TABLE OF CONTENTS – HYPERLINKS TO THE PROCEDURE

1. **PURPOSE:** To describe the procedure for analyzing Illumina short read whole genome sequencing (WGS) data to be used for PulseNet International (PNI) surveillance utilizing the cloud-based Terra.Bio platform.

2. **SCOPE:** This procedure applies to all PulseNet personnel that utilize the Terra.Bio platform to analyze Illumina short read WGS data for surveillance activities within the PulseNet International network. This SOP covers sequence and metadata upload to Terra.Bio, sequence quality evaluation, assembly and genotyping workflows and sequence upload to NCBI. Phylogenetic analyses are covered by the SOP PNID02 (PulseNet International Standard Operating Procedure for Phylogenetic Analysis of WGS Data Using the Terra.Bio Platform).

3. **DEFINITIONS/TERMS:**
   - **3.1 ANI:** Average Nucleotide Identity
   - **3.2 BaseSpace:** Illumina cloud-based computing environment for next generation sequencing data analysis, management and storage, including data sharing.
   - **3.3 Bash Commands:** Bash (Bourne Again Shell) is a command-line interface (CLI) shell used extensively in Linux and macOS. Shell is a computer program that allows direct control of a computer's operating system. Bash commands are used to control the computer or operating system without having to navigate menus, options, and windows within the graphical user interface.

3.4 BioProject: A collection of biological data on NCBI related to a single initiative, originating from a single organization or from a consortium.

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

3.6 Contig: A contiguous consensus sequence derived from the assembly of many short, overlapping DNA fragments.

3.7 Coverage: The average number of reads that includes a given nucleotide in the reconstructed sequence.

3.8 Critical Quality Metrics: coverage (after trimming), average quality (Q score before trimming), assembly length, and secondary genus abundance (contamination detection by MIDAS). Sequences not meeting the minimum thresholds/acceptable ranges for these metrics defined in this document should be re-sequenced.

3.9 CSV: Comma Separated Values

3.10 DeNovo assembly: A sequence assembly generated from the short raw reads without the use of a reference genome.

3.11 FASTA: A text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column.

3.12 FASTQ: a text-based format for storing both a biological sequence and its corresponding quality scores.

3.13 GAMBIT: Genomic Approximation Method for Bacterial Identification and Tracking. A bacterial species identification method utilizing k-mer-based algorithm to search against a large reference database of genomes.

3.14 Gzip: A file format and a software application used for file compression and decompression to transfer data quickly over the Internet

3.15 LIMS: Laboratory Information Management System

3.16 Mash Sketching: Mash is a set of tools for creating and using MinHash sketches, a way of turning a genome into a small signature which can be easily compared with other signatures.

3.17 Metadata: A set of data that describes and gives information about other data.

3.18 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.

3.19 N50: The N50 statistic is commonly used as a rough assessment of genomic assemblies. It represents the contig length (in base pairs) for which half of the genome sequence is assembled in contigs larger than or equal to N50 contig size.

3.20 NCBI: National Center for Biotechnology Information

3.21 PNI: PulseNet International

3.22 QA/QC: Quality Assurance/Quality Control

3.23 Q score: The quality score for each individual base position in a sequence, indicating the accuracy of the base call. Phred scores are used, where Q = -10log (Error Probability). The

higher the quality score, the more reliable the base call. A Q30 means a 1 in 1000 likelihood of an incorrect base call at that position.

3.24 Read: A unit of continuous DNA sequence (based pairs) derived by sequencing a part of the fragmented target DNA.

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

3.26 SOP: Standard Operating Procedure

3.27 SRA: Sequence Read Archive

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

3.29 Terra.Bio: Cloud-based sequence analysis platform developed by the Broad Institute of the Massachusetts Institute of Technology and the Harvard University and utilized by Theiagen Genomics (Highlands Ranch, CO, USA) to offer a for-fee common platform for public health laboratories to host, analyze and share data (Libuit *et al.,* 2023).

3.30 TSV: Tab Separated Values

3.31 WGS: Whole Genome Sequencing

4. RESPONSIBILITIES:

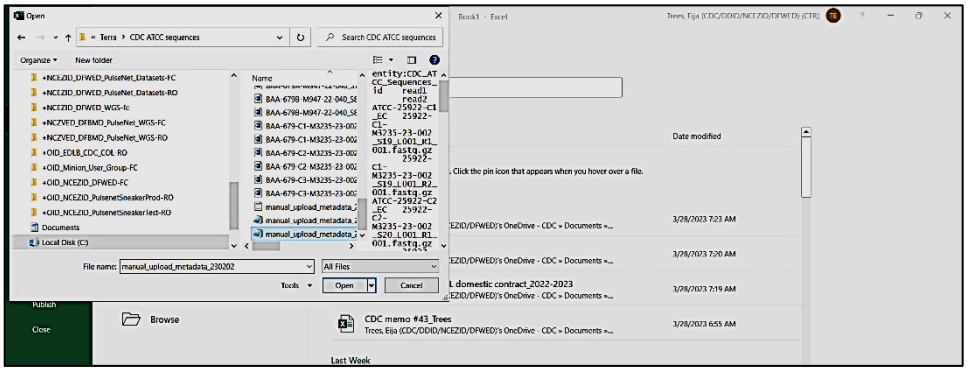
4.1 PulseNet personnel perform quality assessment and genotyping on Illumina short read sequences generated for PulseNet International surveillance utilizing the Terra.Bio platform. Sharing of WGS data with other PNI participants by uploading to NCBI is highly recommended.

5. PROCEDURE:

NOTE: *Sequence data can be imported to Terra in one of the following three ways:* ***(1)*** *Upload from a local network storage (steps 5.1 – 5.3),* ***(2)*** *Direct cloud-to-cloud transfer from Illumina BaseSpace (appendix PNID01-1),* ***(3)*** *Download from NCBI SRA (appendix PNID01-2)*

5.1 **Prepare a metadata file for the sequences to be uploaded to Terra**. The metadata file will link the uploaded FASTQ files to their corresponding database entry names, i.e., entry keys.

5.1.1 Use Excel to open the metadata file template that is in tsv format.



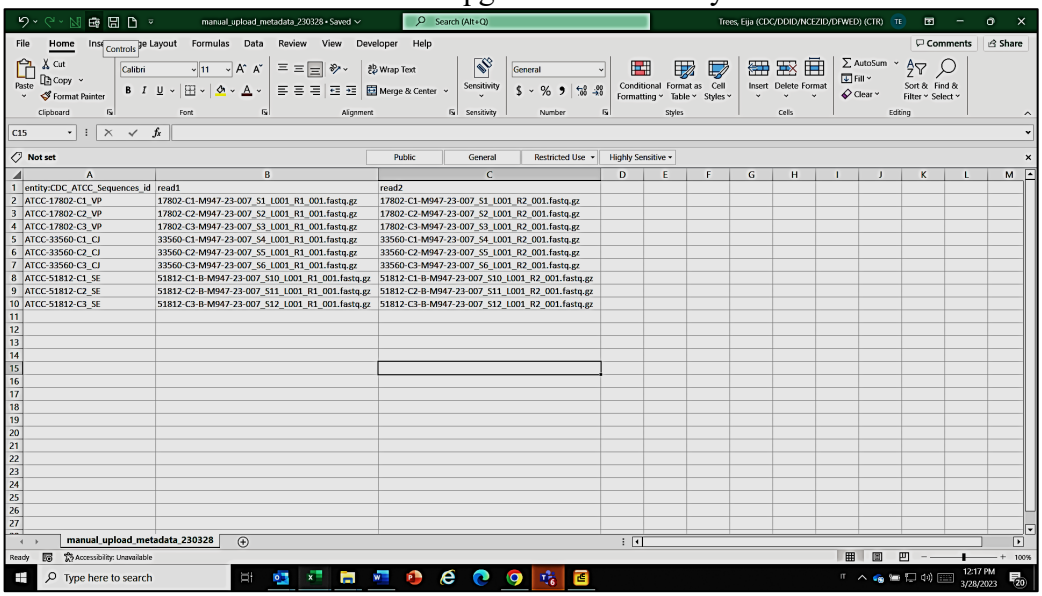
5.1.2 The following three columns are required at a minimum:

5.1.2.1 **Entity:collection_name_id**. This is the name of the Terra data collection (data table) into which you want to upload your sequences, e.g., entity:CDC_ATCC_Sequences_id. The strain IDs (entry keys) are called "entities" in Terra.

