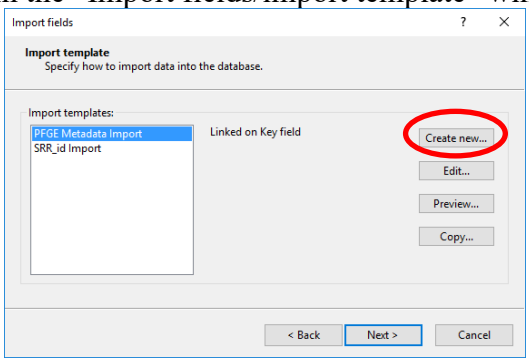


Appendix PND05-2. Instructions to Create a Metadata Import Template

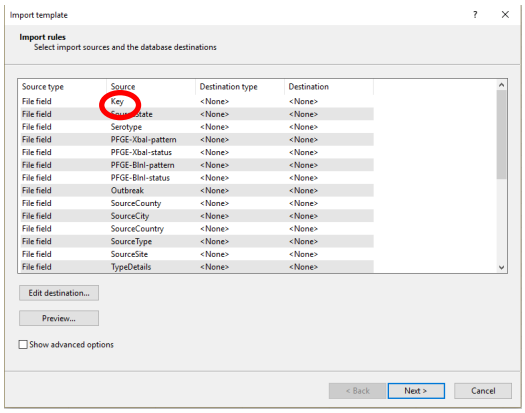
NOTE: At the minimum, the following metadata fields need to be populated in the organism-specific database:

a. SourceState: state of residence
b. SourceCounty, if known
c. SourceCity, if known
d. SourceType: MUST be Animal, Food, Human, Environmental, or Unknown
e. SourceSite: Blood, Stool, Urine, etc.
f. PatientSex
g. PatientAge: either Days, Months, or Years (only fill out one of these fields per sample)
h. IsolatDate: use the format YYYY-MM-DD
i. ReceivedDate: use the format YYYY-MM-DD

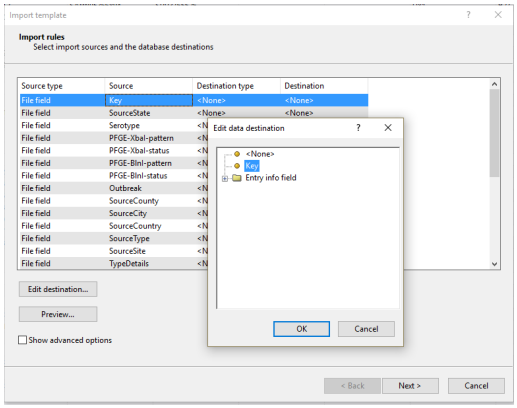
1. After following the steps 6.2.1.-6.2.4. to find the correct Excel file to import, click "Create new" in the "Import fields/import template" window.



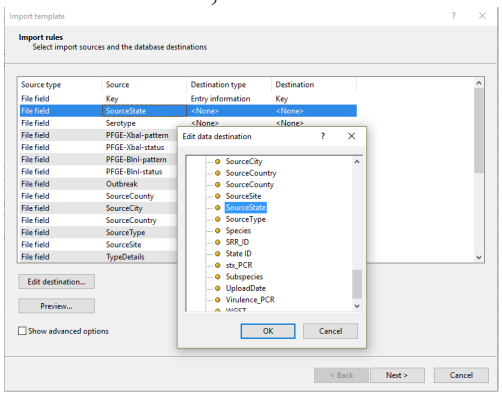
2. "Import template/Import rules" window will appear.
   a. "Source" column refers to the fields in the Excel file to be imported.
   b. "Destination" column refers to the information fields in the BioNumerics database.
   c. Double-click on "Key" in the "Source" column. **NOTE:** the key should be unique and have a match in the database.



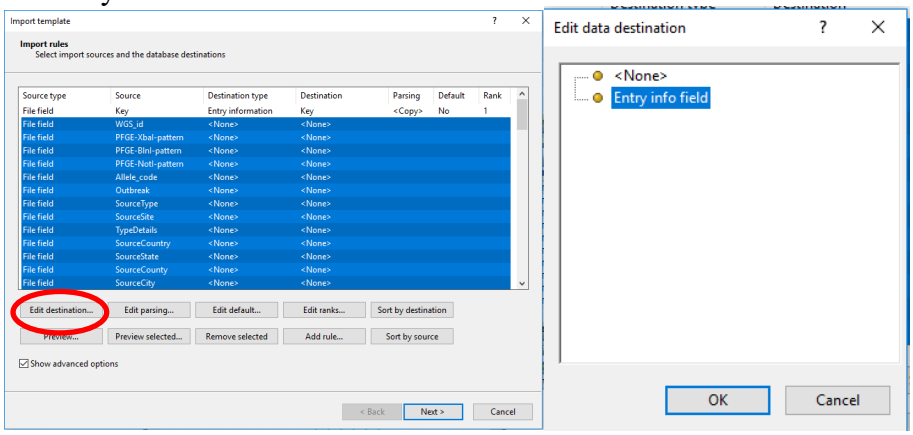
3. "Edit data destination" window will appear. Select "Key" and click "OK".



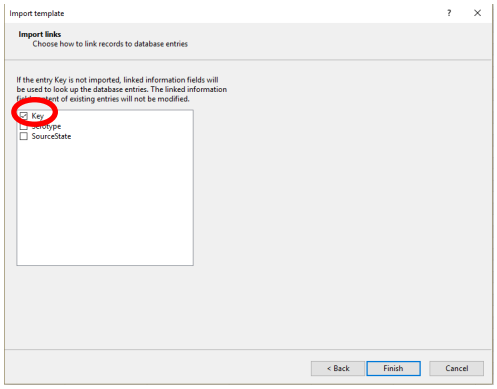
4. In the "Import template/Import rules" window, double-click on the next source field in your import file, in this case "SourceState".
5. In the "Edit data destination" window, expand the "Entry info field" option, select the corresponding destination field, in this case "SourceState" and click "OK".



