1. **PURPOSE:** The purpose of this document is to describe a standardized procedure for analyzing sequences using PulseNet 2.0 and uploading sequence data to NCBI.

2. **SCOPE:** This procedure applies to whole genome sequence data generated by PulseNet participating laboratories for PulseNet surveillance. All genomic sequences generated by PulseNet participating laboratories are uploaded in real-time to the sequence read archive (SRA) located at NCBI and the analyzed data are published to the organism-specific surveillance databases included in the PulseNet 2.0 cloud-based analysis platform.

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 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., SALM1.1-95.247.305.381.443.35038 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 publishing WGS data to the PulseNet National Database and NCBI using PulseNet 2.0.
   - **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. Bioproject:** A collection of biological data related to a single initiative, originating from a single organization or from a consortium.

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

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

4.10. CDC: Centers for Disease Control and Prevention.

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

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

4.13. Cluster (outbreak) code: A 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.

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

4.15. Coverage: The average number of reads that include a given nucleotide in the reconstructed sequence. In PulseNet 2.0, average de novo coverage which is the calculated after quality-based trimming, is used as the critical (pass/fail) quality metric.

4.16. Critical Quality Metrics: Genus & Species (ANI percent aligned, ANI score), average de novo coverage, average quality (Q score), assembly length, MIDAS secondary coverage, percent core alleles called. Failing to meet the minimum thresholds/ acceptable range for any one of these metrics will cause the sequence to stop progression through the PulseNet 2.0 pipeline and will require re-sequencing of the isolate.

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

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

4.19. Grid view: Main PulseNet 2.0 view where the metadata and pipeline status icons are displayed.

4.20. 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.

4.21. MIDAS: Metagenomic Intra-species Diversity Analysis System. Integrated computational pipeline for quantifying bacterial species abundance and coverage from shotgun metagenomes based on blast alignment against a panel of universal single copy genes.

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. PCR: Polymerase Chain Reaction.

4.24. Pipeline: End to end workflow that includes de novo assembly, identification of the organism (genus and species), quality determinations, allele calling, and genotyping from sequence files.

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 2.0: Web-based, open-source platform used to analyze sequences, publish PulseNet organism sequences to the national databases, and upload sequences to NCBI.

4.28. PulseNet Central: PulseNet team at CDC comprising of the PulseNet Response and Outbreak Management Team (PulseNet@cdc.gov) and the PulseNet NGS Laboratory (PulseNetNGSlab@cdc.gov).

4.29. QC: Quality Control.

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

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

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

4.33. Secure Access Management System (SAMS): System used to control and manage access to PulseNet 2.0.

4.34. SeqSero2: A pipeline for *Salmonella* serotype determination from raw sequencing reads or genome assemblies.

4.35. SOP: Standard Operating Procedure.

4.36. 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.

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

4.38. 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.

4.39. WGS: Whole Genome Sequencing.

4.40. WGS_id: An anonymized identification number assigned to all PulseNet sequences as part of the national database publishing 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 published to the national database.

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

5. RESPONSIBILITIES:

5.1. Analysis certified PulseNet public health laboratory personnel analyze sequences and publish analyzed data and raw sequences to the PulseNet National Databases and NCBI using the PulseNet 2.0 workflow.

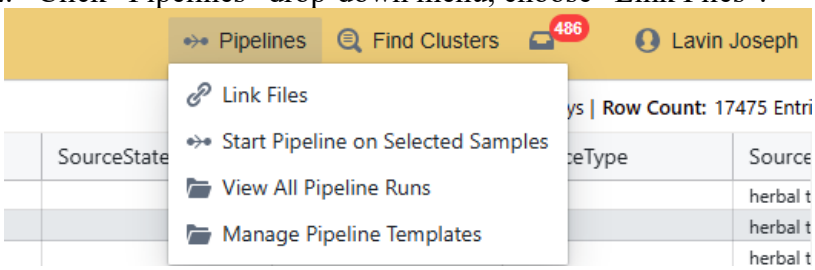
5.2. PulseNet Central personnel perform additional quality control checks of the data submitted by the participating laboratories and request re-sequencing if necessary.

6. PROCEDURE:

6.1. Import Entries and Data into PulseNet 2.0 and Start Pipeline.

6.1.1. Log into PulseNet 2.0 by going to the following URL: https://pulsenet-usa.cdc.gov/ and authenticate through SAMS.

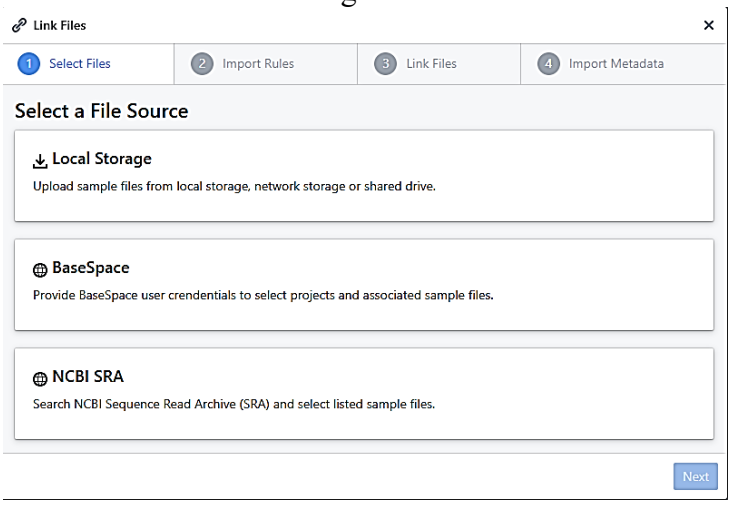
6.1.2. Click "Pipelines" drop-down menu, choose "Link Files".



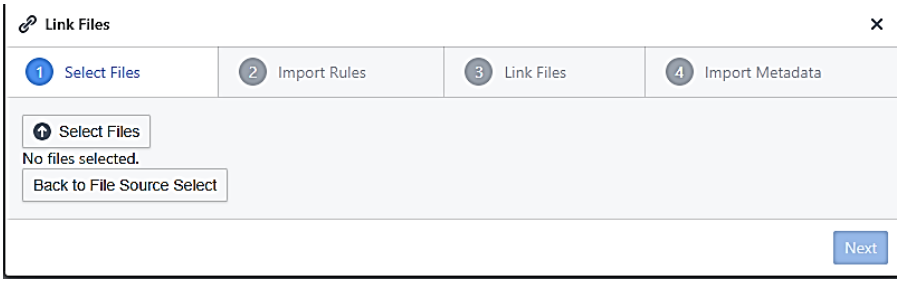
6.1.3. The "Link Files" window will appear.
NOTE: You will see "Local Storage", "BaseSpace", or "NCBI SRA". "Local Storage" is for sequence files stored on the desktop or a file share, "BaseSpace" is for sequences stored on your BaseSpace account and "NCBI SRA" is for sequences to be linked from NCBI.

6.1.4. **Locally Linked Sequence Files.**

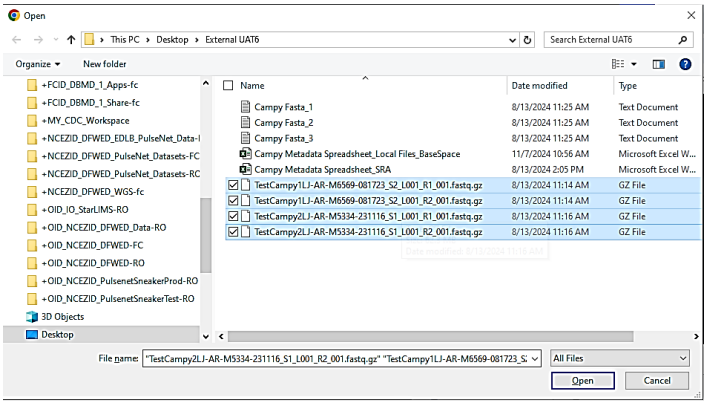
6.1.4.1. Click on "Local Storage"



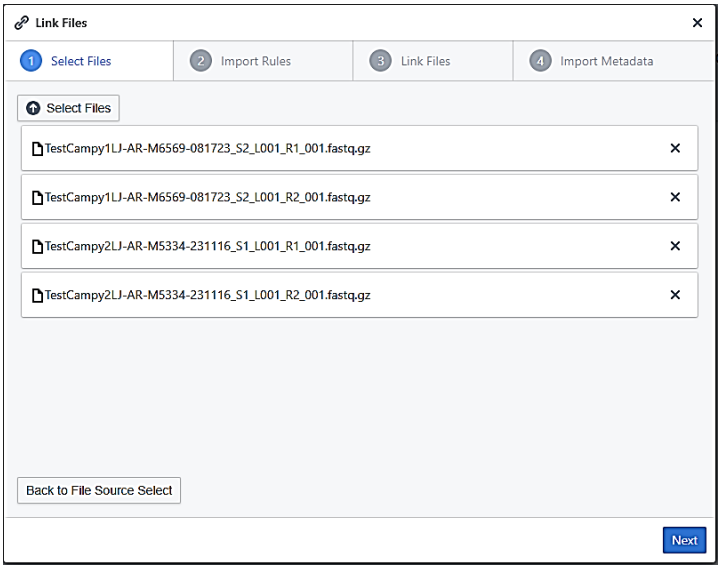
6.1.4.2. Click "Select Files"



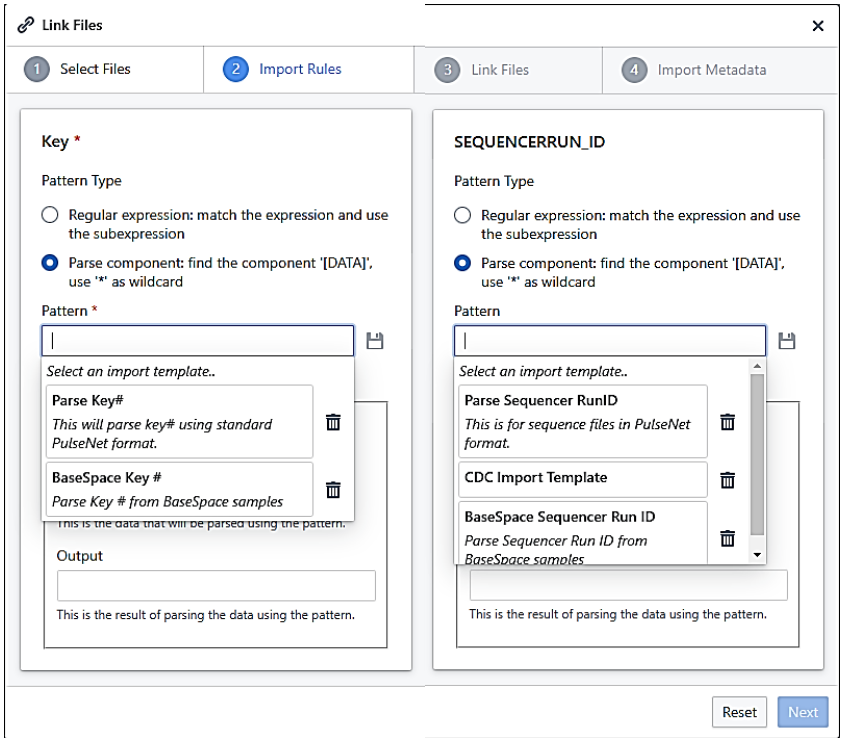
6.1.4.3. Navigate to the location of the sequence files; select the R1 and R2 files for each sequence and click "Open".



6.1.4.4. Click “Next”

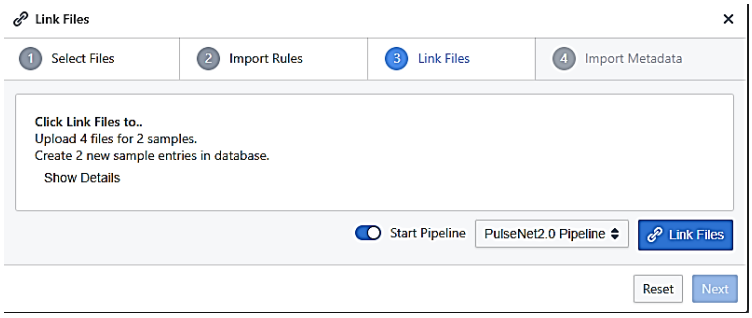


6.1.4.5. Select a saved import rule template for “Key” and “SEQUENCERRUN_ID” by clicking on the corresponding “Pattern” box and click “Next”.



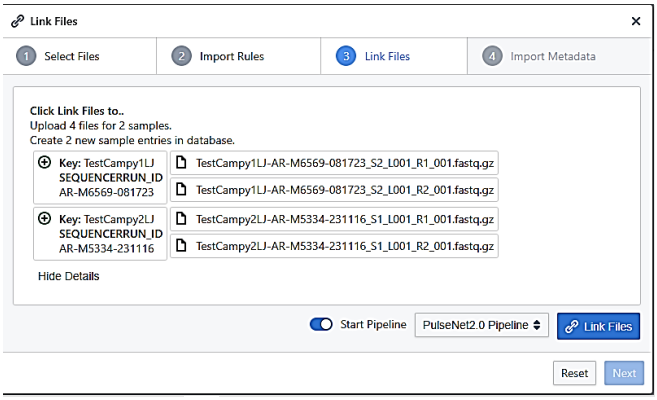
NOTE: See Appendix PND05-1 for instructions on creating and saving import rules/parsing templates for Key and SequencerRun_ID.