NOTE: *no spaces or dashes are allowed. "entity:" and "_id" are required by Terra in the column name.*

5.1.2.2 **Read1.** The name of the read 1 fastq.gz file for the entry.

5.1.2.3 **Read2.** The name of the read 2 fastq.gz file for the entry.



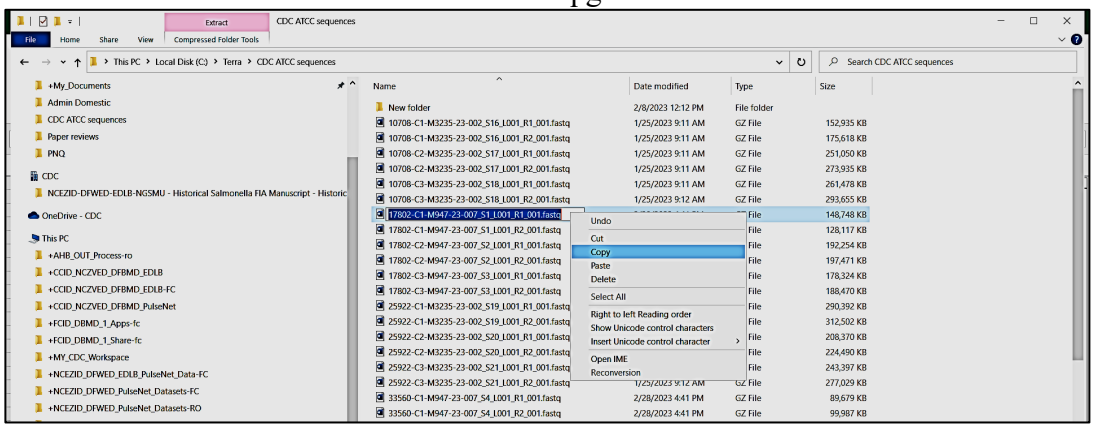
entity:CDC_ATCC_Sequences_id	read1	read2
ATCC-17802-C1_VP	17802-C1-M947-23-007_S1_L001_R1_001.fastq.gz	17802-C1-M947-23-007_S1_L001_R2_001.fastq.gz
ATCC-17802-C2_VP	17802-C2-M947-23-007_S2_L001_R1_001.fastq.gz	17802-C2-M947-23-007_S2_L001_R2_001.fastq.gz
ATCC-17802-C3_VP	17802-C3-M947-23-007_S3_L001_R1_001.fastq.gz	17802-C3-M947-23-007_S3_L001_R2_001.fastq.gz
ATCC-33560-C1_CJ	33560-C1-M947-23-007_S4_L001_R1_001.fastq.gz	33560-C1-M947-23-007_S4_L001_R2_001.fastq.gz
ATCC-33560-C2_CJ	33560-C2-M947-23-007_S5_L001_R1_001.fastq.gz	33560-C2-M947-23-007_S5_L001_R2_001.fastq.gz
ATCC-33560-C3_CJ	33560-C3-M947-23-007_S6_L001_R1_001.fastq.gz	33560-C3-M947-23-007_S6_L001_R2_001.fastq.gz
ATCC-51812-C1_SE	51812-C1-B-M947-23-007_S10_L001_R1_001.fastq.gz	51812-C1-B-M947-23-007_S10_L001_R2_001.fastq.gz
ATCC-51812-C2_SE	51812-C2-B-M947-23-007_S11_L001_R1_001.fastq.gz	51812-C2-B-M947-23-007_S11_L001_R2_001.fastq.gz
ATCC-51812-C3_SE	51812-C3-B-M947-23-007_S12_L001_R1_001.fastq.gz	51812-C3-B-M947-23-007_S12_L001_R2_001.fastq.gz

NOTE: *additional metadata can be added at this point or at later date. Refer to [appendix PNID01-6](#) for guidance on metadata.*

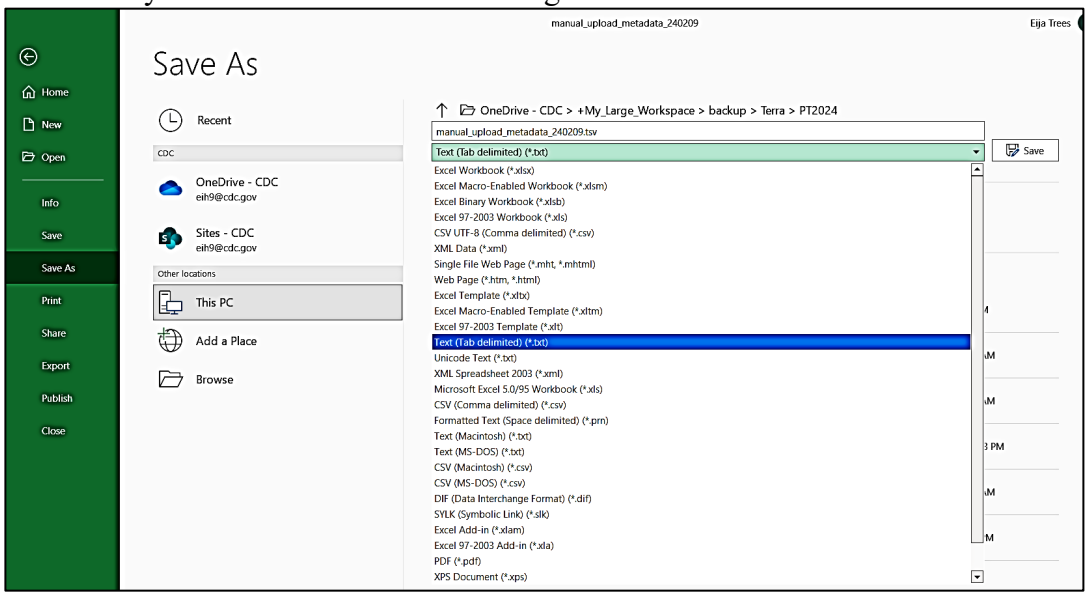
5.1.3 Type in the strain IDs in the "entity" column the way you want them to appear in Terra.

NOTE: *the strain ID does not need to match any part of the fastq.gz file name.*

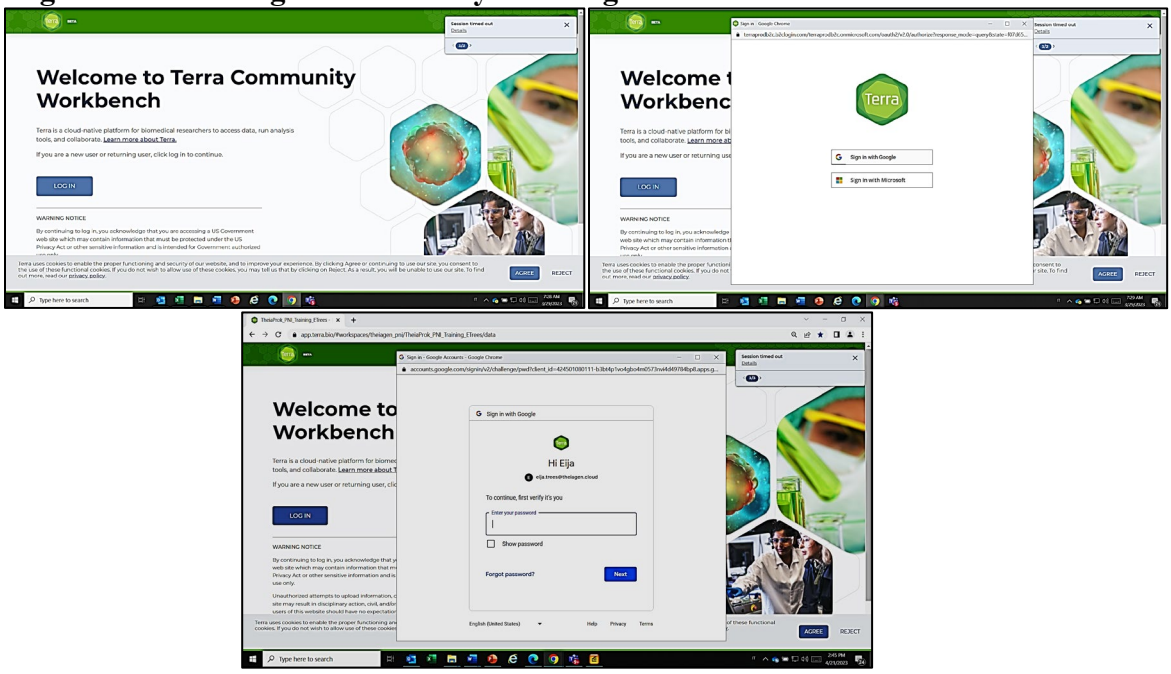
5.1.4 Copy and paste the fastq.gz file names for each strain in the columns "read1" and "read2". Make sure the file names end "fastq.gz".