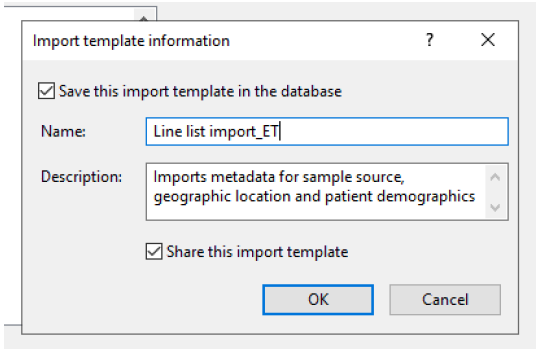
6. Repeat the process for the remaining source fields in your import file, and when you are finished, click "Next". **NOTE:** if all source fields in the import Excel file match exactly the destination fields in the BioNumerics database and there are no extraneous fields in the source file, you can select all source fields by holding down the "shift" key while selecting. Then click "Edit destination", and in the "Edit data destination" window highlight "Entry info field" and click "OK".



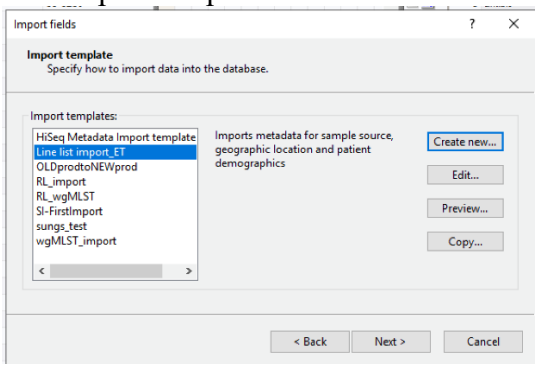
7. In the "Import template/Import links" window, select the source you want to link on, i.e., "Key" and click "Finish".



8. "Import template information" window will appear.
    a. Make sure the box for "Save this import template in the database" is checked.
    b. Give the template a unique name.
    c. In the description field, briefly describe how the template works.
    d. If you want to share the template with others using the database, check the box for "Share this import template".
    e. Click "OK".



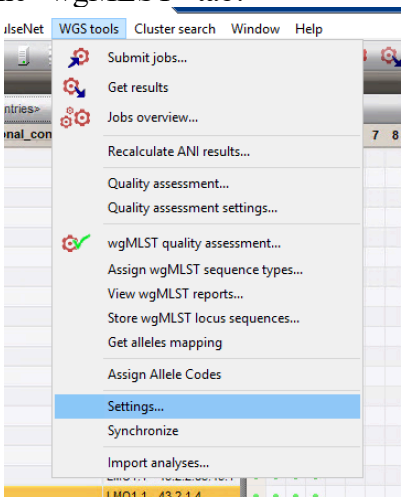
9. The import template you created now appears in the "Import fields/Import template" window under the "Import templates" list.



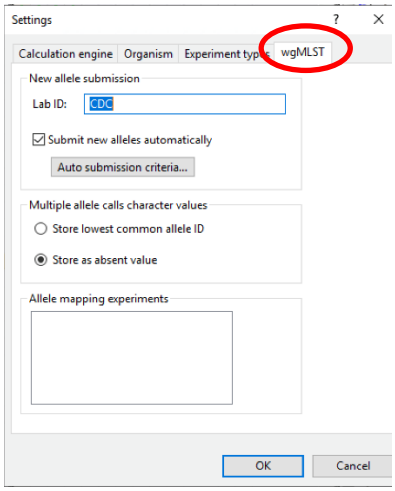
Appendix PND05-3. Initial Set-up and Settings for Allele Calling and Genotyping in Organism-Specific Databases

NOTE: Default settings are provided by CDC and should not be changed.

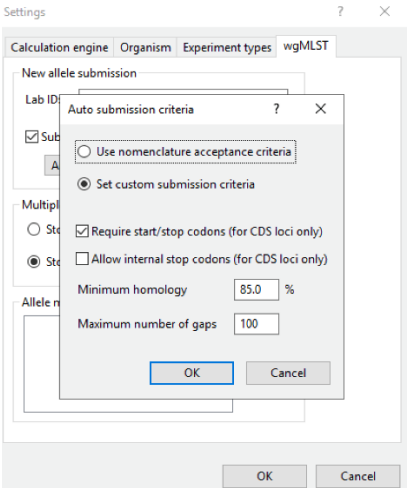
1. From the "WGS tools" drop-down menu, choose "Settings" and verify that the settings are correct in the "wgMLST" tab:



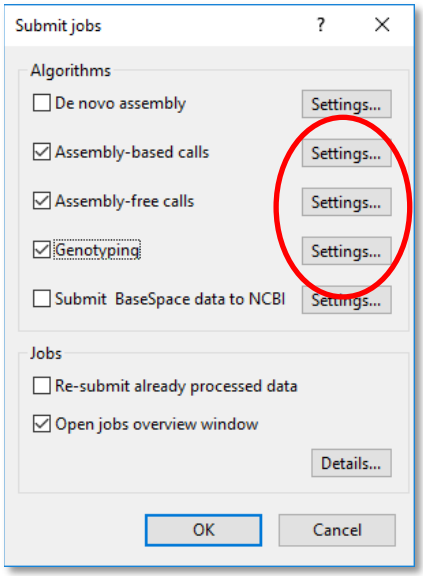
   a. For "Multiple allele calls character values", choose "Store as absent value". Make sure the "Submit new alleles automatically" box is checked and click on "Auto submission criteria". The "Auto submission criteria" window will appear.