6.1.4.6. Click “Show Details” to show details about each sample being linked (optional).

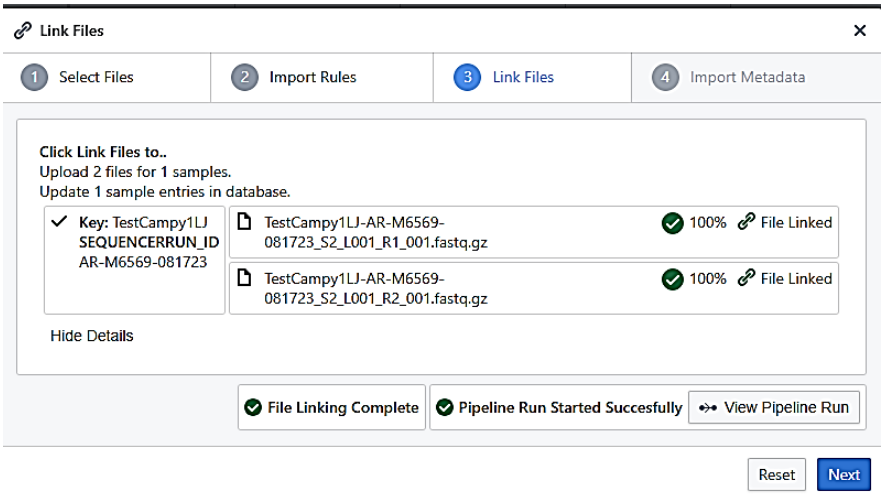


NOTE: The “Link Files” tab will show how many files are being uploaded and how many files are being created or updated.

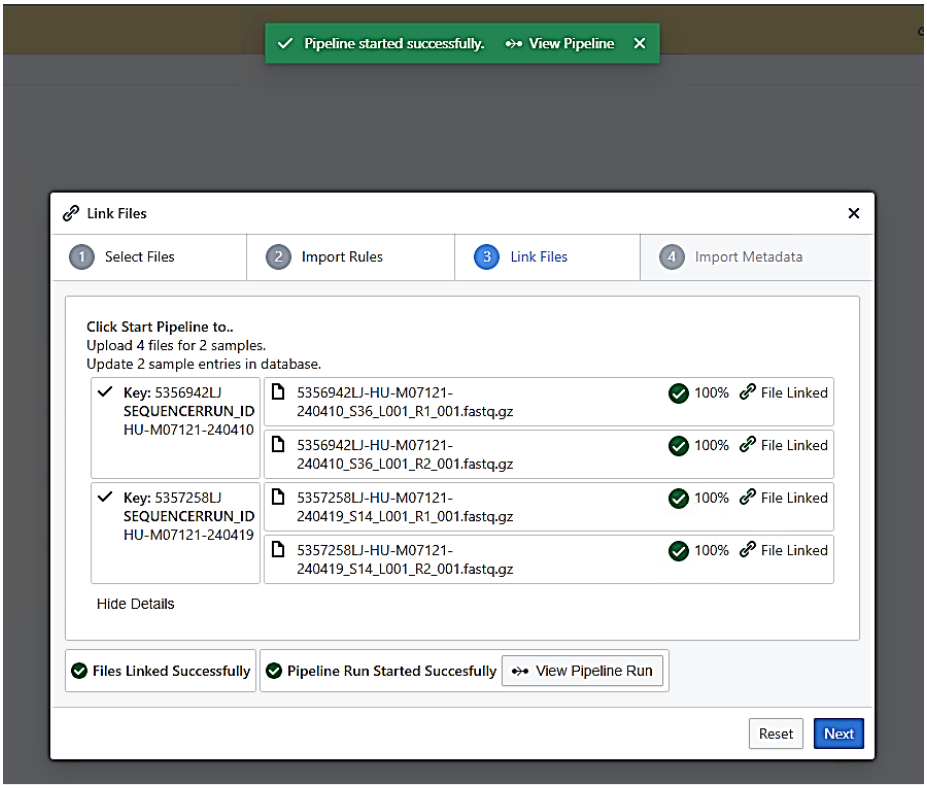
6.1.4.7. Make sure “Start Pipeline” is toggled on and “PulseNet2.0 Pipeline” is selected”. Click “Link Files”.



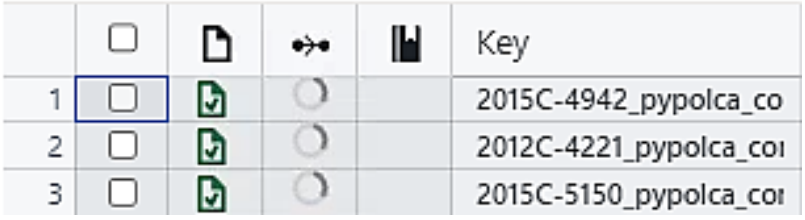
NOTE1: ⊕ indicates a new sample is being created and ✓ indicates a sample is being updated.
NOTE2: The upload progress window will show for each file. Users can close out of the "**Link Files"** window while the files are linking without interrupting the upload process; however, DO NOT close the PulseNet 2.0 web application until the linking completes. Click "Pipelines" in the top navigation panel and select "Link Files" to reopen the "Link Files" window.



6.1.4.8. After the files have been linked successfully, click "Next".
NOTE: In order to link more files in PulseNet 2.0, click "Reset" in the "Link Files" window and go to step 6.1.4.1.



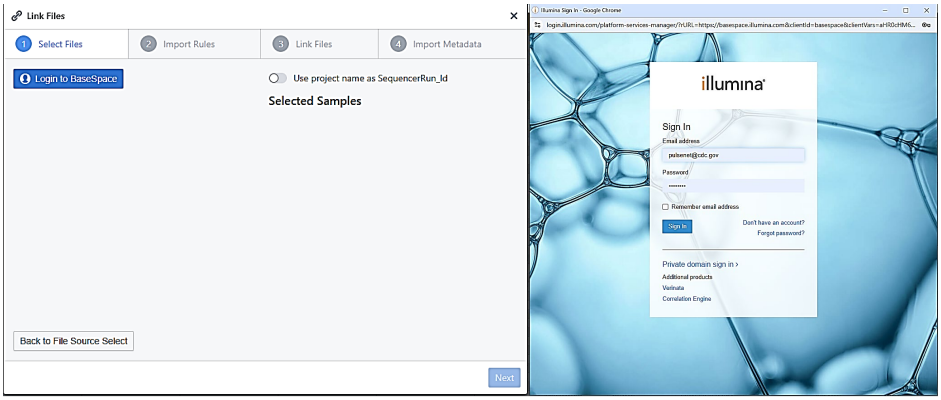
6.1.4.9. Click “Select Import File”. See Appendix PND05-2 for instructions on importing metadata.
NOTE: The imported samples will be selected in the grid view and the pipeline status icon will indicate the samples are running in the pipeline. The green indicator will be present showing that the sequence is linked to a local file source.



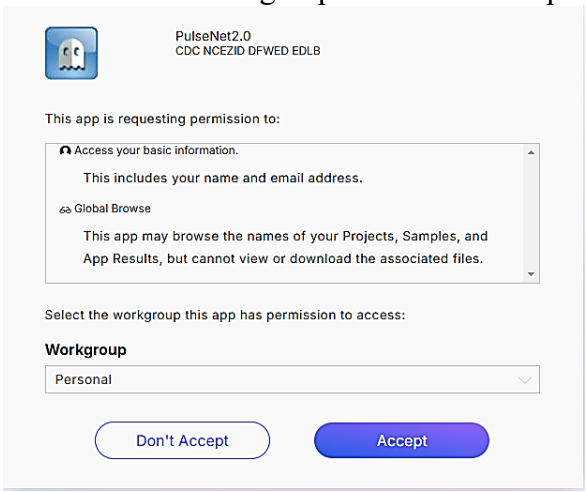
6.1.5. **BaseSpace Linked Sequence Files**

6.1.5.1. Under “Select a File Source” in the “Link Files” window, click on “BaseSpace”

6.1.5.2. Click “Login to BaseSpace” and enter your login credentials.

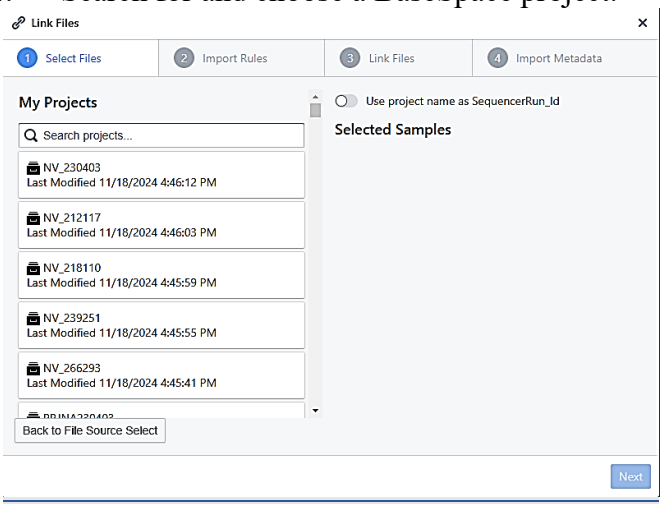


6.1.5.3. Select the workgroup and click “Accept”.

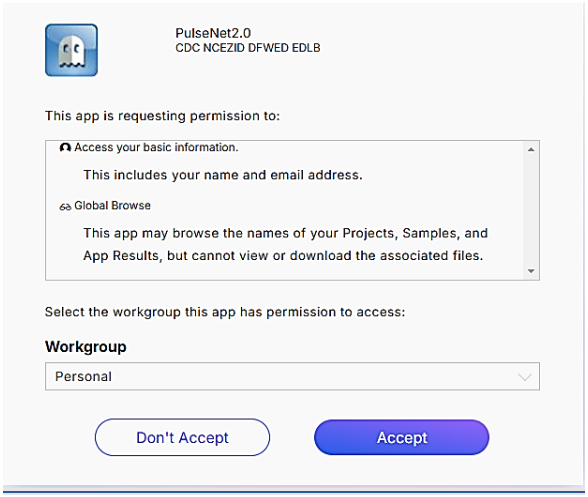


NOTE: BaseSpace projects should be saved in Personal Workgroups since PulseNet 2.0 cannot locate files in other workgroups.

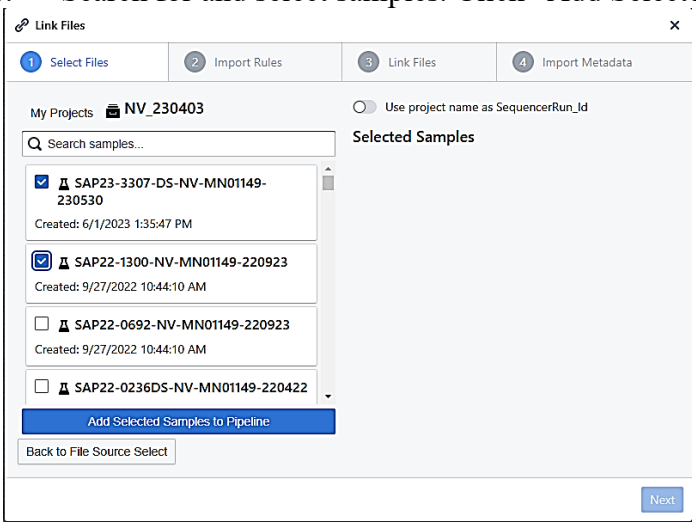
6.1.5.4. Search for and choose a BaseSpace project.



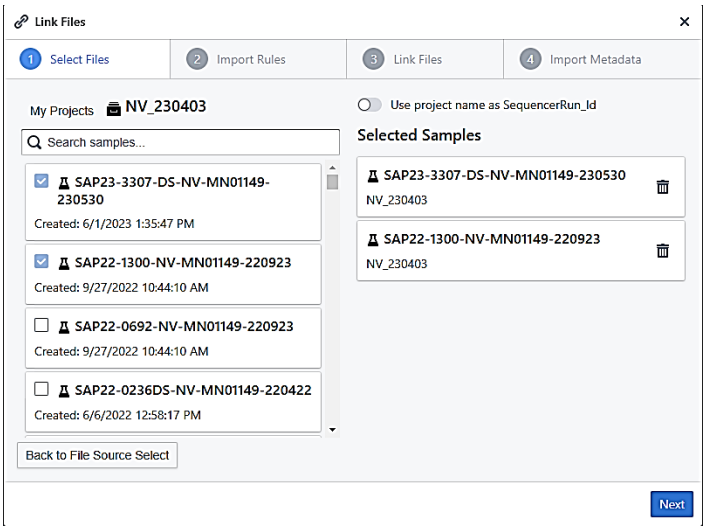
6.1.5.5. Select the Project folder and click “Accept”.



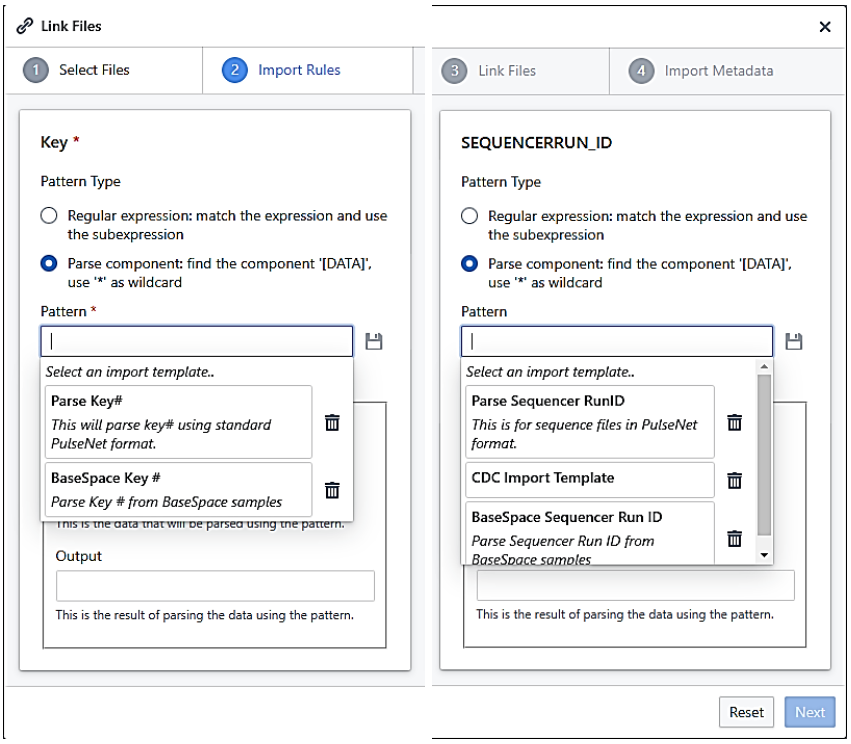
6.1.5.6. Search for and select samples. Click “Add Selected Samples to Pipeline”.