5.1.5 Save the file in the **tsv format**: choose "Save As" and "Text (Tab delimited) (*.txt)". Make sure your file name has "**.tsv**" ending.

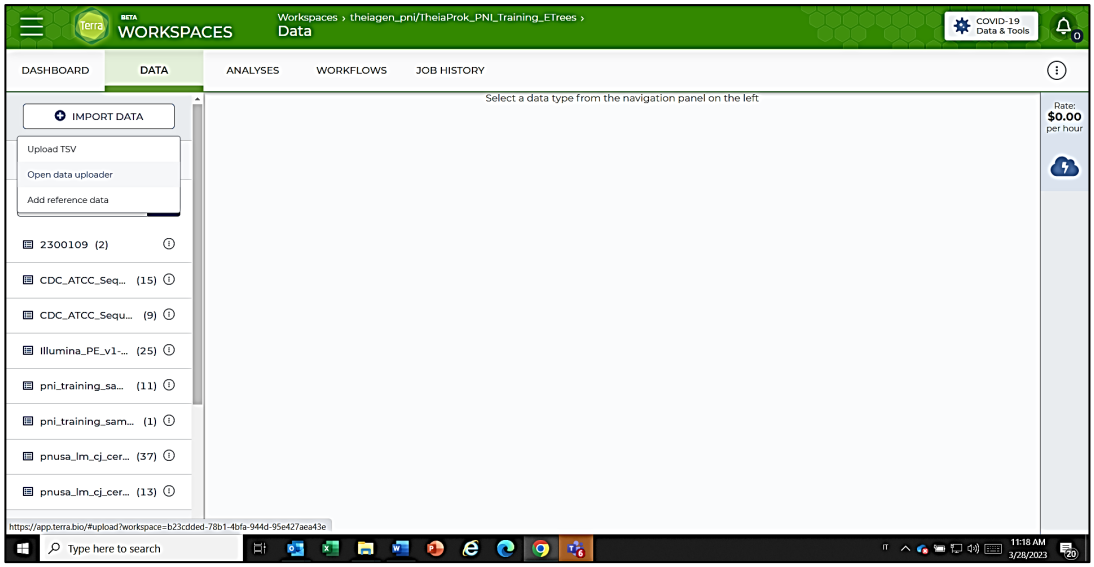


5.2 Log into Terra using Chrome and your Google account

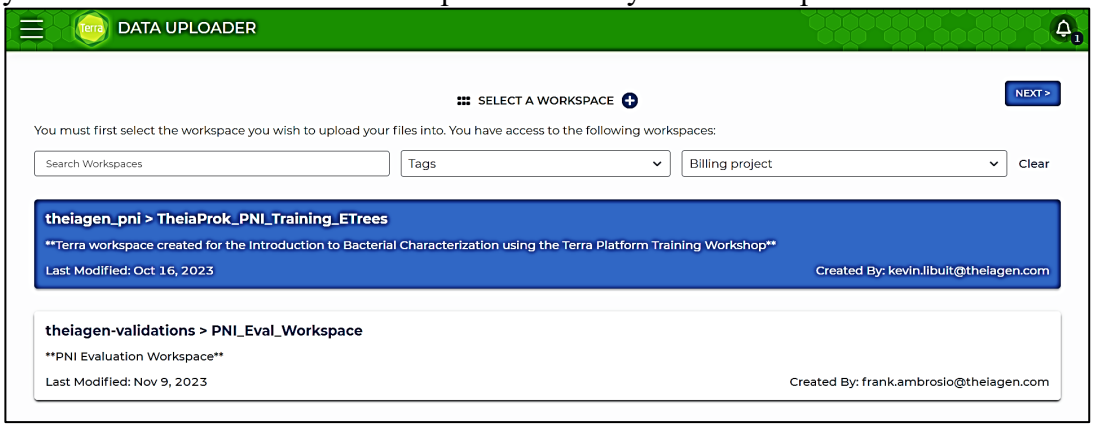


5.3 Upload sequence files and metadata into Terra

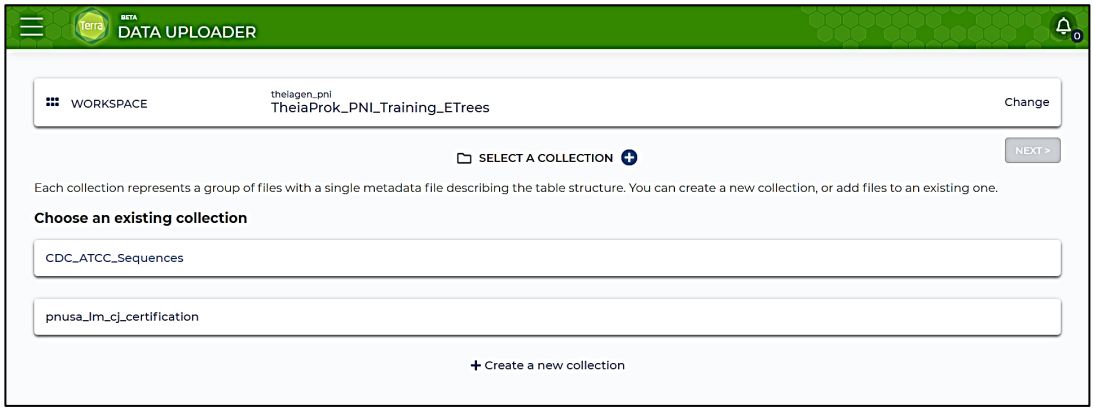
5.3.1 Under the "Terra Workspaces", select the "Data" tab, click "Import data" and select "Open data uploader" from the drop-down menu.



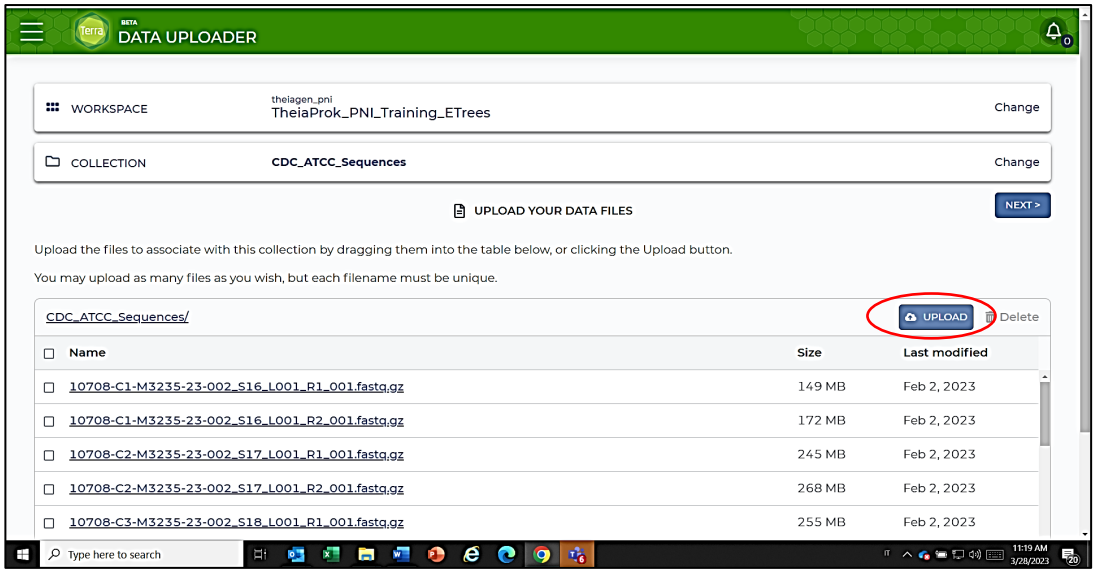
5.3.2 “Data uploader” screen will open. If your account has access to multiple workspaces, you first need to choose the workspace to which you wish to upload the data.