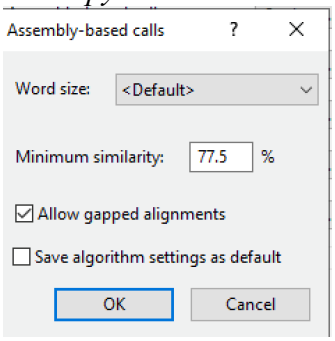
   b. Select "Set custom submission criteria".
   c. Check the box for "Require start/stop codons (for cds loci only)".
   d. Enter the minimum homology:
      i. *Listeria, Escherichia* and *Salmonella*: 85%
      ii. *Campylobacter*: 70%
   e. Enter maximum number of gaps for all organisms: 100
   f. Click "OK". This will take you back to the "wgMLST settings" tab. Click "OK" again.



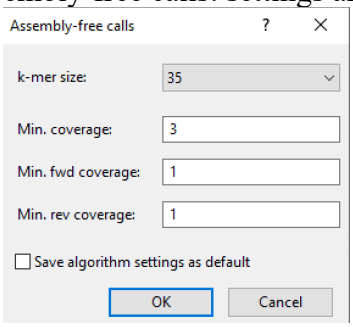
2. After you select entries and click on "Submit jobs to calculation engine", a "Submit jobs" window will appear. From there you can verify that the allele calling and genotyping settings are correct by clicking on "Settings" for each item.



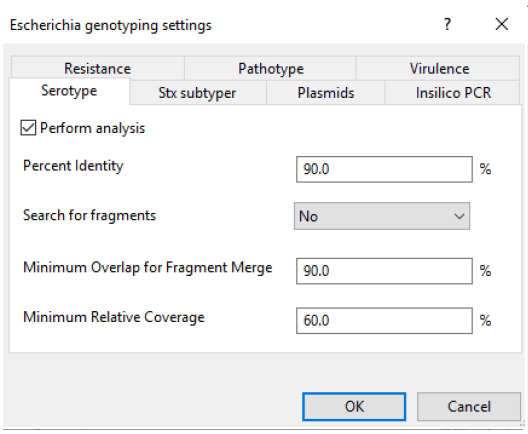
a. Assembly-based calls: minimum similarity varies by the organism database:
   i. *Escherichia:* 77.5
   ii. *Salmonella:* 75
   iii. *Campylobacter and Listeria:* 70

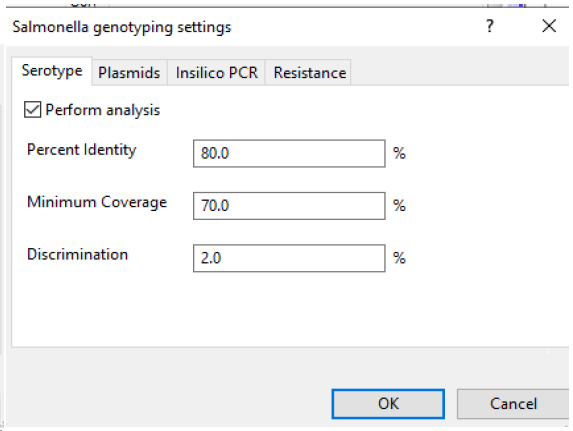


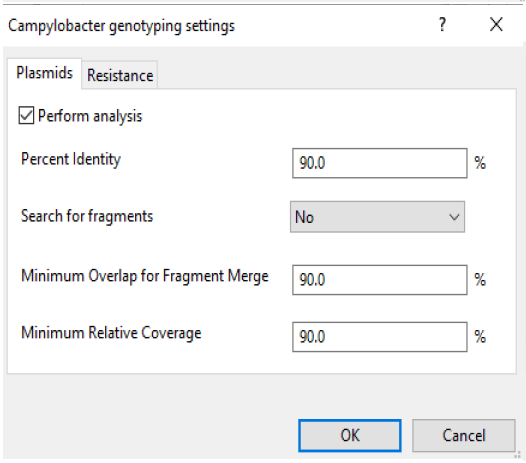
b. Assembly-free calls: settings are the same in all databases:

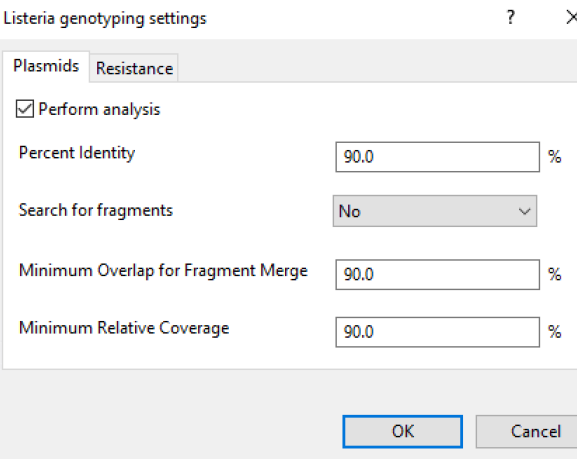


c. Genotyping: genotyper tools vary by organism database (screenshots below). Settings vary by organism for serotyping and in silico PCR and are the same for detection of plasmids and resistance for all organisms. Detection of virulence, stx subtype and pathotype apply to *Escherichia* only. See Table 2 below.