6.1.5.7. Click “Next”.

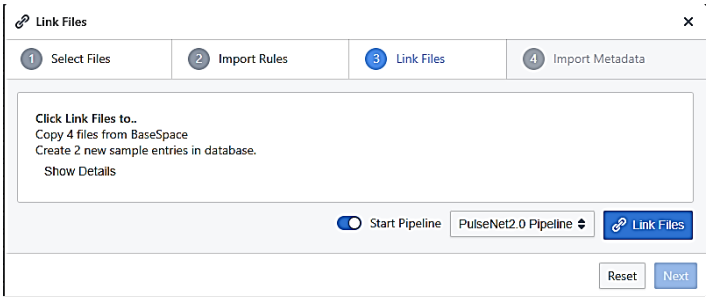


6.1.5.8. Select a saved import rule template for "Key" and "SEQUENCERRUN_ID" by clicking on the corresponding "Pattern" box and click "Next".



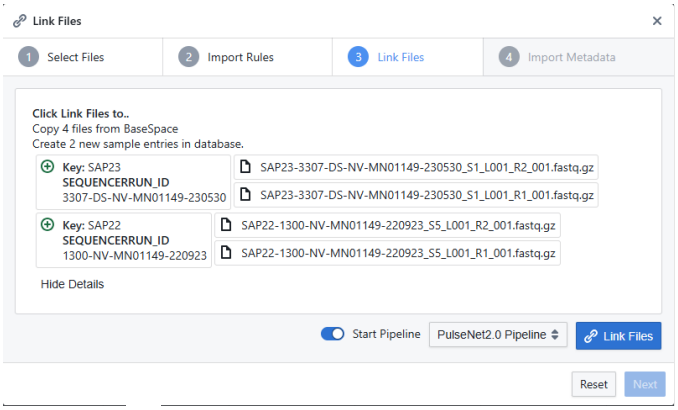
NOTE: See Appendix PND05-1 for instructions on creating and saving import rules/parsing templates for Key and SequencerRun_ID.

6.1.5.9. Click "Show Details" to show details about each sample being linked.



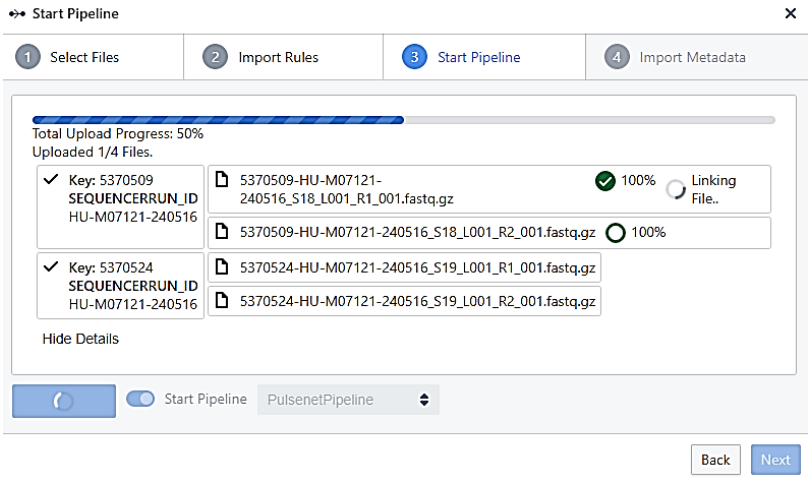
NOTE: The "**Link Files**" tab will show how many files are being uploaded and how many files are being created or updated.

6.1.5.10. Make sure "Start Pipeline" is toggled on and "PulseNet2.0 Pipeline" is selected". Click "Link Files".

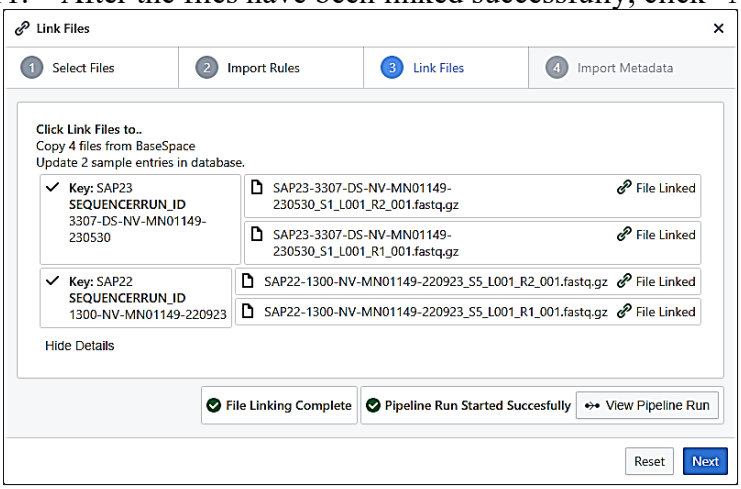


NOTE1: ⊕ indicates a new sample is being created and ✓ indicates a sample is being updated.

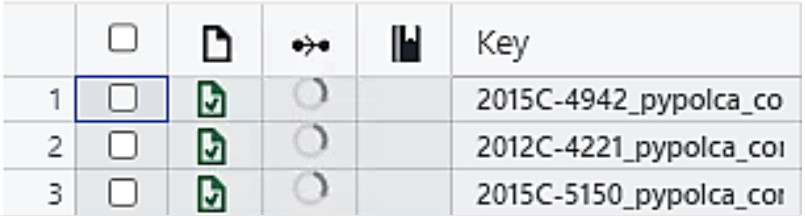
NOTE2: The upload progress window will show for each file. Users can close out of the "**Start Pipeline"** window while the files are linking without interrupting the upload process; however, DO NOT close the PulseNet 2.0 web application until the linking completes. Click "Pipelines" in the top navigation panel and select "Link Files" to reopen the "Start Pipeline" window.



6.1.5.11. After the files have been linked successfully, click "Next".

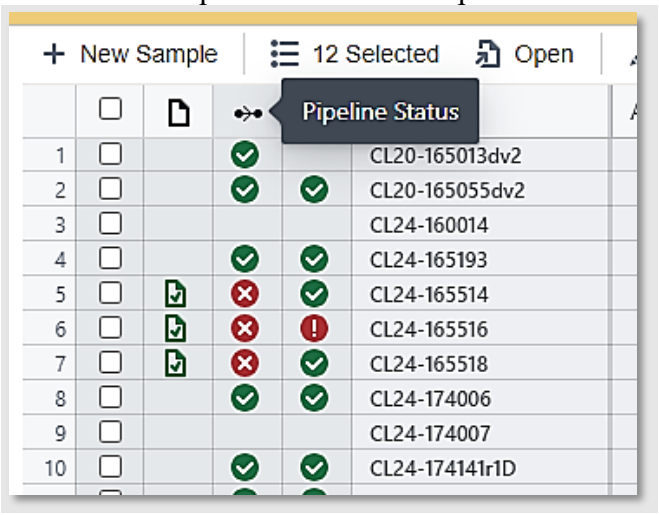


6.1.5.12. Click "Select Import File". See Appendix PND05-2 for instructions on importing metadata.
NOTE: The imported samples will be selected in the grid view and the pipeline status icon will indicate the samples are running in the pipeline. The green indicator will be present showing that the sequence is linked to a local file source.



6.2. View Status of Pipelines

6.2.1. After the pipeline has completed, you can view the status of the pipeline for each submitted sequence under the "Pipeline Status" icon in the grid view.



NOTE: Sequences that successfully pass the pipeline will be filtered to the corresponding organism-specific database views based on genus and species identification by ANI if PulseNet surveillance organisms. If ANI determined that the genus and species is not that of a PulseNet surveillance organism, the sequence will remain in the Reference_ID view. If ANI determines that the organism is not one of those listed below, the ANI step of the pipeline will fail. See Table 1 below.

Database	Organism	Database	Organism
Campylobacter	*Campylobacter coli*	Reference ID	*Eschcericia fergusonii*
	Campylobacter fetus		*Eschcericia marmotae*
	Campylobacter hyointestinalis		*Escherichia ruysiae*
	Campylobacter jejuni		*Grimontia hollisae*
	Campylobacter upsaliensis		*Listeria innocua*
Clostridium	*Clostridium botulinum*		*Listeria ivanoii*
Cronobacter	*Cronobacter condimenti*		*Listeria marthii*
	Cronobacter dubliensis		*Listeria seeligeri*
	Cronobacter malonaticus		*Listeria welshimeri*
	Cronobacter muytjensii		*Photobacterium damselae*
	Cronobacter sakazakii		*Vibrio alginolyticus*
	Cronobacter turicensis		*Vibrio cidicii*
	Cronobacter universalis		*Vibrio cincinnatiensis*
Escherichia	*Escherichia albertii*		*Vibrio fluvialis*
	Escherichia coli		*Vibrio furnisii*
Listeria	*Listeria monocytogenes*		*Vibrio harveyi*
Salmonella	*Salmonella bongori*		*Vibrio metoecus*
	Salmonella enterica		*Vibrio metschnikovii*
Vibrio	*Vibrio cholerae*		*Vibrio mimicus*
	Vibrio parahaemolyticus		*Vibrio navarrensis*
	Vibrio vulnificus		*Vibrio paracholerae*
Yersinia	*Yersinia enterocolitica*		*Vibrio tarriae*
			Yersinia aldovae
			Yersinia bercovieri
			Yersinia frederiksenii
			Yersinia intermedia
			Yersinia kristensenii
			Yersinia massiliensis
			Yersinia mollaretii
			Yersinia pseudotuberculosis
			Yersinia rohdei
			Yersinia species

Table 1. List of organisms identified in Reference ID and Organism Specific Databases

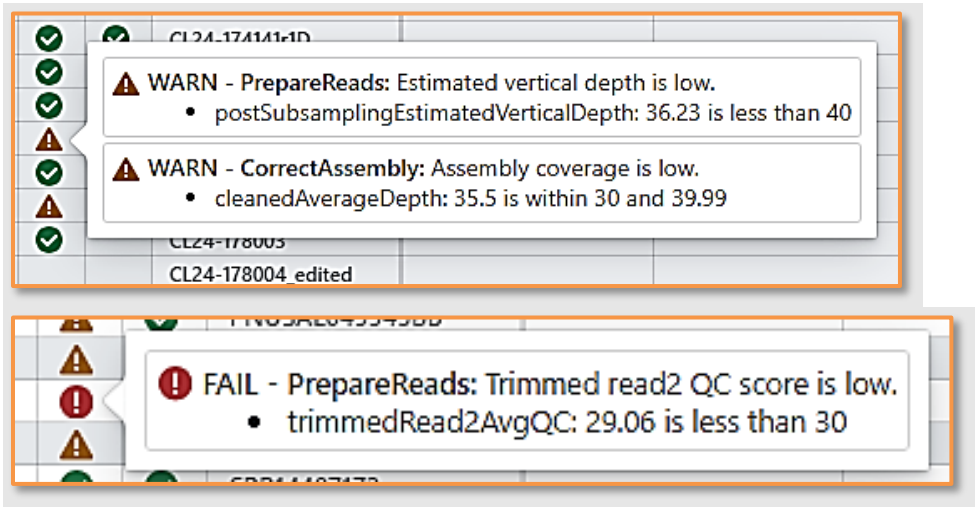
***** *Shigella* species are identified as *Escherichia coli* by ANI.

6.2.2. Analyzed entries will display one of these five status icons:

Icon	Status	Definitions
	Pipeline Completed	Quality is good, sequence data can be published.
	Pipeline Completed with Errors	Pipeline did not complete and needs to be restarted or troubleshooted.
	Warn	Sequence is passing but has minor quality issues; we recommend investigating the reason(s) for warning(s) before publishing.

	Pipeline Cancelled	Pipeline was cancelled by the user and can be restarted.
	Fail	Sequence failed quality and cannot be published; will likely need to be re-sequenced.

NOTE: You can hover over these icons to see further details.