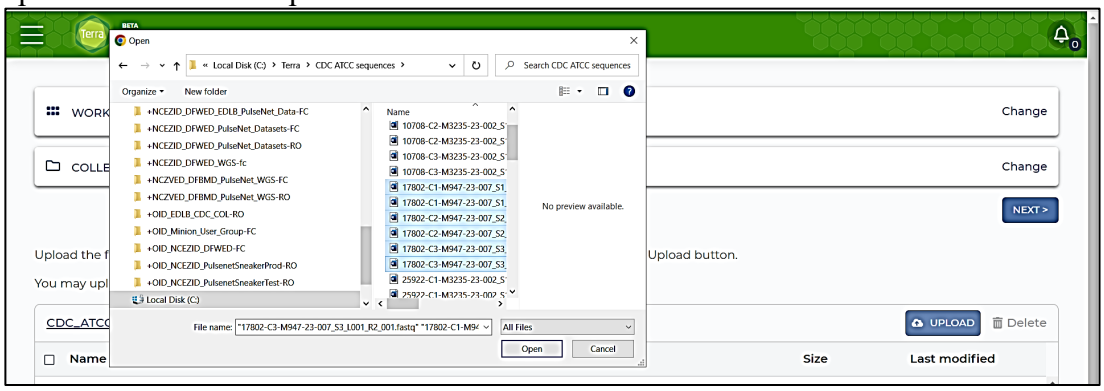
5.3.3 Either choose an existing collection by clicking on the collection name on the list or create a new collection.



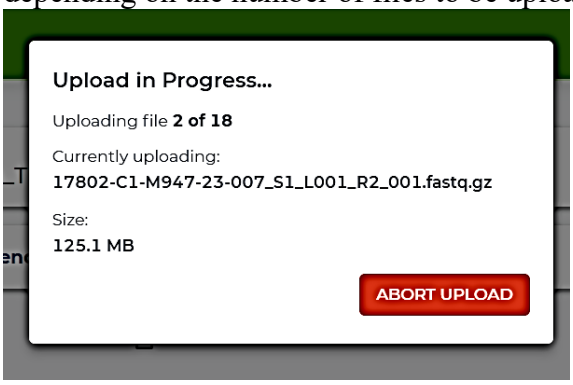
5.3.4 Under “Upload your data files”, click on “Upload”.



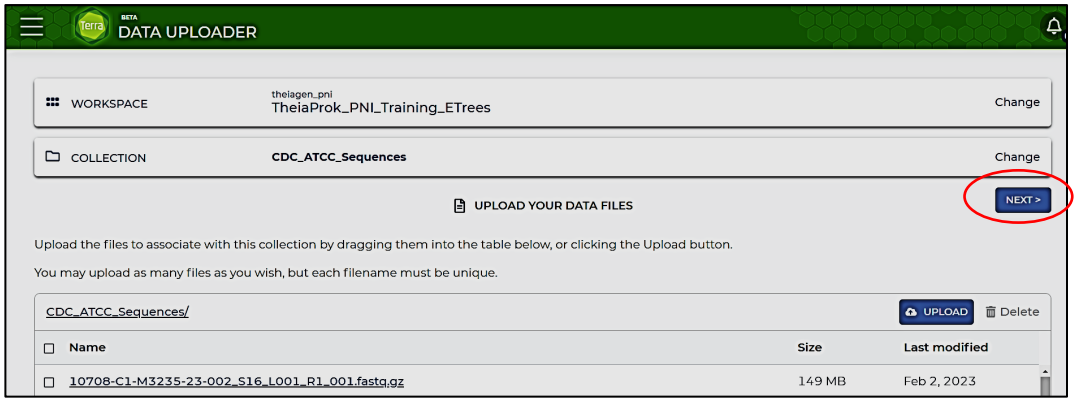
5.3.5 Navigate to the location where the FASTQ files are saved, select the files to be uploaded and click “Open”.



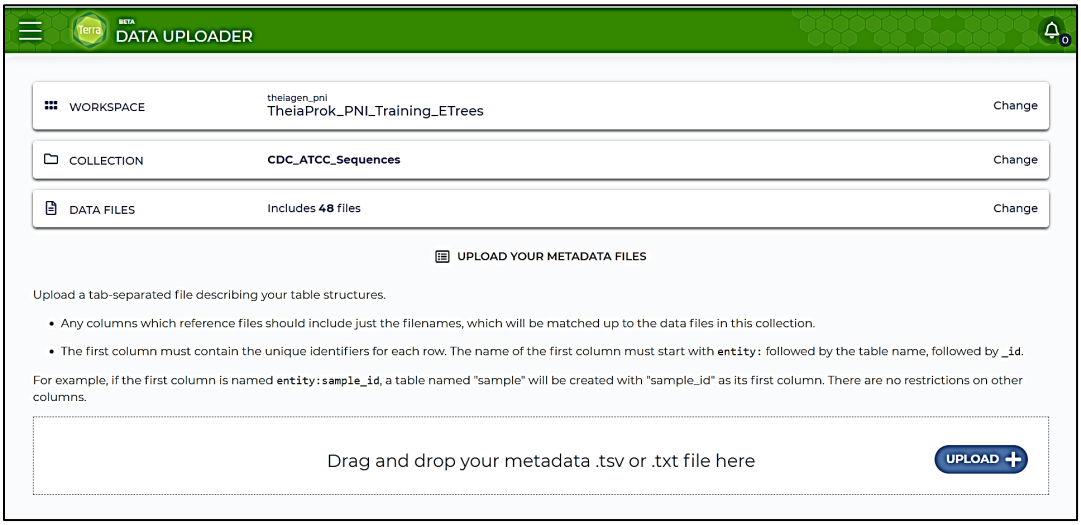
5.3.6 “Upload in progress” pop-up window will appear. The upload may take a few minutes depending on the number of files to be uploaded and your Internet bandwidth.



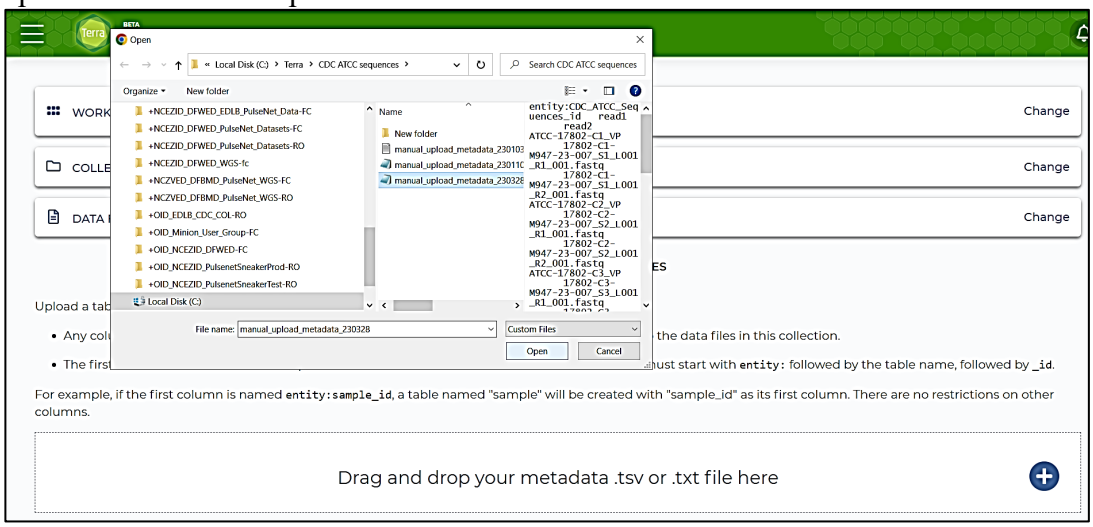
5.3.7 After the pop-up window disappears, click on “Next”.