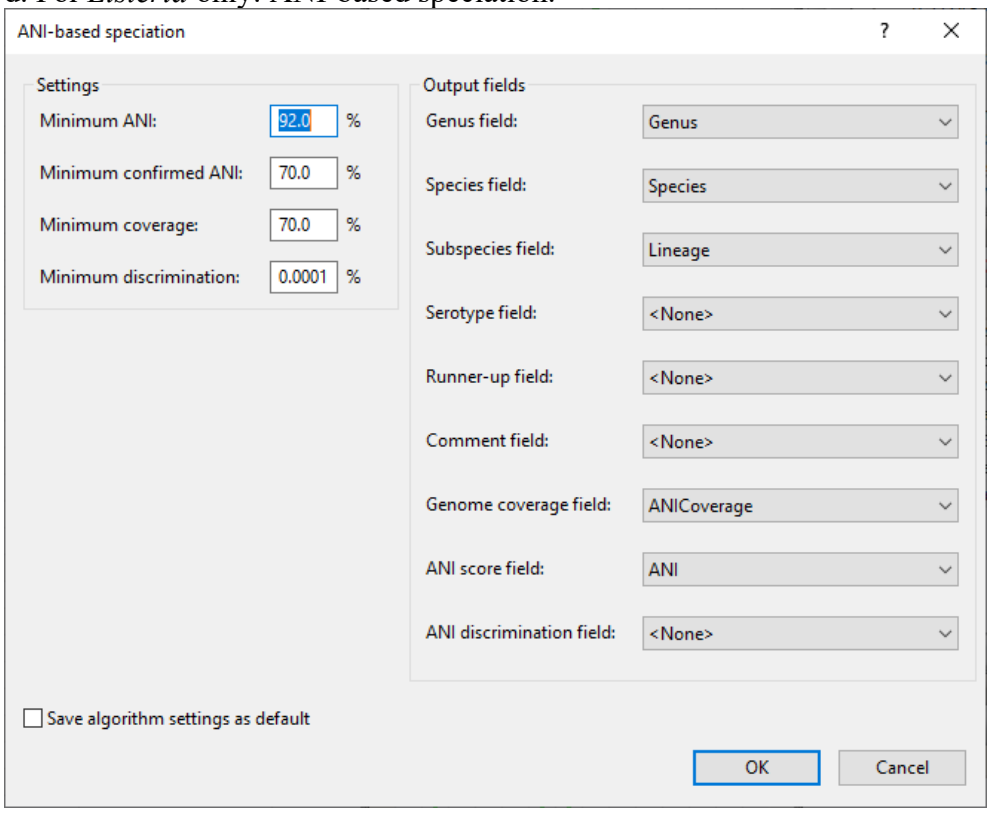




	Escherichia serotyping	*Salmonella* serotyping	*Escherichia* in silico PCR	*Salmonella* in silico PCR	Plasmids[2]	Resistance[2]	*Escherichia* pathotype	*Escherichia* virulence	*Escherichia* stx subtyping
Percent identity (%)	90.0	80.0	90.0	70.0	90.0	90.0	90.0	90.0	NA
Search for fragments	No	NA	NA	NA	No	Yes	Yes	Yes	NA
Min overlap for fragment merge (%)	90.0	NA	NA	NA	90.0	90.0	90.0	90.0	NA
Min relative coverage (%)	60.0	70.0	NA	NA	90.0	60.0	60.0	60.0	NA
Discrimination (%)	NA[1]	2.0	NA	NA	NA	NA	NA	NA	NA
Max non-UPAC nucleotides	NA	NA	2	3	NA	NA	NA	NA	NA
Max nucleotide mismatches	NA	NA	1	3	NA	NA	NA	NA	NA
Max amplicon length deviation (%)	NA	NA	10.0	10.0	NA	NA	NA	NA	NA

Table 2. Settings for the genotyper tools in the organism-specific databases. [1]NA, not applicable, [2]all organism databases

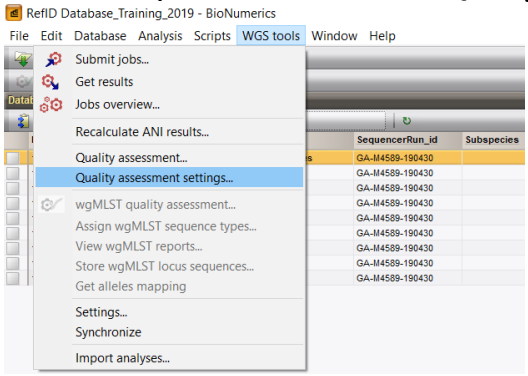
d. For *Listeria* only: ANI-based speciation.



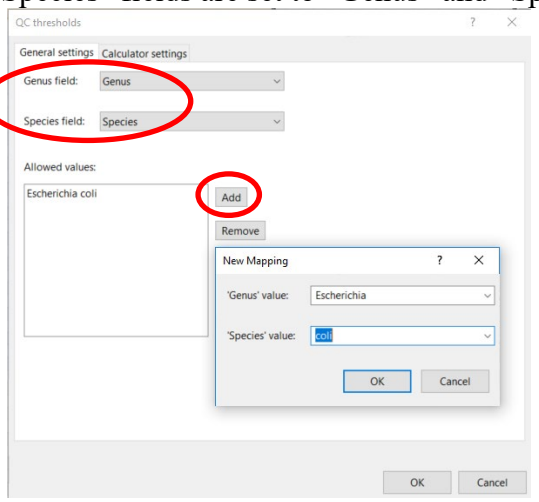
Appendix PND05-4. Configuration of the Quality Assessment Window in the Organism-Specific Databases

NOTE: The quality assessment settings only need to be configured once for each organism-specific database.