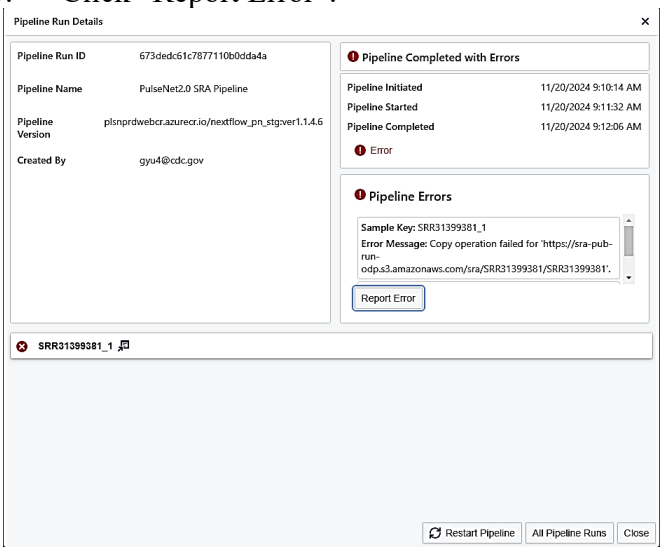


6.2.3. For any jobs that have status "Pipeline Completed with Errors", capture the error message.

6.2.3.1. Click on "Pipelines" and select "View All Pipeline Runs".

6.2.3.2. Select the pipeline that had errors.

6.2.3.3. Click "Report Error".



6.2.3.4. Click "Send" on the email to send the error message to PulseNet@cdc.gov for troubleshooting.

6.3. Review Quality Metrics. Refer to PND05.JA1 (PulseNet Job Aid for Illumina Data Quality Control).

6.4. Review Genotyping Results.

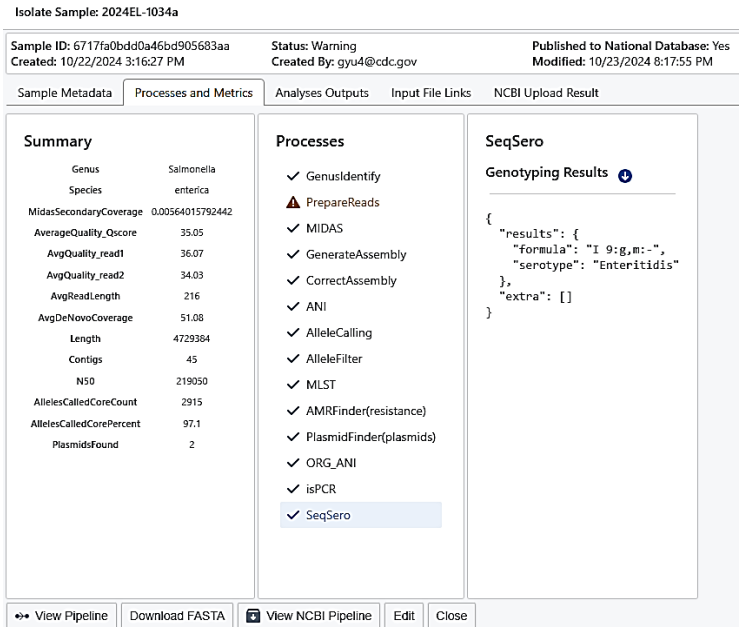
6.4.1. *Salmonella* serotyping – the Genotyper tool populates the following two fields in the grid view:

6.4.1.1. "AntigenForm_wgs": antigenic formula detected by SeqSero2.

6.4.1.2. "Serotype_wgs": serotype name determined by comparing the antigenic formula to a lookup 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.

Key	AntigenForm_wgs	Serotype_wgs
2024EL-1030a	I 9:g,m:-	Enteritidis
2024EL-1031a	I 9:g,m:-	Enteritidis
2024EL-1032a	I 9:g,m:-	Enteritidis
2024EL-1033a	I 9:g,m:-	Enteritidis
2024EL-1034a	I 9:g,m:-	Enteritidis
2024EL-1035a	I 9:g,m:-	Enteritidis
2024EL-1036a	I 9:g,m:-	Enteritidis
2024EL-1037a	I 9:g,m:-	Enteritidis

6.4.1.3. *Salmonella* genotyper information can also be observed by double-clicking the "Key" field of an entry to view the "Isolate Sample" window, clicking on the "Processes and Metrics" tab and selecting "SeqSero" under "Processes".



6.4.1.3.1. The "Serotype_wgs" field will be populated in the grid view with the antigenic formula if the serotype is not identified.

6.4.2. *Escherichia* serotyping – the Genotyper tool populates the "Serotype_wgs" field with O and/or H antigens based on the CGE SerotypeFinder v2 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_wgs
PNUSAE202104	O13/O129/O135:H14
PNUSAE202105	O26:H11
PNUSAE202106	O13/O129/O135:H14
PNUSAE202107	O157:H7
PNUSAE202108	O26:H11
PNUSAE202109	O26:H11
PNUSAE202110	O26:H11
PNUSAE202111	O103:H25
PNUSAE202112	O157:H7
PNUSAE202113	O182:H21
PNUSAE202114	O15:H45

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

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

O106/O17/O77:H45
O17/O44/O77:H18
O169/O183:H41
O123/O186:H2
O123/O186:H11
O118/O151:H16
O13/O129/O135:H14

6.4.2.3. Sequences with missing O antigens or multiple O antigens can be published to the national database.

6.4.3. *Escherichia* pathotype – the Genotyper tool populates the "Pathotype" field in the PulseNet 2.0 grid view based on the combination of genes detected by the VirulenceFinder v2 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	Adherence and invasion
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

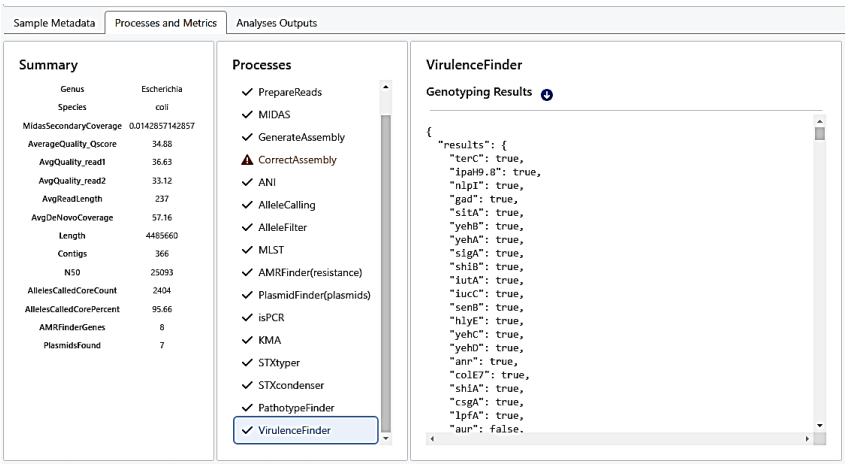
Table2. Escherichia pathotypes assigned based on the combination of genes detected by the VirulenceFinder v2 tool

6.4.4. *Escherichia* group – The Escherichia group is populated automatically and indicates whether the sequence is from STEC O157, STEC NonO157, Non STEC EC, *Shigella sonnei*, *Shigella flexneri, Shigella dysenteriae, Shigella boydii, Shigella other*. Occasionally, "CHECK" will populate before the classification, which is an indication to the data analyst at PulseNet Central to manually determine and input the correct group.
NOTE: *Escherichia* group is predicted based on sequence characteristics. It may not be taxonomically accurate and may differ from species identification based on traditional methods.

Escherichia_group	Pathotype	Toxin_wgs
Shigella flexneri	EIEC/Shigella	
Shigella sonnei	EIEC/Shigella	
STEC NonO157	STEC	stx1a
Shigella sonnei	EIEC/Shigella	
Shigella flexneri	EIEC/Shigella	
STEC NonO157	STEC	stx1a
Shigella flexneri	EIEC/Shigella	
STEC O157	STEC	stx2a; stx2c
STEC NonO157	STEC	stx1a
STEC NonO157	STEC	stx1a
STEC NonO157	STEC	stx1a
STEC NonO157	STEC	stx1a
STEC O157	STEC	stx2a; stx2c
STEC NonO157	STEC	stx2a; stx2d

6.4.5. *Escherichia* stx subtype – the Genotyper tool populates the "Toxin_wgs" field in the grid view based on the stx1 and stx2 subtypes detected by the *in silico* PCR tool.
NOTE1: 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.
NOTE2: Stx subtypes identified using the current version of the tool includes stx1a, stx1c, stx1d, stx2a, stx2b, stx2c, stx2d, stx2e, stx2f, and stx2f. Other rare/newer subtypes that either already exist/could potentially be discovered are not identified by the stxtyper at this point in which Toxin_wgs will be blank; however, the Pathotype field has "STEC" as a pathotype because the pathotype field is populated based on the VirulenceFinder results and the VirulenceFinder is capable of detecting the newest subtypes.

6.4.6. *Escherichia* genotyper information can also be observed by double-clicking the entry to view the "Isolate Sample" window, clicking on the "Processes and Metrics" tab and selecting "STXtyper", STXcondenser", "PathotypeFinder", or "VirulenceFinder" under "Processes".

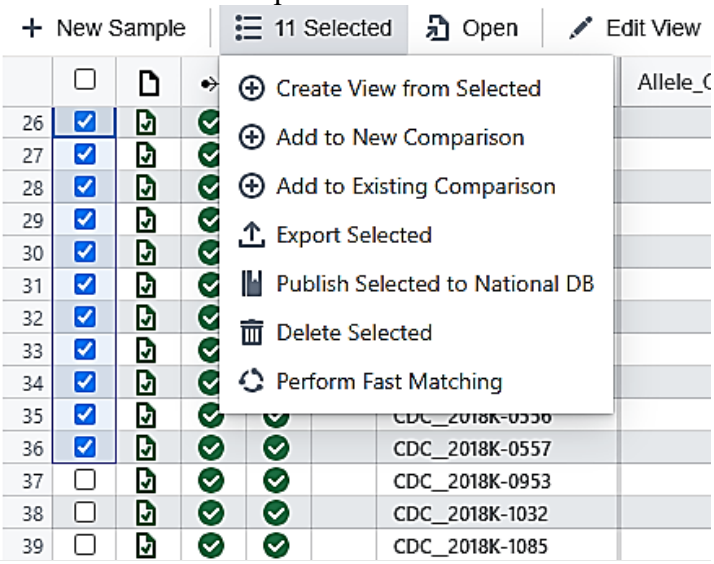


6.4.7. *Vibrio* serotyping – the Genotyper tool populates the "Serotype_wgs" field The O1 Serogroup (either O1 or O139) based on a custom built tool.
NOTE: "Serotype_wgs" is only completed for *V. cholerae.*

6.4.8. The experiments for resistance and plasmids can be viewed in two ways:

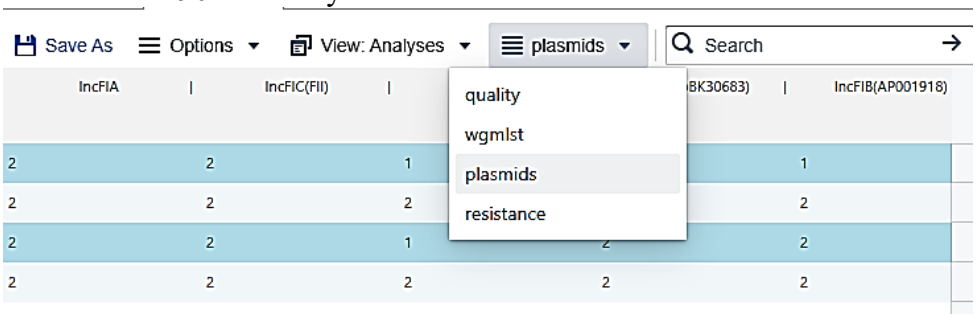
6.4.8.1. Multiple samples at the same time in a comparison:

6.4.8.1.1. Select the entries in PulseNet 2.0. Choose “X Selected” and click “Add to New Comparison”.



6.4.8.1.2. In the comparison window, change “quality” to “plasmids” or “resistance” using the dropdown menu.

NOTE: 1= gene present with >90% identity to the reference, 2= gene absent or <90% identity to the reference

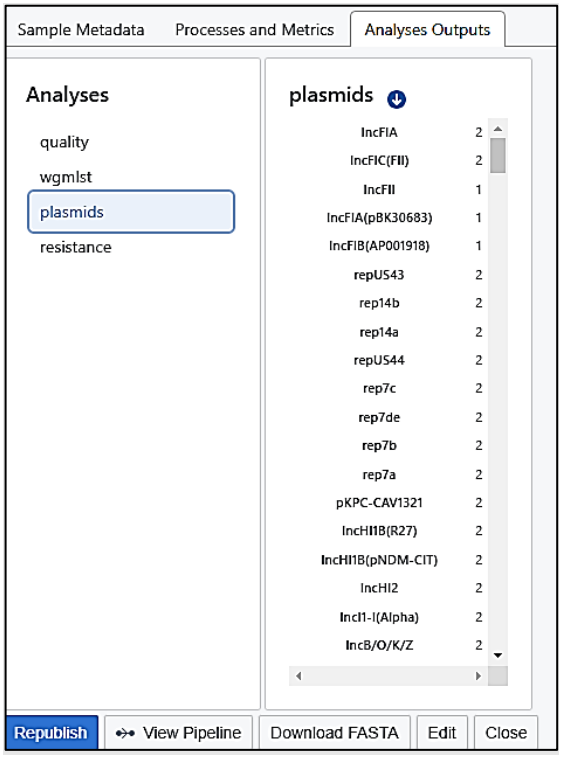


6.4.8.1.3. If desired, you can export the values as a .csv file by clicking “Options” and “Export Analysis”.

6.4.8.2. One sample at a time:

6.4.8.2.1. Double-click on an entry to open the “Isolate Sample” window.

6.4.8.3. Click on the “Analyses Outputs” tab and select “plasmids” or “resistance” under “Analyses”. Click on the down arrow to export these metrics as a .csv file.



6.4.9. *Listeria* lineage – the Genotyper tool will populate the "Lineage" field in the grid view.