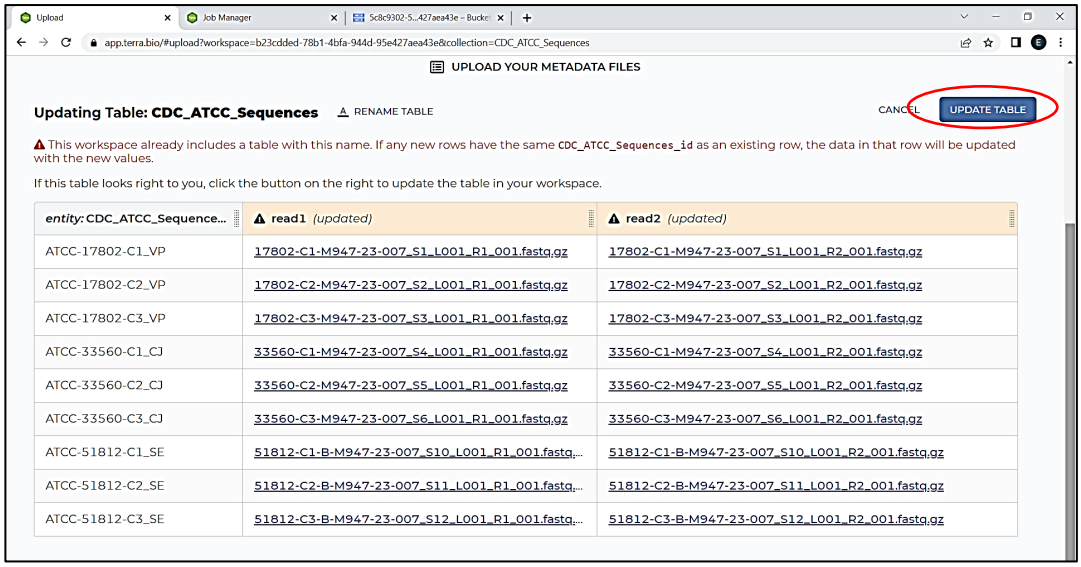
5.3.8 Under “Upload your metadata files”, click on “Upload”.



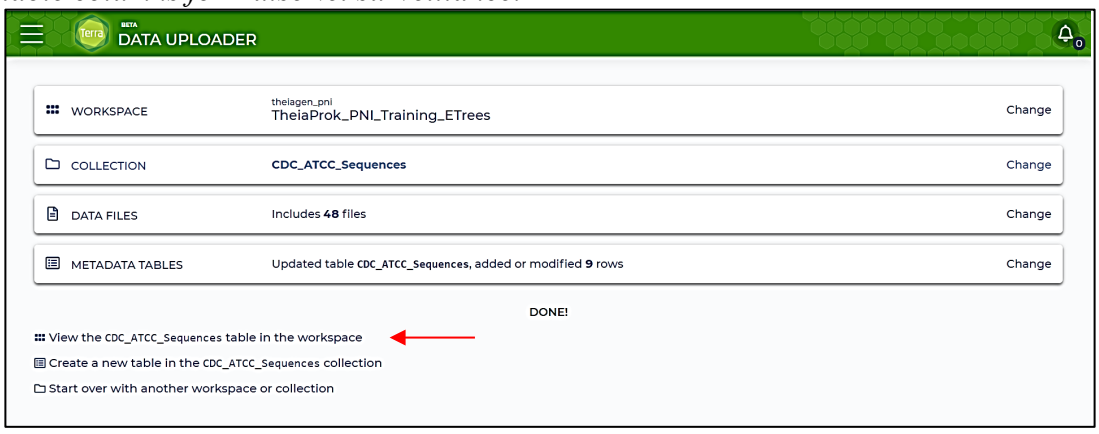
5.3.9 Navigate to the location where the metadata tsv file is saved, select the file to be uploaded and click “Open”.

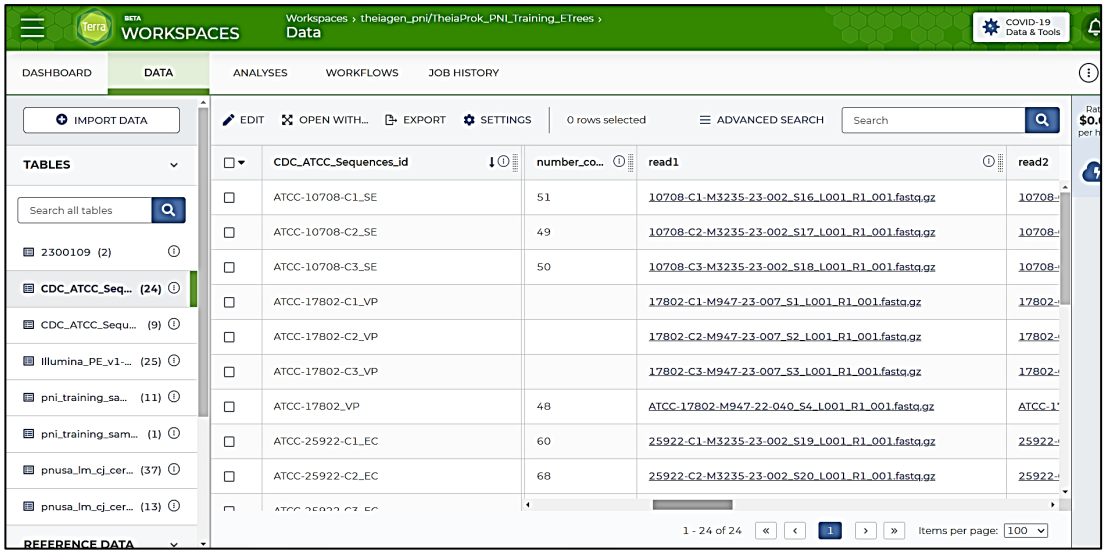


5.3.10 On the following screen, check that the fastq.gz files are linked to the correct entry keys and click “Update table”.



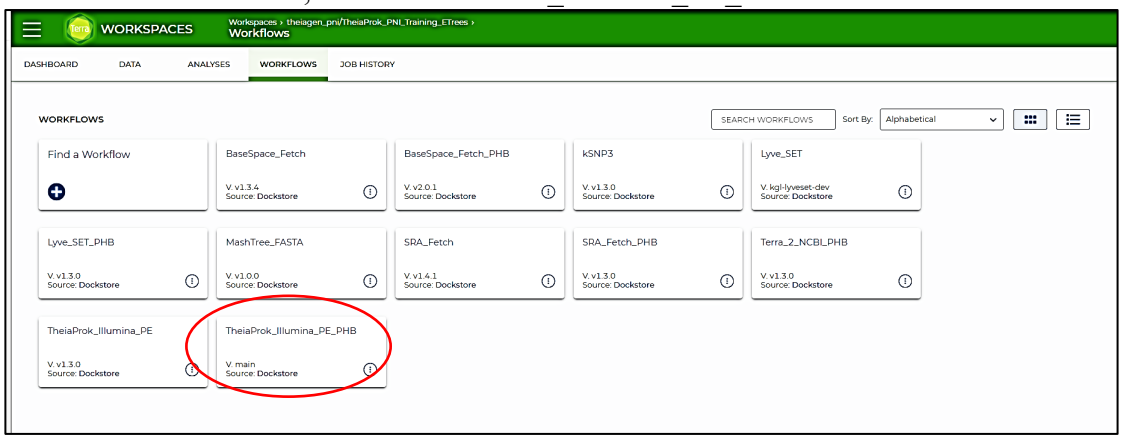
5.3.11 A "Done" message will appear on the "Data uploader" screen. You can view the updated data table by clicking on the link that appears on the screen.
NOTE: *The columns visible in the data table can be customized. For the ease of navigation, it is recommended to create separate views for QC metrics, genotyping results and metadata. Refer to appendices PNID01-3 (QC metrics), PNID01-5 (Genotyping) and PNID01-6 (Metadata) for guidance on how to customize the data table columns for PulseNet surveillance.*