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



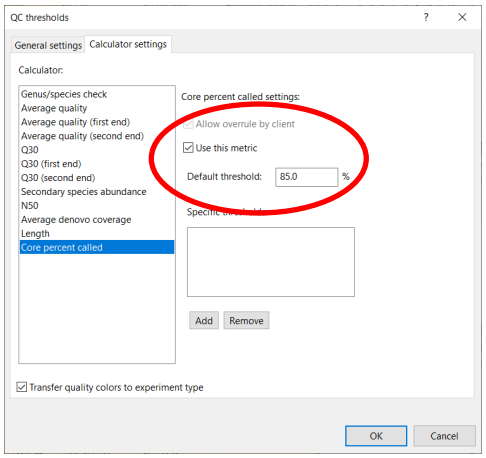
2. In the appearing "QC thresholds" window & "General settings" tab, make sure the "Genus" and "Species" fields are set to "Genus" and "Species".



   a. Click "Add", enter the appropriate Genus values and Species values listed below one at a time and click "OK":
      i. *Campylobacter* database:
         1) *Campylobacter jejuni*
         2) *Campylobacter coli*
         3) *Campylobacter lari*
         4) *Campylobacter upsaliensis*
         5) *Campylobacter fetus*
      ii. *Escherichia* database:
         1) *Escherichia coli* **NOTE:** includes *Shigella spp*. The ANI method in the RefID database identifies *Shigella* as *Escherichia*.
      iii. *Salmonella* database:
         1) *Salmonella bongori*
         2) *Salmonella enterica*

iv. *Vibrio* database:
    1) *Vibrio cholerae*
    2) *Vibrio parahaemolyticus*
    3) *Vibrio vulnificus*
    NOTE: database still under development
v. *Listeria* database:
    1) *Listeria monocytogenes*

3. Switch to the "Calculator settings" tab, select "Core percent called" from the "Calculations" list, make sure the box for "Use this metric" is checked, and enter the "Default threshold":
    a. *Campylobacter*, *Escherichia* and *Salmonella*: 85%
    b. *Listeria:* 95%

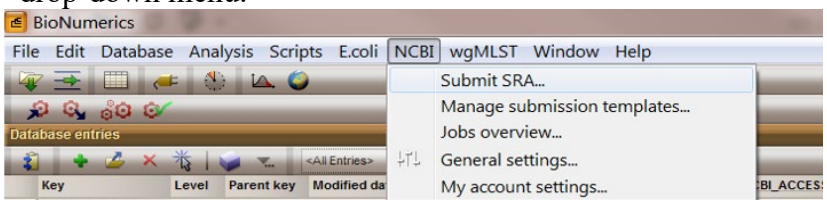


4. Manually uncheck "Use this metric" for all other quality metrics listed in the "Calculator" except core percent called. After you are done, click "OK". **NOTE1:** if you do not use RefID database for raw sequence quality assessment, then follow the instructions in PND20 Appendix 2 (PulseNet standard operating procedure for the reference identification database workflow) to set up the quality assessment window for Average quality, Average denovo coverage and Length. **NOTE2:** Genus/species check and Secondary species abundance can only be performed in the RefID database.

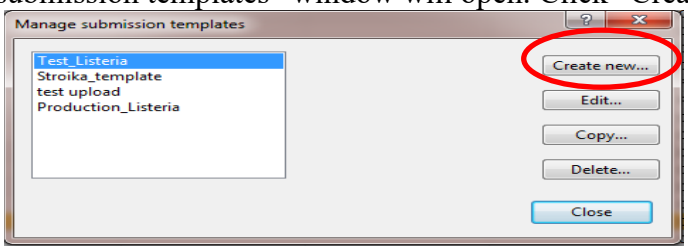
Appendix PND05-5. Instructions to Create NCBI Submission Templates

NOTE: Separate submission templates are required for clinical vs. non-clinical isolates, different species, different library preparation kits, and different instrument types.

1. In the main BioNumerics window, choose "Manage submission templates" from the "NCBI" drop-down menu.



2. The "Manage submission templates" window will open. Click "Create new".

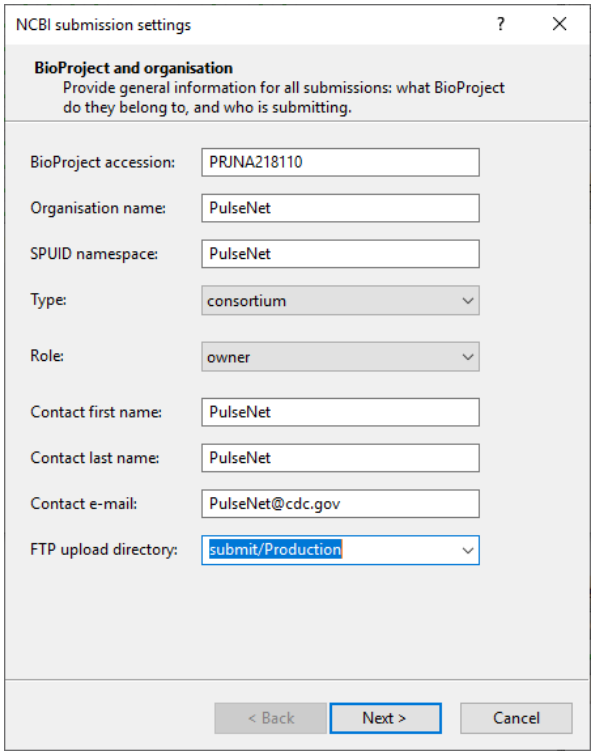


3. The "NCBI submission settings/Bioproject and organization" window will open.
   a. Enter the organism-specific bioproject accession number. Refer to Table 3 for PulseNet bioproject numbers. **NOTE:** please contact PulseNet@cdc.gov when you need to upload sequences for *Clostridium botulinum, Cronobacter* sp. or *Yersinia* sp.