Genus	Species	Lineage	ANICoverage	ANI
Listeria	monocytogenes		93.9041	99.5184
Listeria	monocytogenes	I	95.7204	99.5135
Listeria	monocytogenes	I	94.4139	99.4805
Listeria	monocytogenes	I	95.2519	99.5062
Listeria	monocytogenes		94.6926	99.4942
Listeria	monocytogenes	I	94.8022	99.483
Listeria	monocytogenes	I	95.2445	99.5095
Listeria	monocytogenes	I	94.3862	99.4777
Listeria	monocytogenes	III	97.0993	99.0259
Listeria	monocytogenes	III	97.0986	99.0257
Listeria	monocytogenes	I	95.3506	99.4906
Listeria	monocytogenes	II	95.3547	98.7849

6.5. Review 7-Gene MLST Results.

6.5.1. 7-gene MLST schemes available in different organism-specific databases:

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

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

6.5.1.3. *Listeria:* Pasteur

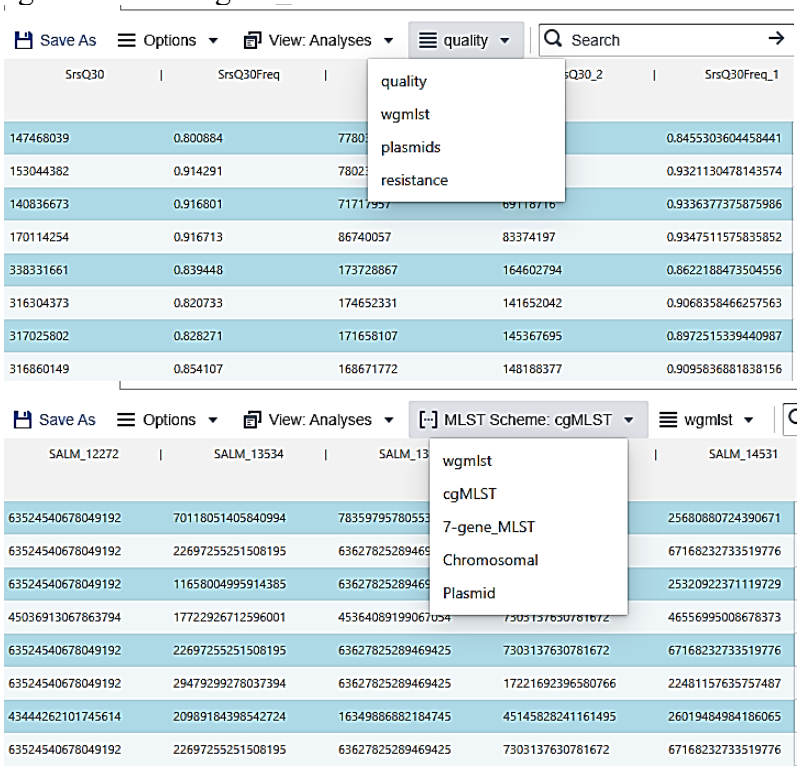
6.5.1.4. *Salmonella:* Achtman

6.5.1.5. *Vibrio*: PubMLST

6.5.2. "MLST_ST" and/or "MLST_CC" fields will be automatically populated with the public ST nomenclature using the PulseNet 2.0 pipeline.

Genus	Species	MLST_CC	MLST_ST
Campylobacter	jejuni	ST-48 complex	ST918
Campylobacter	jejuni	ST-48 complex	ST918
Campylobacter	jejuni	ST-48 complex	ST918
Campylobacter	coli	ST-828 complex	ST1068
Campylobacter	coli	ST-828 complex	ST1068
Campylobacter	coli	ST-828 complex	ST1068
Campylobacter	coli	ST-828 complex	ST1068
Campylobacter	coli	ST-828 complex	ST1068
Campylobacter	jejuni	ST-45 complex	ST45
Campylobacter	jejuni	ST-52 complex	ST52
Campylobacter	jejuni	ST-21 complex	ST8
Campylobacter	jejuni	ST-179 complex	ST5453
Campylobacter	jejuni	ST-21 complex	ST8
Campylobacter	jejuni	ST-21 complex	ST8
Campylobacter	jejuni	ST-21 complex	ST8
Campylobacter	jejuni	ST-21 complex	ST8

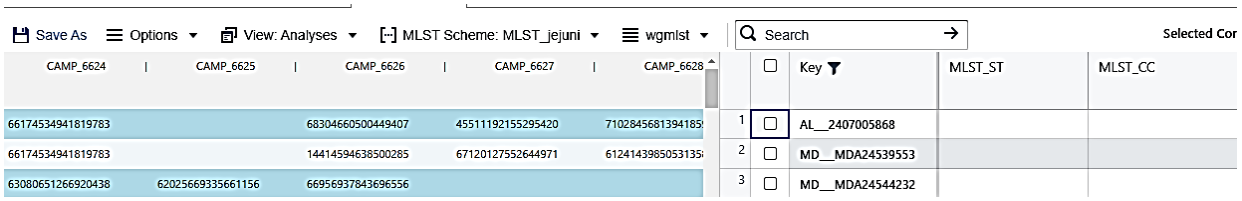
6.5.2.1. Unmatched/undetected sequence types are left blank. You can further investigate the reason for this by creating a comparison for the isolates with missing STs, now you see only the 7 housekeeping genes. In the comparison, change “quality” to “wgMLST” and change “MLST Scheme: cgMLST” to “7-gene_MLST”.



The ST is usually left blank for the following reasons:

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

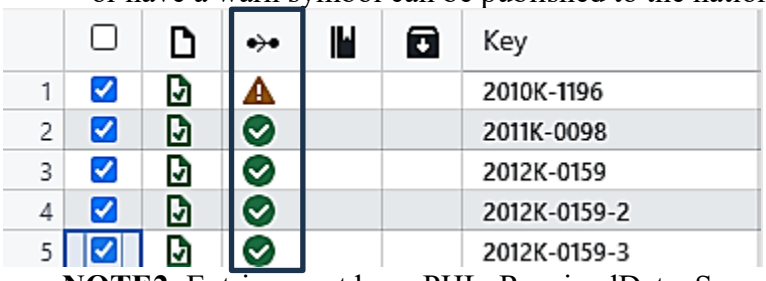
6.5.2.1.2. One or more of the genes are missing from the profile.



6.6. Publish to the National Database.

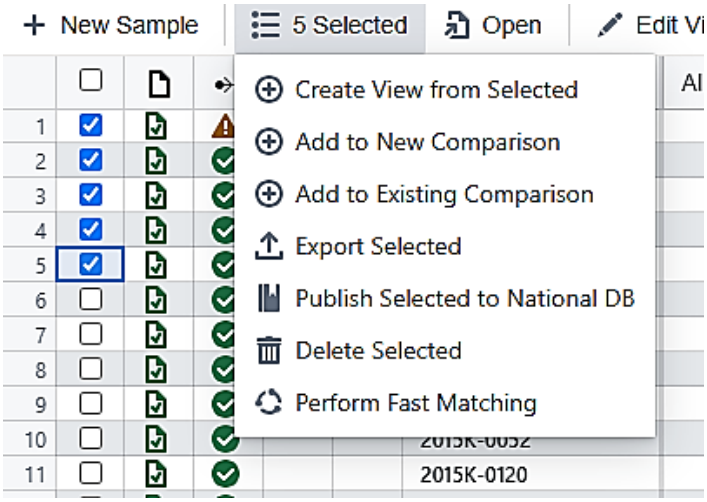
6.6.1. Select the entries to be published to the national database.

NOTE1: Make sure the sequence has successfully completed the pipeline by looking at the pipeline status. Only sequences that pass the pipeline successfully or have a warn symbol can be published to the national database.



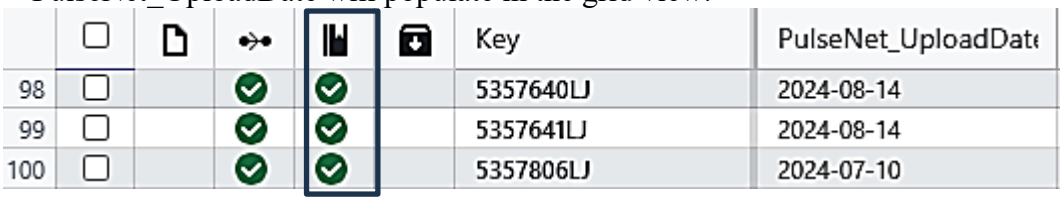
NOTE2: Entries must have PHL_ReceivedDate, SourceType, and PatientSex (human samples only) to be published to the national database.

6.6.2. Click on the "X Selected" menu option and choose "Publish Selected to National DB".



6.6.3. A confirmation window will open to confirm the publishing of the sample(s) to the national database. Click "OK."

6.6.4. A notification will appear that the sample(s) has/have been published successfully. The published icon in the grid view will have a green check and the PulseNet_UploadDate will populate in the grid view.

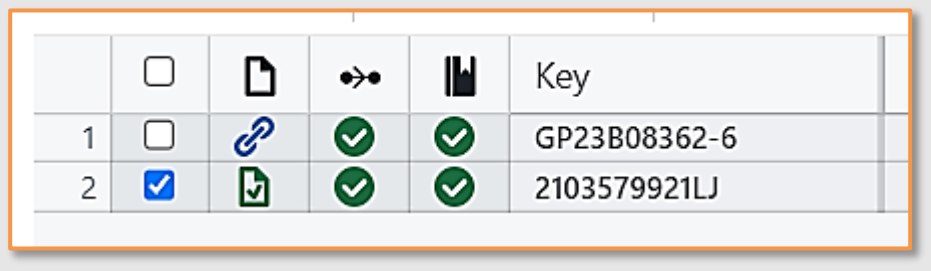


NOTE1: Once samples are published to the NDB, most fields are no longer editable by your lab.
NOTE2: There are a handful of fields which will remain editable for PHLs after publishing to the NDB: Purpose_of_sampling, Animal_env, Food_origin, Intended_consumer, Food_processing_method, Facility_type, Food_type_processed, Env_local_scale, Env_medium, Local_Outbreak_Code, PHL_Comment1, PHL_Comment2, CIDT, NCBI_Submission_Date
NOTE3: Other edits to samples require them to be unpublished by a CDC database manager. Contact PulseNet to make edits by sending an email to PulseNet@cdc.gov.

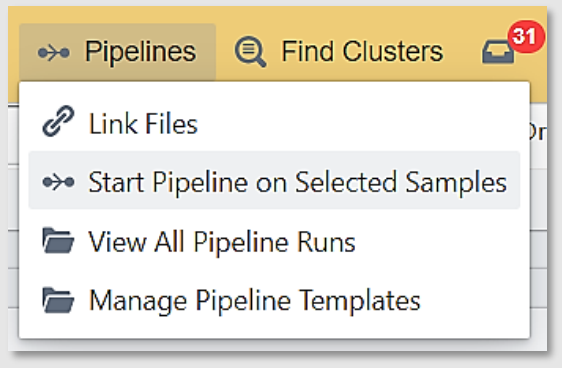
6.6.5. Outbreak codes, REP codes, and allele codes are automatically copied to the local database views when changes are made in the national database view.

6.7. Upload to NCBI.

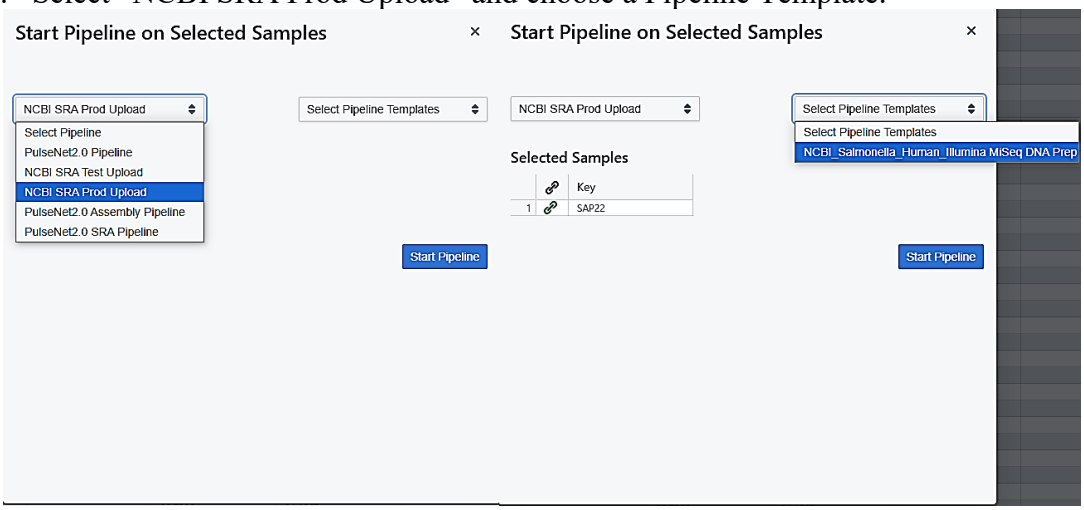
6.7.1. In the PulseNet 2.0 grid view, select the entries to be uploaded. **NOTE1:** verify that the entries have the "WGS_id" field populated. **NOTE2:** Before submitting sequences to NCBI, verify that data is not already uploaded to NCBI. The green page symbol indicates that files are linked locally or on BaseSpace and have not received an SRR_id. The blue link symbol indicates that an SRR_id is already present in the database. Choose entries with the green page symbol and have been published to the national database.



6.7.2. Select "Start Pipeline on Selected Samples" in the Pipelines menu.



6.7.3. Select "NCBI SRA Prod Upload" and choose a Pipeline Template.



NOTE: See Appendix PND05-3 to create NCBI Submission Templates.

6.7.4. Verify that metadata is complete and correct.

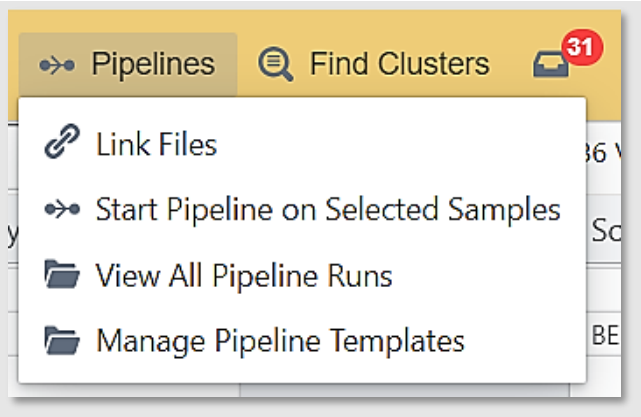
NOTE: Missing metadata will produce the error shown below and will require you to update that content.