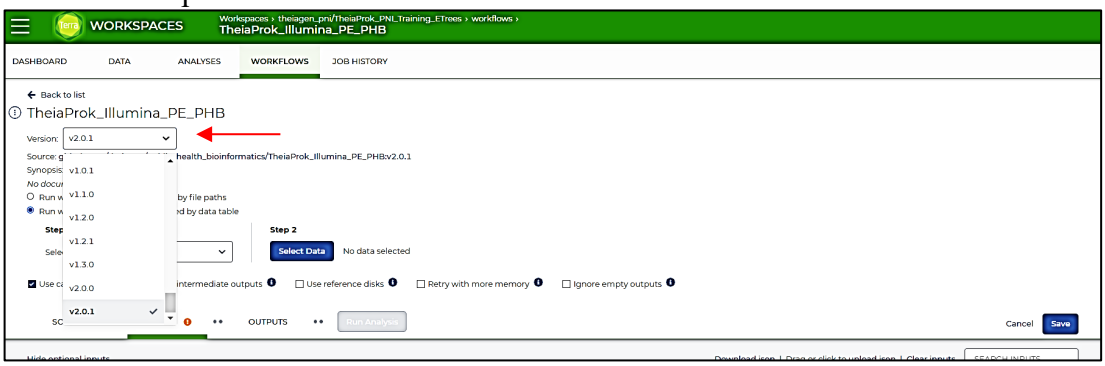
5.4 **Run the QC and genotyping workflow.** The TheiaProk workflow performs the QC of the raw sequence reads, de novo assembles the raw reads using Skesa and then performs the QC of the assembly and identification of the species. A variety of genotyping assays appropriate to the species are also available.

5.4.1 In the "Workflows" tab, select "TheiaProk_Illumina_PE_PHB".

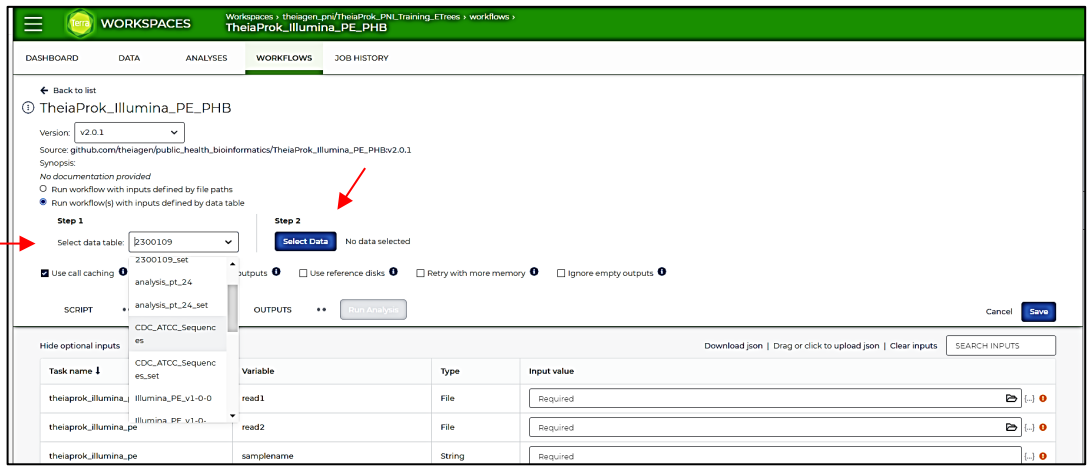


5.4.2 In the "TheiaProk_Illumina_PE_PHB" screen:

5.4.2.1 Select the latest version of the TheiaProk_Illumina_PE_PHB workflow from the "Version" drop-down menu.

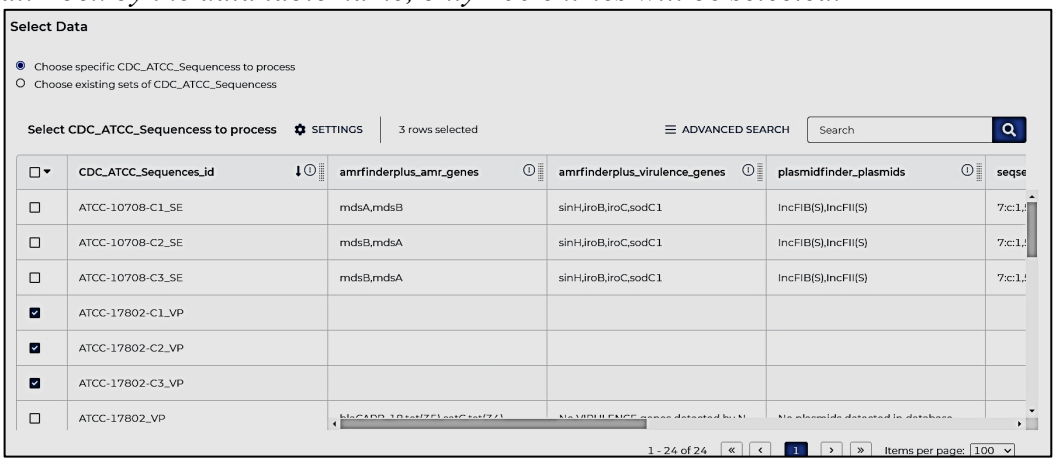


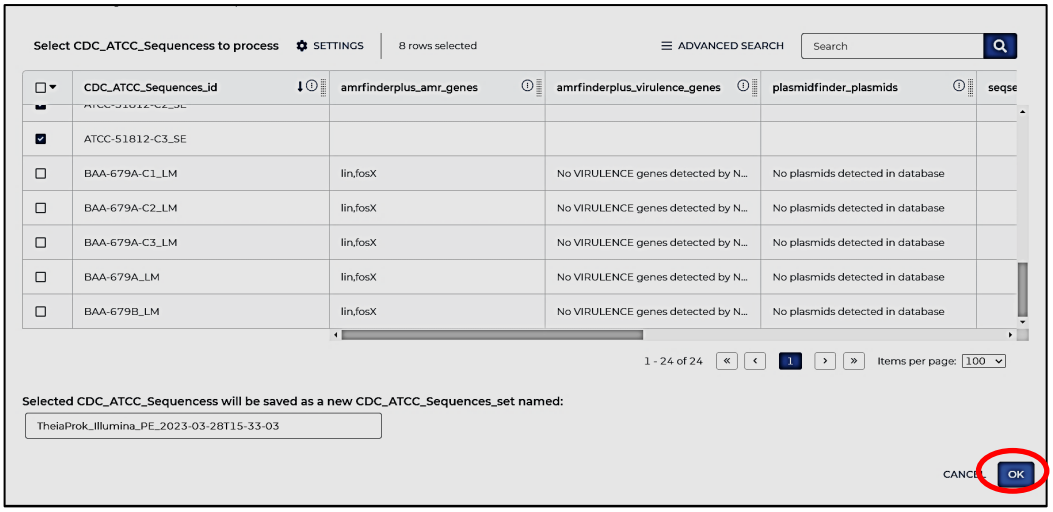
5.4.2.2 Under "Step 1", select the "Root type", i.e., the data table in which the samples are located, e.g., "CDC_ATCC_Sequences".
NOTE: *For most data tables, there are two options: the **main** data table containing the individual sample entries and the **set** data table containing sample sets used for phylogenetic analyses, NCBI uploads, etc (e.g., CDC_ATCC_Sequences_set). Make sure to select the **main** data table.*