Organism	**Bioproject accession number**
Campylobacter[1]	PRJNA239251
Escherichia and *Shigella*	PRJNA218110
Listeria	PRJNA212117
Salmonella	PRJNA230403
Vibrio[1]	PRJNA266293

Table 3.NCBI bioproject numbers for PulseNet organisms. [1]Campylobacter species other than coli, fetus, jejuni, lari, and upsaliensis and Vibrio species other than cholerae, parahaemolyticus and vulnificus are uploaded to NCBI from the RefID database.

   b. Enter the contact information for CDC (screenshot below) in order to anonymize your submission.
   c. Make sure the "FTP upload directory" is "submit/Production". Click "Next".



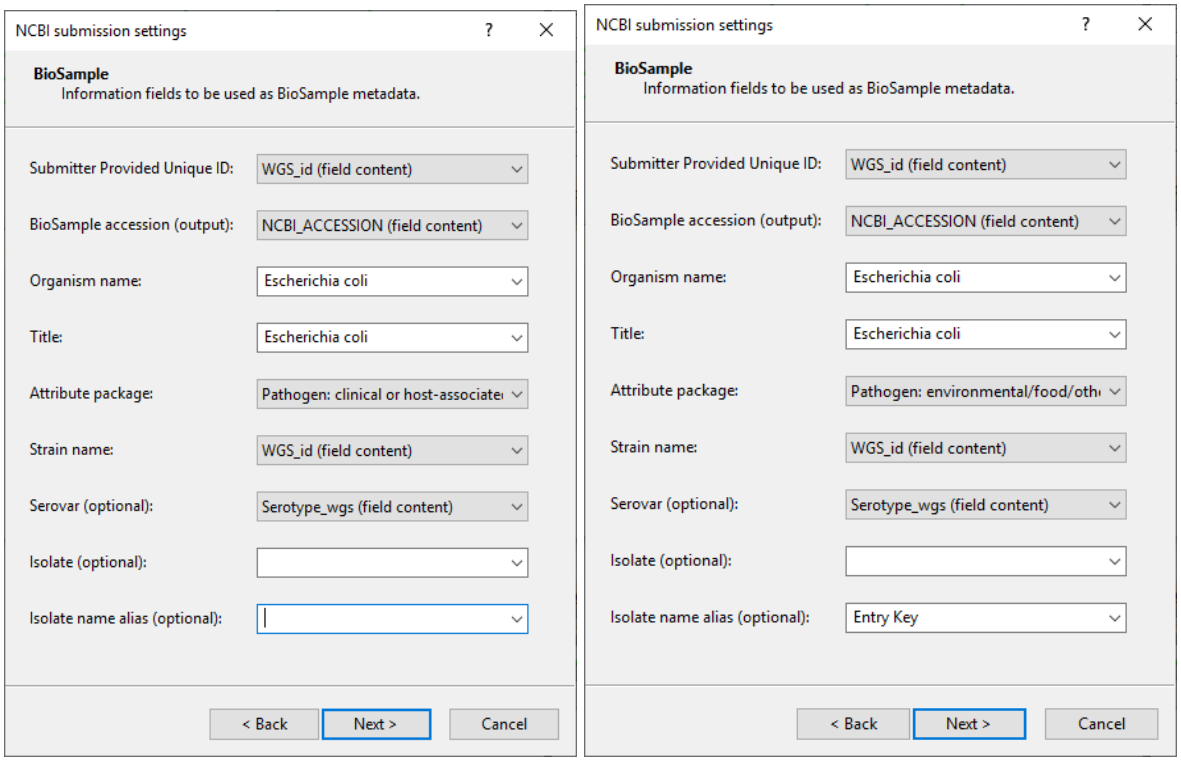
4. The "NCBI submission settings/Biosample" window will open.
    a. Submitter Provided Unique ID: from the drop-down menu, choose "WGS_id (field content)".
    b. Biosample accession (output): from the drop-down, choose "NCBI_ACCESSION (field content)".
    c. Organism name and title should match. Genus and species are preferred, therefore in databases with multiple species, such as the *Escherichia* database, species-specific templates are required for upload. If the species is unknown, it is acceptable to enter the genus name and "sp." in the organism and title fields, e.g., "*Shigella* sp.". Refer to Table 4 for the species-specific submission templates needed for PulseNet NCBI uploads. **NOTE:** when submitting a sequence with unknown species, the following error message will likely appear: "*<Message error_code="33" severity="warning" error_source="data">Submission processing may be delayed due to necessary curator review. Please check spelling of organism, current information could not be resolved automatically and will require a taxonomy consult.</Message>".* Anything submitted to NCBI with "sp." in the organism name field will trigger a curator review of the submission. **The submission will eventually go through, no additional action is required from the submitter.**