6.7.5. Click "Start Pipeline". When the pipeline is successfully initiated, the "Pipeline started successfully" message will appear.

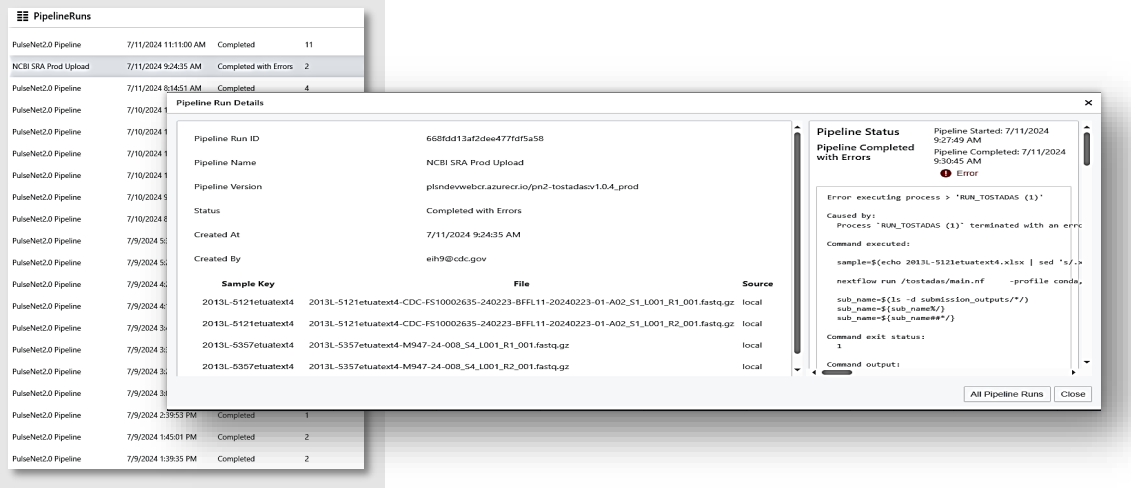
6.7.6. Once the pipeline completes successfully, the SAMN and SRR accession numbers will auto-populate in the NCBI_ACCESSION and SRR_id database fields and synchronize to the national database. Green checkmarks will appear in the "NCBI Upload Status" column.

Key	WGS_id	NCBI_ACCESSION	SRR_id
2021K-0250			
2021K-0656			
2024EL-1030a	2024EL-1030a	SAMN44406829	SRR31094052
2024EL-1031a	2024EL-1031a		
2024EL-1032a	2024EL-1032a		
2024EL-1033a	2024EL-1033a	SAMN44406906	SRR31094231
2024EL-1034a	2024EL-1034a	SAMN44406886	SRR31094201
2024EL-1035a	2024EL-1035a	SAMN44406960	SRR31094433
2024EL-1036a	2024EL-1036a	SAMN44406905	SRR31094214
2024EL-1037a	2024EL-1037a	SAMN44406924	SRR31094233

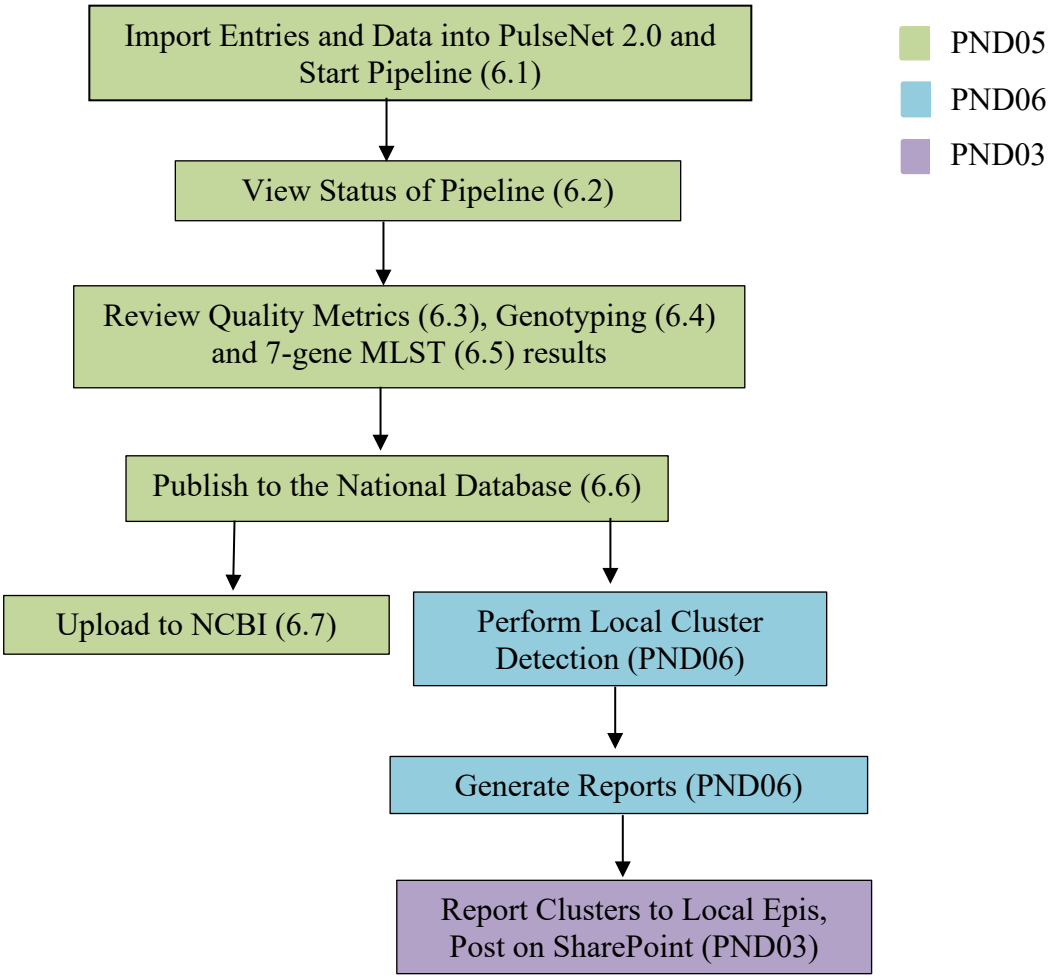
6.7.7. To check the status of a pipeline or if a pipeline errors, click on "View All Pipeline Runs" under the "Pipelines" menu option.



6.7.8. Double-click on the "NCBI SRA Prod Upload Pipeline" that has an error and email the Pipeline Run ID to PulseNet@cdc.gov.



7. FLOW CHART:


Import Entries and Data into PulseNet 2.0 and Start Pipeline (6.1)
View Status of Pipeline (6.2)
Review Quality Metrics (6.3), Genotyping (6.4) and 7-gene MLST (6.5) results
Publish to the National Database (6.6)
Upload to NCBI (6.7)
Perform Local Cluster Detection (PND06)
Generate Reports (PND06)
Report Clusters to Local Epis, Post on SharePoint (PND03)
PND05
PND06
PND03

8. RELATED DOCUMENTS:

Document Number	Title
PNQ08	PulseNet WGS Certification SOP
PND03	PulseNet SharePoint Access and Use SOP
PND06	Local Cluster Detection SOP
PND05.JA1	PulseNet Job Aid for Illumina Data Quality Control

9. REFERENCES:

10. CONTACTS:

10.1. CDC PulseNet Response and Outbreak Management 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

11.3. 03/10/2025

–Revised to use PulseNet 2.0 workflow for sequence data analysis

–Updated Related Document PNQ07: Illumina Sequence Data SOP to be a job aid to PND05 (PND05.JA1)

11.4. 05/10/2025

–Updated Table 1: Organism identification for Reference ID and Organism specific databases.

–Added Table numbering throughout document.

12. APPROVAL SIGNATURES:

Approved By: PATRICIA C. LAFON -S Digitally signed by PATRICIA C. LAFON -S Date: 2025.05.13 10:40:29 -04'00' Date: 05/13/2025

PulseNet QA/QC Personnel

Approved By: Kelley B. Hise -S Digitally signed by Kelley B. Hise -S Date: 2025.05.13 06:55:58 -04'00' Date: 05/13/2025

PulseNet Response and Outbreak Management Team Lead

Approved By: Peyton A. Smith -S Digitally signed by Peyton A. Smith -S Date: 2025.05.13 09:37:35 -04'00' Date: 05/13/2025

PulseNet WGS Laboratory Technical Lead

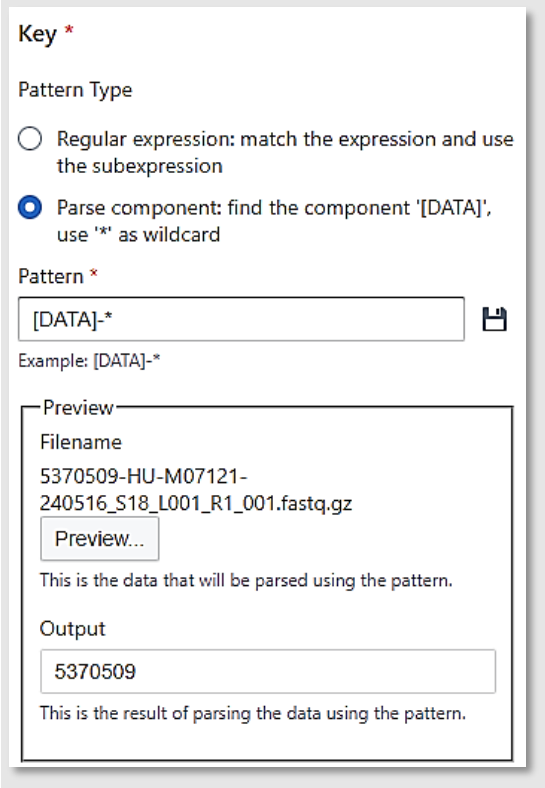
Approved By: CHRISTINE C. LEE -S Digitally signed by CHRISTINE C. LEE -S Date: 2025.05.12 22:00:41 -04'00' Date: 05/12/2025

PulseNet Reference Outbreak and Surveillance Team Lead

13. APPENDICES:

Appendix PND05-1. Create and save import rules/parsing templates for "Key" and "SEQUENCERRUN_ID"

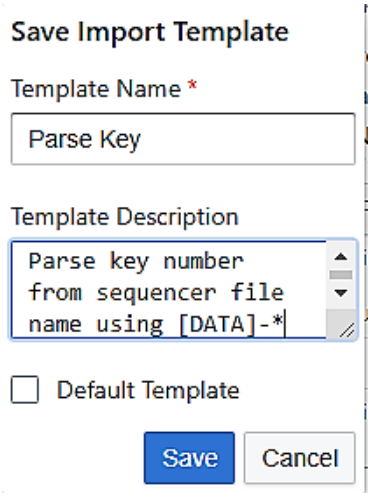
1. Under "Import Rules" of "Start Pipeline", select "Parse Component..." under the pattern type for Key.
2. Enter "[DATA]-*" in the "Pattern" field and click "Preview…" to verify the key is parsed out from the sequence file name.
   NOTE: See table below for PulseNet Key parsing options.



PulseNet Key	Parsing Pattern[1]
No dashes	[DATA]-*
One or more dashes	[DATA]-X*
With or without dashes when BaseSpace Toolkit is used	[DATA]_*

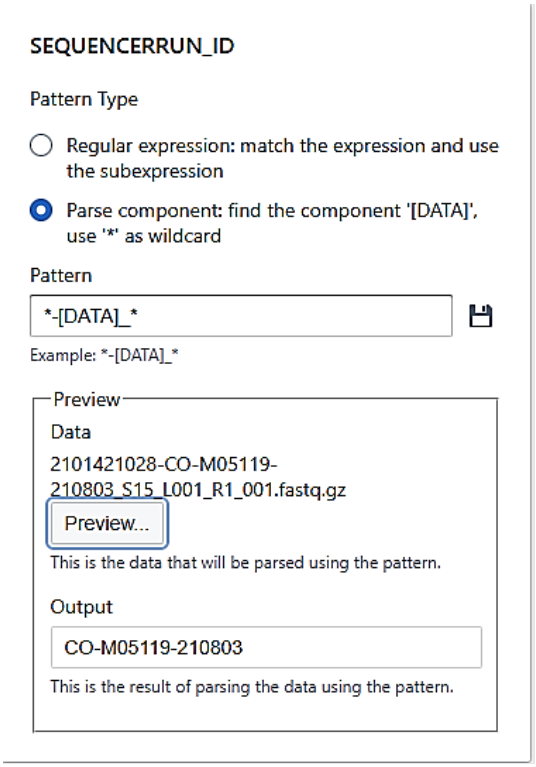
[1] X= The first letter of the LabID

3. Save the import rule template by clicking on the disk icon, entering a "Template Name", and a "Template Description". Click "Save".
   NOTE: Template description is optional and checking "Default Template" enables this template to be automatically chosen in the "Import Rules" step.



NOTE: [DATA]-* is used when parsing a key that does not have dashes. If using a key that has dashes (i.e., 2024D-0024), use the parsing string [DATA]-M*.

4. Select "Parse Component" under the pattern type for SEQUENCERRUN_ID.
5. Enter "*-[DATA]_*" in the "Pattern" field and click "Preview…" to verify the sequencer run ID is parsed out from the sequence file name.
   NOTE: See Table 3 below for PulseNet SequencerRun_Id parsing options.