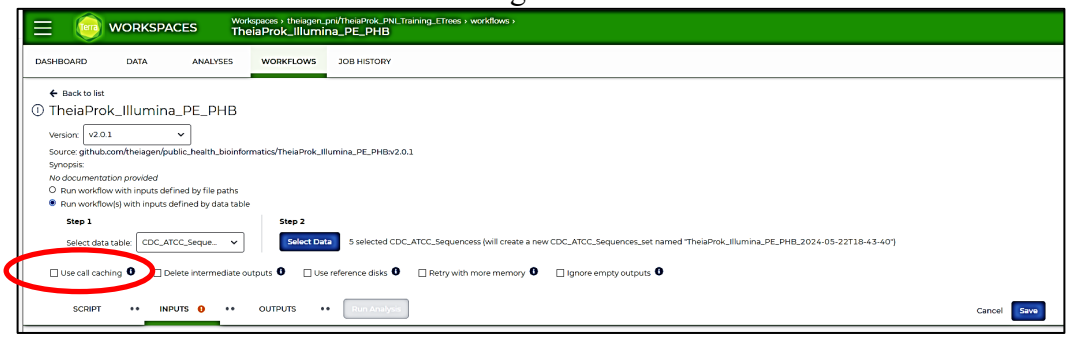
5.4.2.3 Under "Step 2", click "Select data" (screenshot above). This will take you to the data table specified in Step 1.

5.4.2.4 Select the strains to be analyzed and scroll all the way to the bottom to click "OK".
NOTE: *if the data table contains more than 100 entries and you check the "Select all" box by the data table name, only 100 entries will be selected.*





5.4.2.5 De-select the box for “Use call catching”.



5.4.2.6 In the “**Inputs**” tab, specify the following input values in the “Attribute” column (scroll down the list):

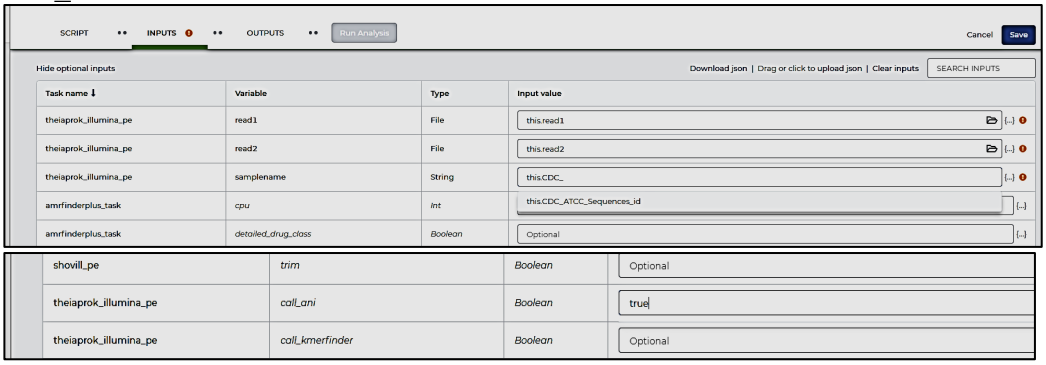
NOTE: *When you fill in the Attribute column, clicking inside the cell will bring up a drop-down menu of attributes that you can select to avoid typos (screenshot below).*

5.4.2.6.1 Read1: “This.read1”.

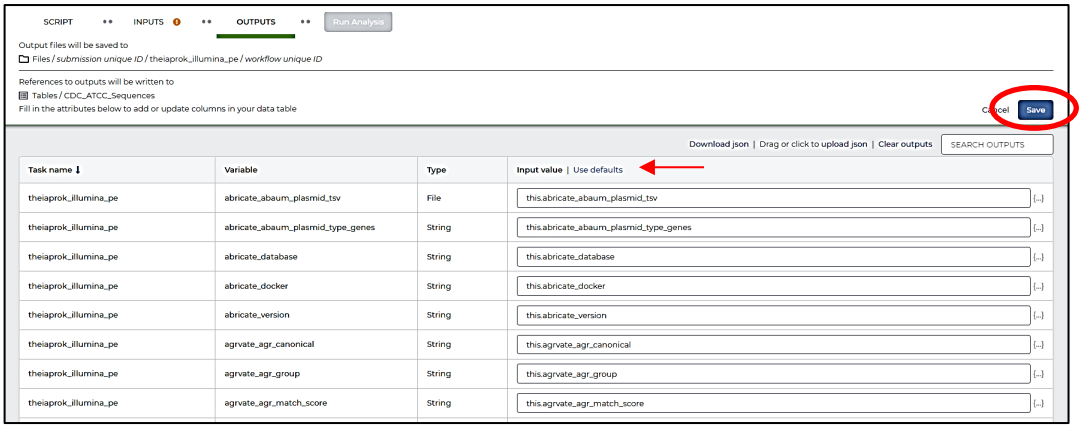
5.4.2.6.2 Read2: “This.read2”.

5.4.2.6.3 Samplename: “**This**.*data table name*_**id**”, e.g., This.CDC_ATCC_Sequences_id.

5.4.2.6.4 Call_ani: “True”.



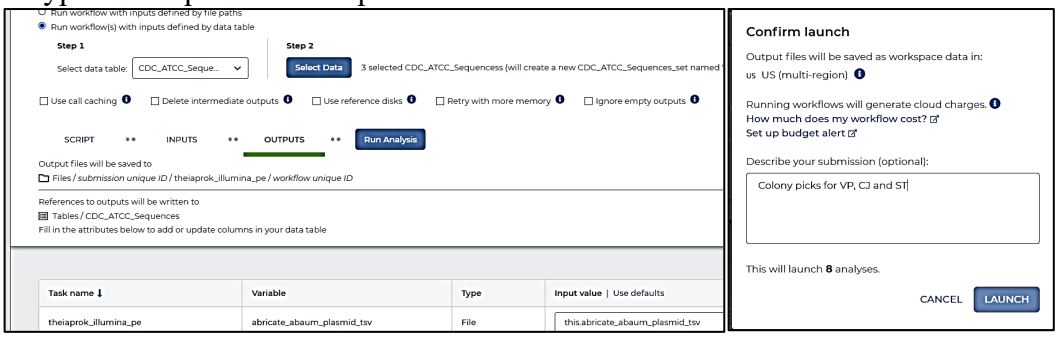
5.4.2.7 In the “**Outputs**” tab, click “Use defaults”.



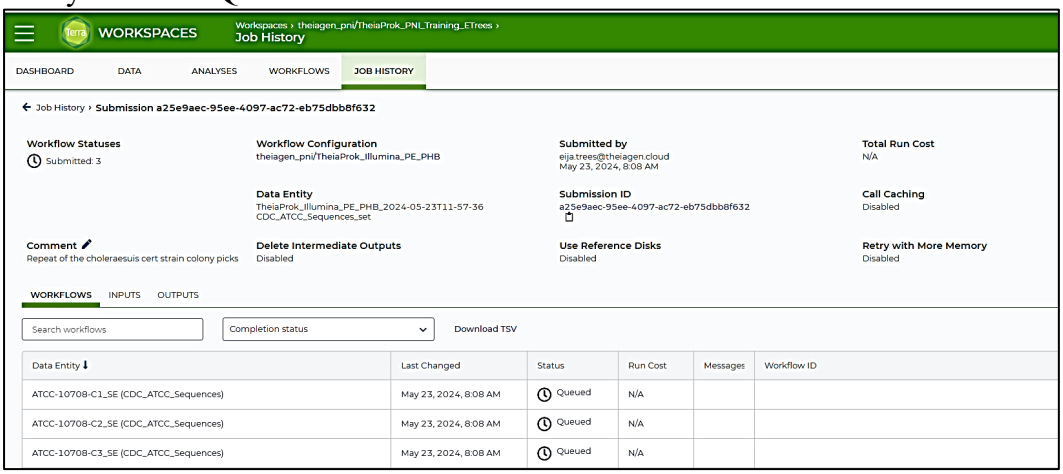
5.4.2.8 Click “Save” (screenshot above).