Genus	**Species**
Campylobacter	*coli*
	fetus
	jejuni
	lari
	upsaliensis

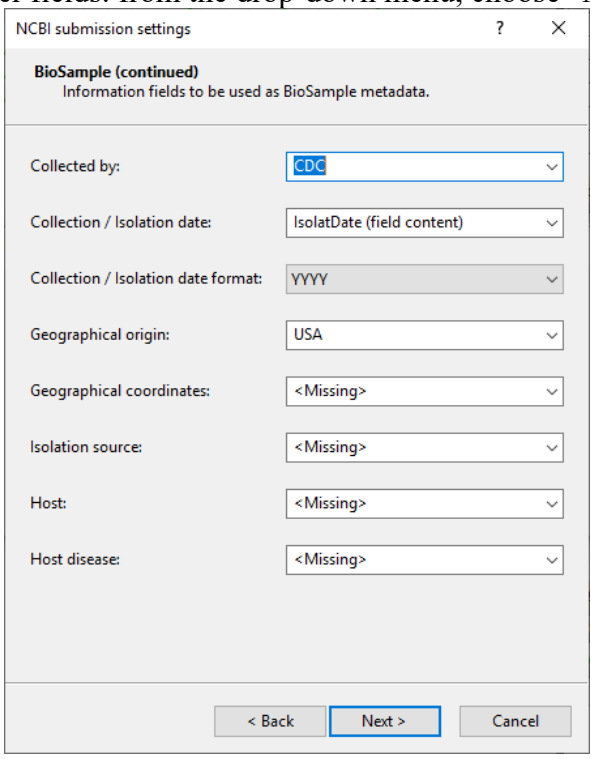
	sp.[1]
Escherichia *Shigella*	*coli*
	flexneri
	sonnei
	dysenteriae
	boydii
	sp.
Listeria	*monocytogenes*
Salmonella	*bongori*
	enterica
Vibrio	*cholerae*
	parahaemolyticus
	vulnificus
	sp.[1]

Table 4. Species included in the PulseNet organism-specific databases. [1]Uploaded from the RefID database.

d. Attribute package: from the drop-down menu, choose "Pathogen: clinical or host associated" for clinical (human) isolates or "Pathogen: environmental/food/other" for food, environmental and animal isolates.
e. Strain name should match the Submitter Provided ID: "WGS_id (field content)".
f. Serovar: from the drop-down menu, choose "Serotype_wgs (field content)".
g. For clinical samples, leave the remaining optional fields blank. For non-clinical samples, choose "Entry Key" from the drop-down menu for the "Isolate name alias" field.
h. Click "Next".

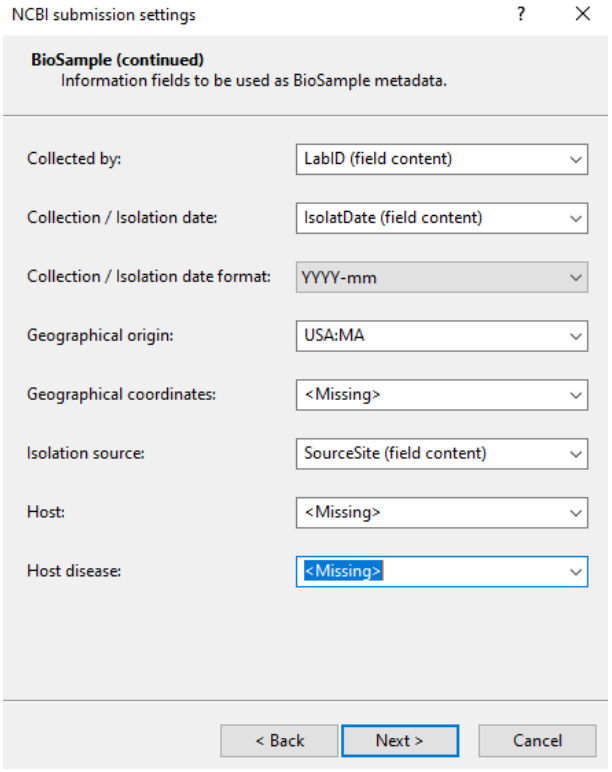


5. The "NCBI submission settings/Biosample (continued)" window will open.
    a. For clinical samples:
        i. Collected by: enter "CDC".
        ii. Collection/Isolation date: from the drop-down menu, choose "IsolatDate (field content)".
        iii. Collection/Isolation date format: from the drop-down menu, choose "YYYY".
        iv. Geographical origin: enter the country of origin, e.g., "USA" or alternatively choose "SourceCountry (field content)" from the drop-down menu if you routinely populate this field in your database.
        v. All other fields: from the drop-down menu, choose "Missing".

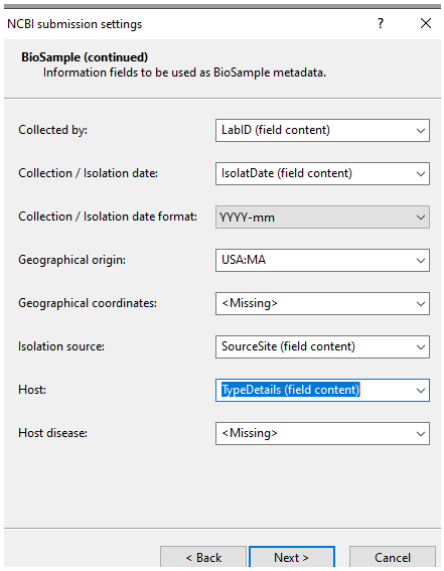


    b. For food and environmental samples:
        i. Collected by: from the drop-down menu, choose "LabID (field content)".
        ii. Collection/Isolation date: from the drop-down menu, choose "IsolatDate (field content)".
        iii. Collection/Isolation date format: from the drop-down menu, choose "YYYY-mm".
        iv. Geographical origin:
            1) For US isolates, enter the source country and state in the format "USA:State", e.g., "USA:MA".
            2) For non-US isolates, enter the country name, e.g., "Mexico" or alternatively choose "SourceCountry (field content) from the drop-down menu if you routinely populate this field in your database.
        v. Isolation source: from the drop-down menu, choose "SourceSite (field content)".

vi. All other fields: from the drop-down menu, choose “Missing”.