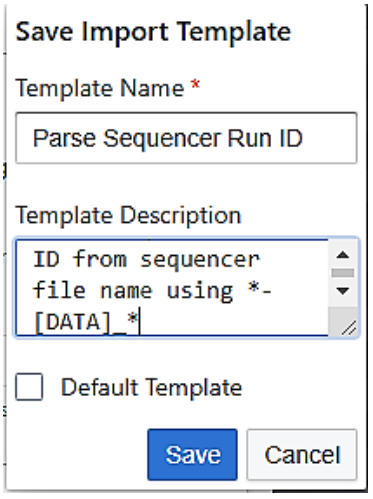
PulseNet Key	Sequencing Platform	Parsing Pattern
No dashes	MiSeq, MiniSeq, iSeq, NextSeq with no BaseSpace Toolkit used	*-[DATA]_*
One dash	MiSeq, MiniSeq, iSeq, NextSeq with no BaseSpace Toolkit used	*-*-[DATA]_*
Two dashes	MiSeq, MiniSeq, iSeq, NextSeq with no BaseSpace Toolkit used	*-*-*-[DATA]_*
Three dashes	MiSeq, MiniSeq, iSeq, NextSeq with no BaseSpace Toolkit used	*-*-*-*-[DATA]_*
With or without dashes when BaseSpace Toolkit is used	Utilize the "Use the project name as the SequencerRun_Id" option on the "Link Files" screen	N/A[1]
No dashes[2]	ClearLabs	*-[DATA]-B*
One dash[2]	ClearLabs	*-*-[DATA]-B*
Two dashes[2]	ClearLabs	*-*-*-[DATA]-B*
Three dashes[2]	ClearLabs	*-*-*-*-[DATA]-B*

[1]Not applicable

[2] In ClearLabs run set up, it is not possible to define the SequencerRun_Id using the standardized PulseNet format. Instead, the Laboratory ID needs to be added AFTER the PulseNet Key separated by dash. B is the first letter of the Clearlabs instrument Id.

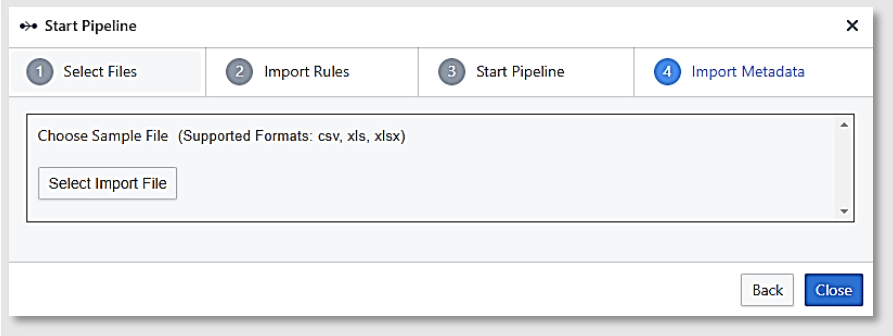
Table 3. PulseNet SequencerRun_Id parsing options

6. Save the import rule template by clicking on the disk icon, entering a "Template Name", and a "Template Description". Click "Save".
   NOTE: Template description is optional and checking "Default Template" enables this template to be automatically chosen in the "Import Rules" step.

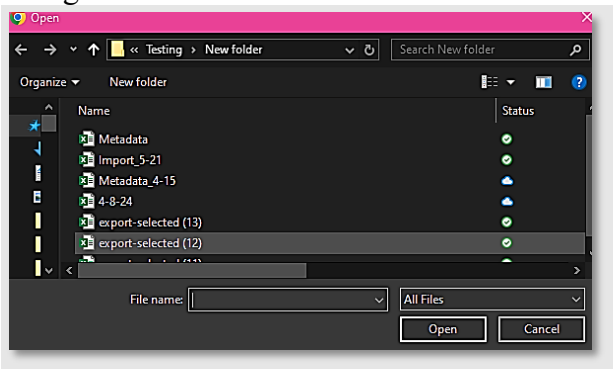


Appendix PND05-2. Instructions to Import Metadata

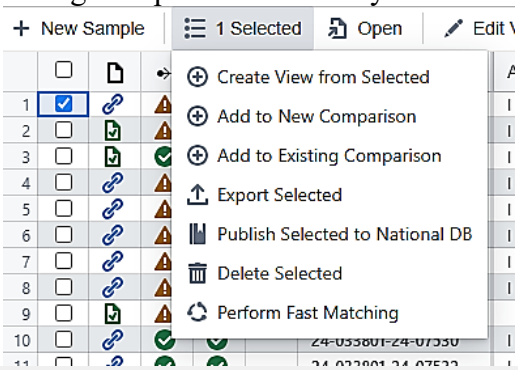
1. Click “Select Import File”.



2. Navigate to the folder location of the metadata file, click “Open”



NOTE1: Metadata import spreadsheets can be generated by exporting the metadata for a single entry from the database to obtain the correct field headings (“X Selected” > “Export Selected”), replacing the metadata with the new metadata (delete the unnecessary columns), and save the spreadsheet OR entering the metadata into the templates found on SharePoint and saving the spreadsheet locally.

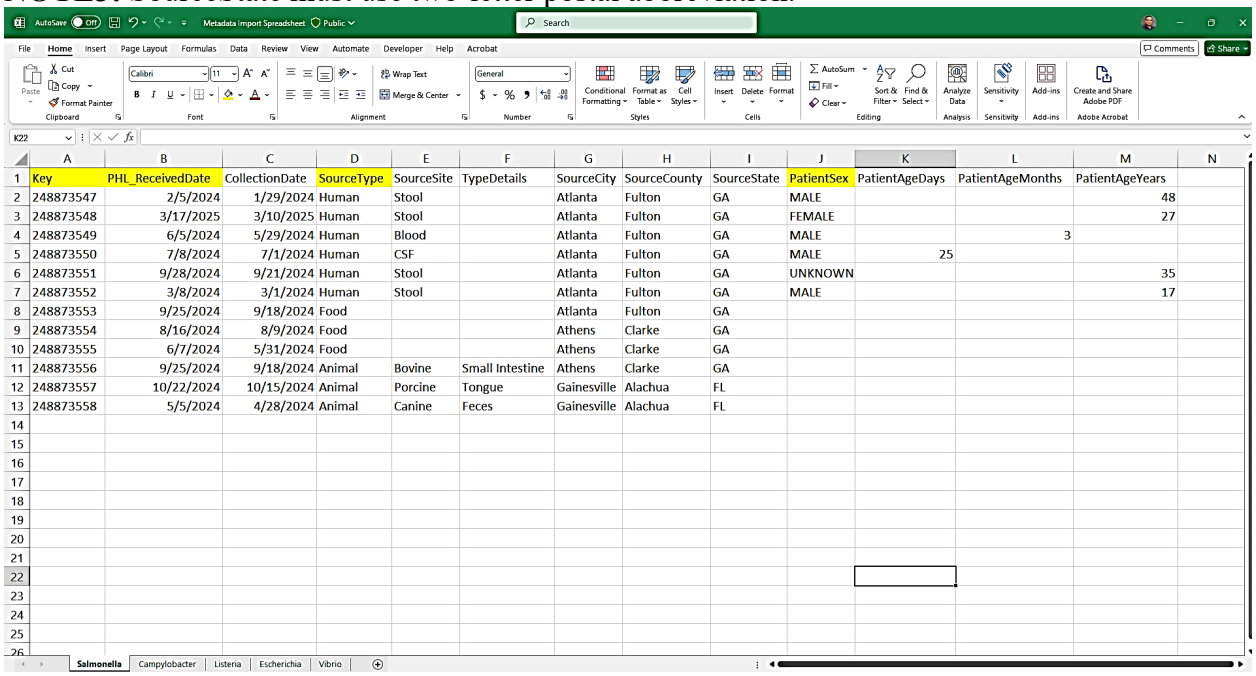


Key	CollectionDate	PatientAgeDays	PatientAgeMonths	PatientAgeYears	PatientSex	PHL_ReceivedDate	SourceCity	SourceCountry	SourceCounty	SourceSite	SourceState	SourceType	TypeDetails
SampleKey1	9/25/2024			35	MALE	10/3/2024	Atlanta	USA	Fulton	Stool	GA	Human	
SampleKey2	10/2/2024		5		FEMALE	10/10/2024	Lilburn	USA	Gwinnett	Stool	GA	Human	
SampleKey3	10/9/2024	10			UNKNOWN	10/17/2024	Marietta	USA	Cobb	Blood	GA	Human	
SampleKey4	10/16/2024					10/24/2024	Macon	USA	Bibb	Chicken	GA	Food	Chicken Breast

NOTE2: At the minimum, the following metadata fields need to be populated in the organism-specific database (highlighted):

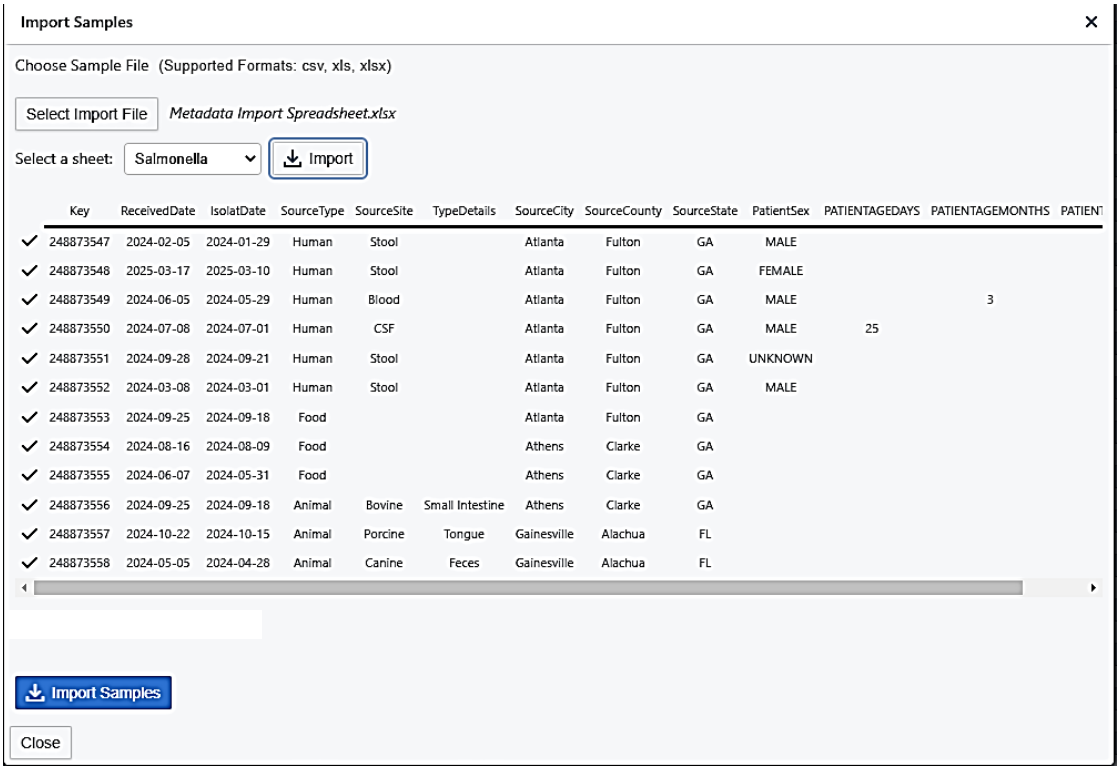
a. SourceType: MUST be Animal, Food, Human, Environmental, or Unknown
b. PatientSex (only if Human SourceType): MUST be MALE, FEMALE, or UNKNOWN.
   NOTE: PatientSex options must be capitalized.
c. PHL_ReceivedDate

NOTE3: SourceState must use two-letter postal abbreviation.

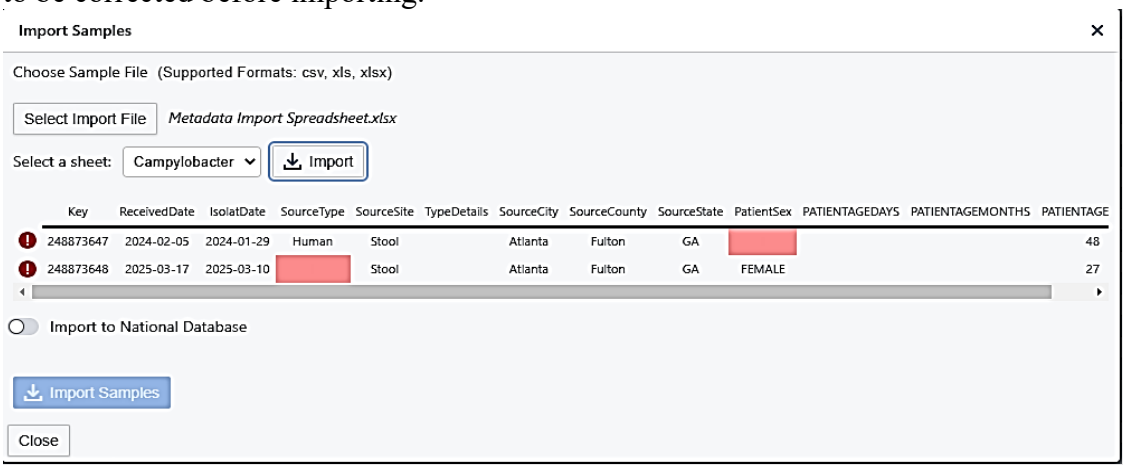


Key	PHL_ReceivedDate	CollectionDate	SourceType	SourceSite	TypeDetails	SourceCity	SourceCounty	SourceState	PatientSex	PatientAgeDays	PatientAgeMonths	PatientAgeYears
248873547	2/5/2024	1/29/2024	Human	Stool		Atlanta	Fulton	GA	MALE			48
248873548	3/17/2025	3/10/2025	Human	Stool		Atlanta	Fulton	GA	FEMALE			27
248873549	6/5/2024	5/29/2024	Human	Blood		Atlanta	Fulton	GA	MALE		3	
248873550	7/8/2024	7/1/2024	Human	CSF		Atlanta	Fulton	GA	MALE	25		
248873551	9/28/2024	9/21/2024	Human	Stool		Atlanta	Fulton	GA	UNKNOWN			35
248873552	3/8/2024	3/1/2024	Human	Stool		Atlanta	Fulton	GA	MALE			17
248873553	9/25/2024	9/18/2024	Food			Atlanta	Fulton	GA				
248873554	8/16/2024	8/9/2024	Food			Athens	Clarke	GA				
248873555	6/7/2024	5/31/2024	Food			Athens	Clarke	GA				
248873556	9/25/2024	9/18/2024	Animal	Bovine	Small Intestine	Athens	Clarke	GA				
248873557	10/22/2024	10/15/2024	Animal	Porcine	Tongue	Gainesville	Alachua	FL				
248873558	5/5/2024	4/28/2024	Animal	Canine	Feces	Gainesville	Alachua	FL				

3. Select the sheet in the file that contains the metadata, click “Import”
4. The samples will appear with a green check mark if all metadata is imported correctly, click “Import Samples”.