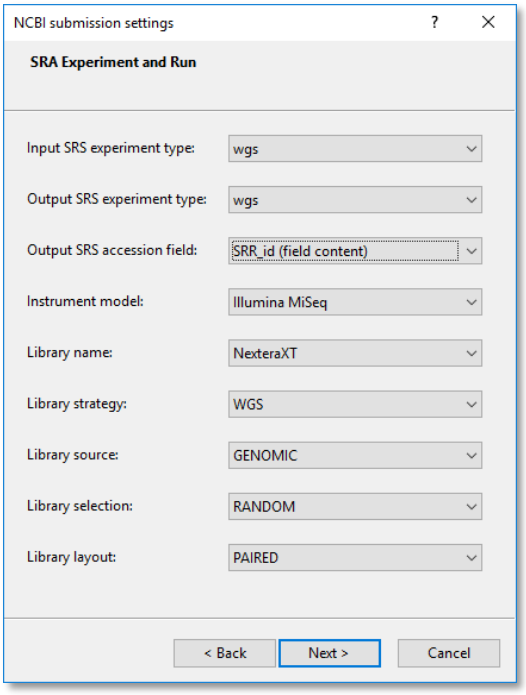


c. For animal samples:
   i. Collected by: from the drop-down menu, choose “LabID (field content)”.
   ii. Collection/Isolation date: from the drop-down menu, choose “IsolatDate (field content)”.
   iii. Collection/Isolation date format: from the drop-down menu, choose “YYYY-mm”.
   iv. Geographical origin:
      1) For US isolates, enter the source country and state in the format "USA:State, e.g, "USA:MA".
      2) For non-US isolates, enter the country name, e.g., "Mexico" or alternatively choose "SourceCountry (field content) from the drop-down menu if you routinely populate this field in your database.
   v. Isolation source: from the drop-down menu, choose “SourceSite (field content)”.
   vi. Host: from the drop-down menu, choose "TypeDetails (field content)" to specify the animal type, e.g., "Bovine".
   vii. All other fields: from the drop-down menu, choose “Missing”.



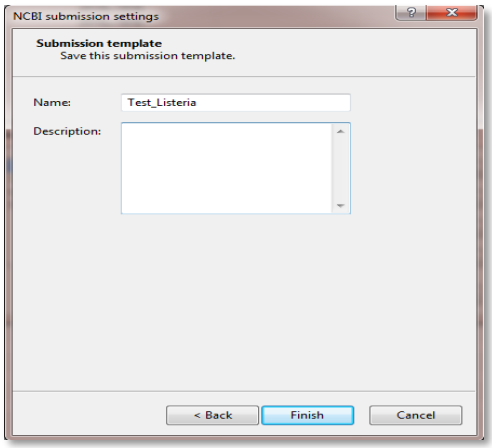
d. Click "Next".

6. The "NCBI submission settings/SRA experiment and run" window opens.
   a. Output SRS accession field: from the drop-down menu, choose "SRR_id (field content)".
   b. Instrument model: from the drop-down menu, choose the appropriate Illumina instrument.
   c. Library name: from the drop-down menu, choose the appropriate library preparation kit.
   d. The remaining fields: as shown in the screenshot below.

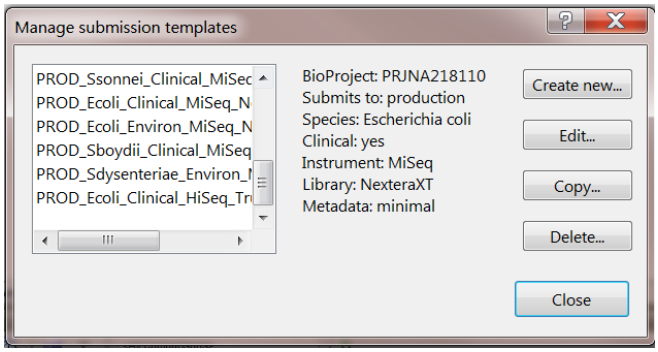


   e. Click "Next".

7. The "NCBI submission settings/Submission template" window opens.
    a. Enter a meaningful name.
    b. Enter a description.
    c. Click "Finish".



8. The saved template is now listed in the "manage submission templates" window.



Appendix PND05-6. Instructions for Resubmitting Sequence Data to NCBI

1. Before resubmitting, ask NCBI to retract the SRR_id by contacting the SRA helpdesk at genomes@ncbi.nlm.nih.gov.
2. Delete the SRR_id (SRR_id field) from the database.
3. Leave the biosample accession number (NCBI ACCESSION field) in the database. There is no need to create a new biosample for the resubmission.
4. In the database, copy the WGS_id from the WGS_id field into the "CDC_id" database field and add "_1", e.g., "PNUSAL002588_1".
5. Select the entry and resubmit using the correct resubmission (SPUID) template. Instructions to create a resubmission (SPUID) template:
    a. Choose "Manage submission templates" under the NCBI drop-down menu in the main database window.
    b. Select the template you used to make the original submission to NCBI and select "Copy".
    c. Add "SPUID" to the new template name to distinguish it from the regular template, e.g., "SPUID_Salmonella enterica_clinical".
    d. Select the copied template and click "Edit".
    e. Click "Next" in the "Bioproject and organization" tab. In the "Biosample" tab under the "Submitter Provided Unique ID" field, choose the "CDC_id" field instead of the "WGS_id" field.
    f. Click "Next" through the subsequent tabs until you reach the end and click "Finish".