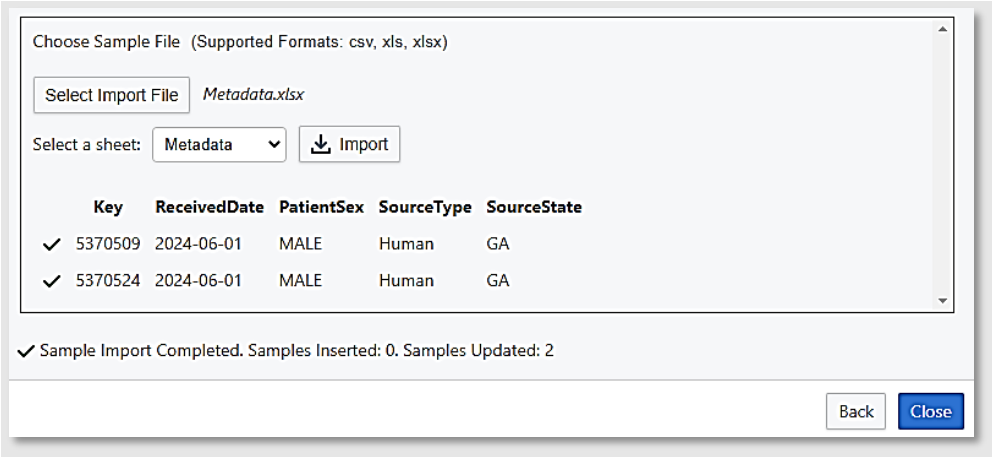


NOTE1: Missing or improperly formatted metadata will be highlighted in red and needs to be corrected before importing.



NOTE2: These screens still use the legacy fields (i.e., ReceivedDate instead of PHL_ReceivedDate and IsolatDate instead of CollectionDate). These field differences will not affect validation of the metadata.

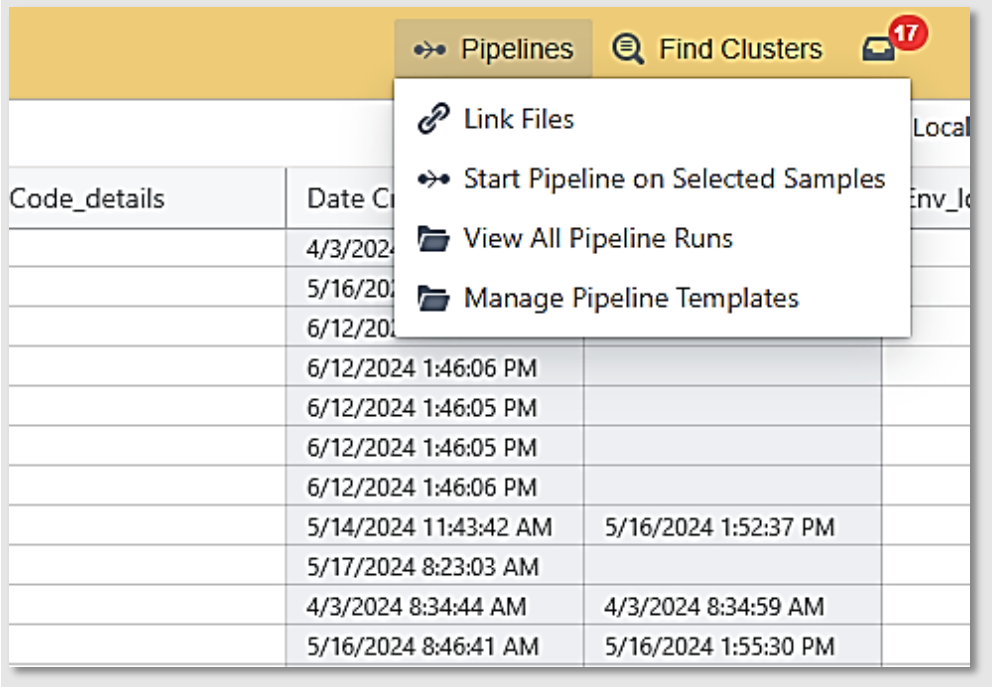
5. After the metadata has imported successfully, click “Close”.


Choose Sample File (Supported Formats: csv, xls, xlsx)
Select Import File
Metadata.xlsx
Select a sheet:
Metadata
Import
Key
ReceivedDate
PatientSex
SourceType
SourceState
5370509
2024-06-01
MALE
Human
GA
5370524
2024-06-01
MALE
Human
GA
Sample Import Completed. Samples Inserted: 0. Samples Updated: 2
Back
Close

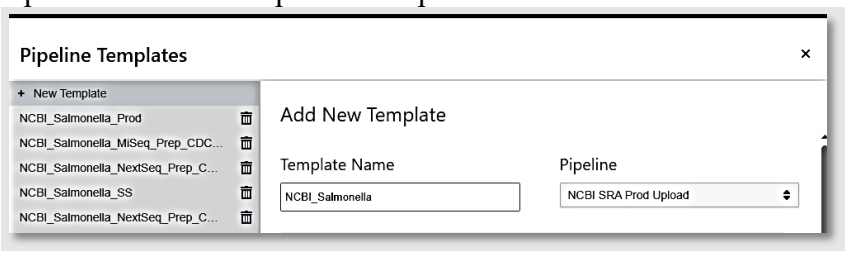
Appendix PND05-3. Instructions to Create NCBI Submission Templates

NOTE: Separate submission templates are required for each organism database and different *Shigella* species

1. In the grid view, click "Pipelines" and choose "Manage Pipeline Templates".

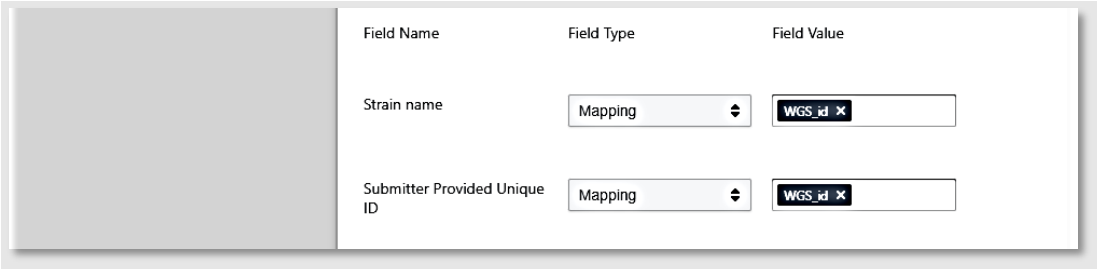


2. Click on "New Template", provide a Template Name, and choose "NCBI SRA Prod Upload" from the "Pipeline" drop-down menu.



NOTE: Existing templates will be visible under "Pipeline Templates" and can be deleted by clicking the trash icon.

3. For Strain name and Submitter Provided Unique ID, select "Mapping" for "Field Type" and choose "WGS_id" as the "Field Value" or alternative field if using a different strain identification number.



4. For SPUID name space, select "Value" for "Field Type" and enter "EDLB-CDC" as the "Field Value" or GenomeTrakr required contact information.
5. For BioProject accession, select "Value" for "Field Type" and enter a PulseNet Bioproject accession from the list below as the "Field Value" or a different accession number if known.

Organism	**Bioproject accession number**
Campylobacter	PRJNA239251
Escherichia and *Shigella*	PRJNA218110
Listeria	PRJNA212117
Salmonella	PRJNA230403
Vibrio	PRJNA266293
Cronobacter	PRJNA420465
C. botulinum	PRJNA428620
Yersinia	PRJNA275974

Table 4. NCBI Bioproject Accesssion Numbers for PulseNet organisms

6. Enter the following information into the template.
   NOTE: The difference in attribute (Human vs non-human) can be determined by hovering over the question mark.
   a. Author: "Field Type"=Value; "Field Value"=CDC
   b. Serovar: "Field Type"=Mapping; "Field Value"=SEROTYPE_WGS
      NOTE: For sequences that do not have serotypes (i.e., *Campylobacter*), use "Field Type"=Value; "Field Value"=missing
   c. SourceType: "Field Type"=Mapping; "Field Value"=SourceType
   d. Isolate Name Alias (if SourceType equals Human): "Field Type"=Value; "Field Value"=missing
   e. Isolate Name Alias (if SourceType does not equal Human): "Field Type"=Mapping; "Field Value"=Key
   f. Isolation source (if SourceType equals Human): "Field Type"=Value; "Field Value"=missing
   g. Isolation source (if SourceType does not equal Human): "Field Type"=Mapping; "Field Value"=SourceSite
   h. Geographical origin (if SourceType equals Human): "Field Type"=Value; "Field Value"=USA or "Field Type"=Mapping; "Field Value"=SourceCountry
   i. Geographical origin (if SourceType does not equal Human): "Field Type"=Value; "Field Value"=USA:*state*
   j. Organism name: "Field Type"=Mapping; "Field Value"=GENUS SPECIES

NOTE: The organism name for *Shigella* species should be entered as "Field Type"=Value and "Field Value"=*genus and species name*. For example, a *Shigella flexneri* NCBI template should have "Field Type"=Value and "Field Value"=Shigella flexneri. A separate template should be created and used for different *Shigella* sequences to be uploaded to NCBI.

k. Collection/Isolation date (if SourceType equals Human): "Field Type"=Mapping; "Field Value"=CollectionDate (YYYY)
l. Collection/Isolation date (if SourceType does not equal Human): "Field Type"=Mapping; "Field Value"=CollectionDate (YYYY-mm)
m. Collected by (if SourceType equals Human): "Field Type"=Value; "Field Value"=CDC)
n. Collected by (if SourceType does not equal Human): "Field Type"=Mapping; "Field Value"=LabID)
o. Instrument model: "Field Type"=Mapping; "Field Value"=Instrument_model
NOTE: This option can be used when the instrument model is populated in the Instrument_model field of PulseNet 2.0. Alternatively, this can be set to: "Field Type"=Value; "Field Value"=*Choose instrument model from drop down menu.* Individual templates will need to be created for each instrument if using this method. When using ClearLabs, iSeq should be chosen as the instrument model, NOT ClearLabs.
p. Library strategy: "Field Type"=Value; "Field Value"=WGS
q. Library source: "Field Type"=Value; "Field Value"=GENOMIC
r. Library selection: "Field Type"=Value; "Field Value"=RANDOM
s. Library layout: "Field Type"=Value; "Field Value"=PAIRED
t. Library name: "Field Type"=Mapping; "Field Value"=Library
NOTE: This option can be used when the library preparation method is populated in the Library field of PulseNet 2.0. Alternatively, this can be set to: "Field Type"=Value; "Field Value"=*Choose Library from drop down menu.* Individual templates will need to be created for each library preparation if using this method.
u. Project name: "Field Type"=Mapping; "Field Value"=Project
NOTE1: This option can be used when the project name is populated in the Project field of PulseNet 2.0. Alternatively, this can be set to: "Field Type"=Value; "Field Value"=*Enter Project name*
NOTE2: Project name can be PulseNet, GenomeTrakr, NARMS.
v. Sequenced by (if SourceType equals Human): "Field Type"=Value; "Field Value"=CDC
w. Sequenced by (if SourceType does not equal Human): "Field Type"=Value; "Field Value"=*enter your full laboratory name*

7. Additional fields can be added to the template by clicking "Add Field" and entering the field name, choosing the "Field Type" and "Field Value".
NOTE: Additional fields are added to align with the One Health Enteric Package at NCBI, see Populating metadata templates for NCBI submissions using PulseNet 2.0
8. Click "Create".

Appendix PND05-4. Instructions for Resubmitting Sequence Data to NCBI

1. Before resubmitting, ask NCBI to retract the Biosample (NCBI_ACCESSION) and SRA (SRR_id) by contacting the Biosample and SRA helpdesk at biosamplehelp@ncbi.nlm.nih.gov, sra@ncbi.nlm.nih.gov, and pd-help@ncbi.nlm.nih.gov.
   NOTE: E-mail all three groups when requesting a biosample and SRA retraction.
2. Relink the entry to the new sequence files and upload to NCBI following steps 6.1-6.3 and 6.7.
   NOTE: This method will NOT link your sample to the same Biosample and SRR identification numbers. Rather, new Biosample and SRR identification numbers will be produced.