NOTE: *the “Save” button is only visible if you have changed the inputs from the previous submission.*

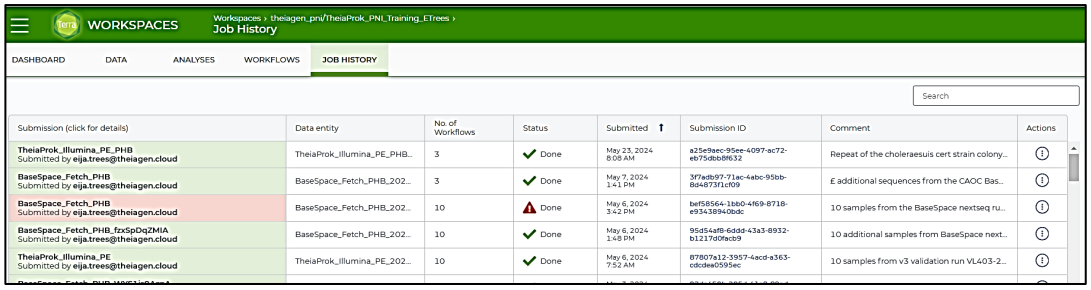
5.4.2.9 Click “Run analysis”. A “Confirm launch” pop-up window appears that allows you to type in an optional description. Click “Launch”.



5.4.3 A “Workflow statuses” screen will appear where your submitted jobs should be initially listed as “Queued”.

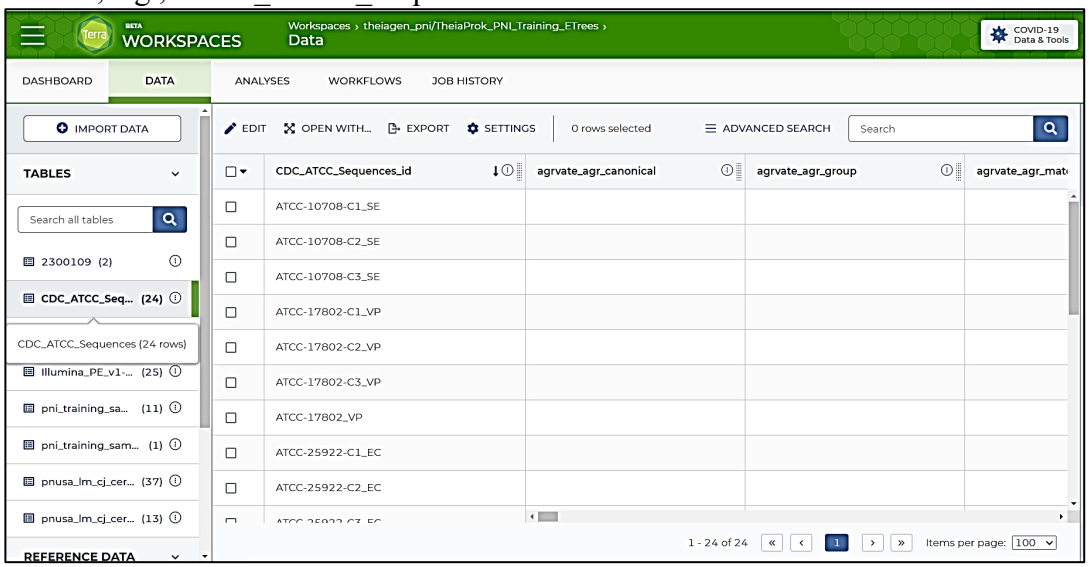


5.4.4 Go to the “Job history” tab to check the status of your job submission. A successfully finished job is indicated by a green check mark.

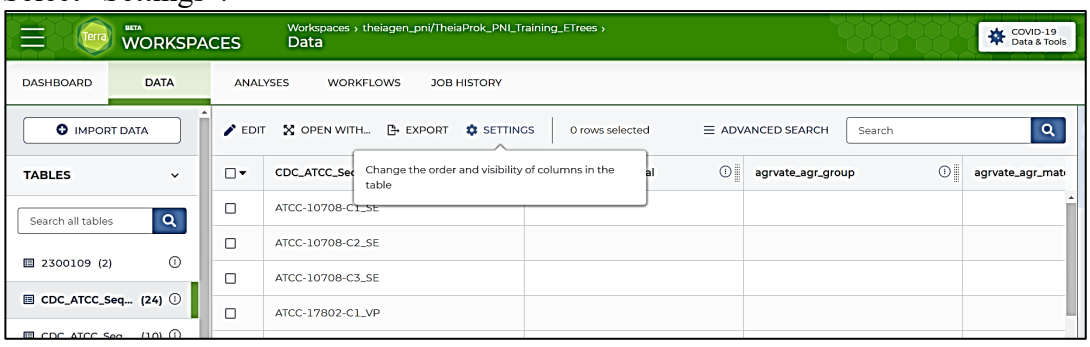


5.5 **Evaluate the QC metrics for the sequences:** QC metrics can be viewed either directly on the data table (5.5.1-5.5.3) or they can be exported to Excel for the selected entries (5.5.4).

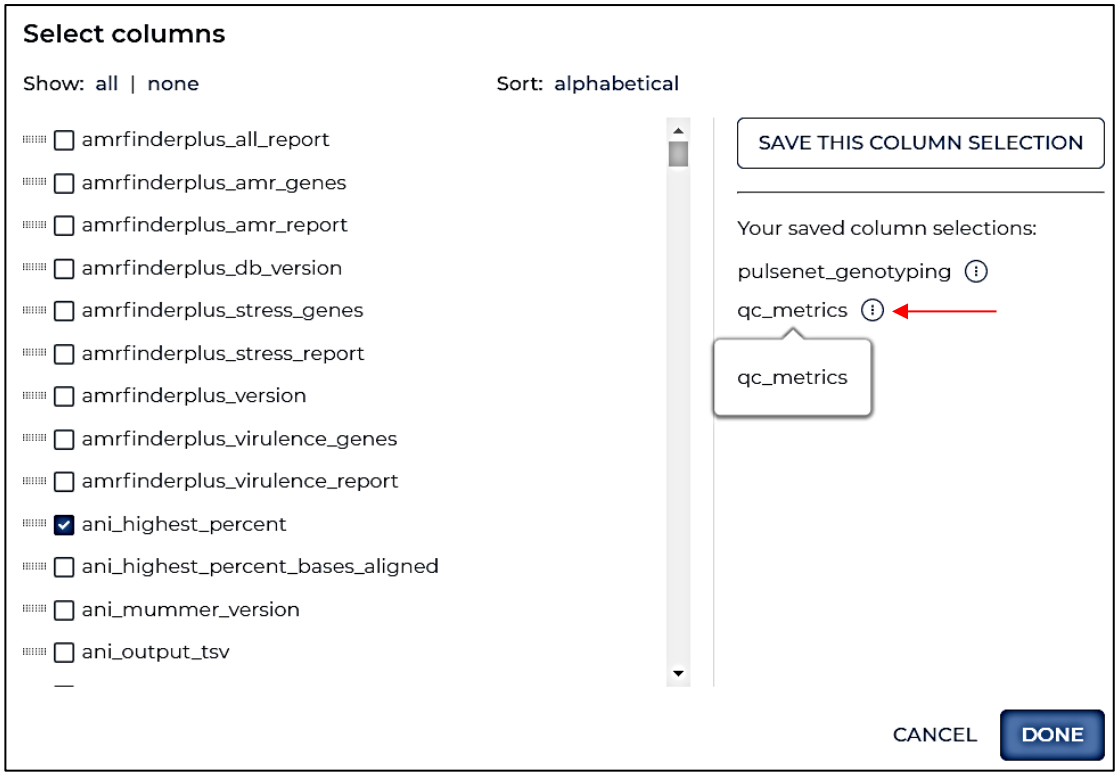
5.5.1 Under the "Terra Workspaces", select the "Data" tab, then select the data table of interest, e.g., "CDC_ATCC_Sequences".

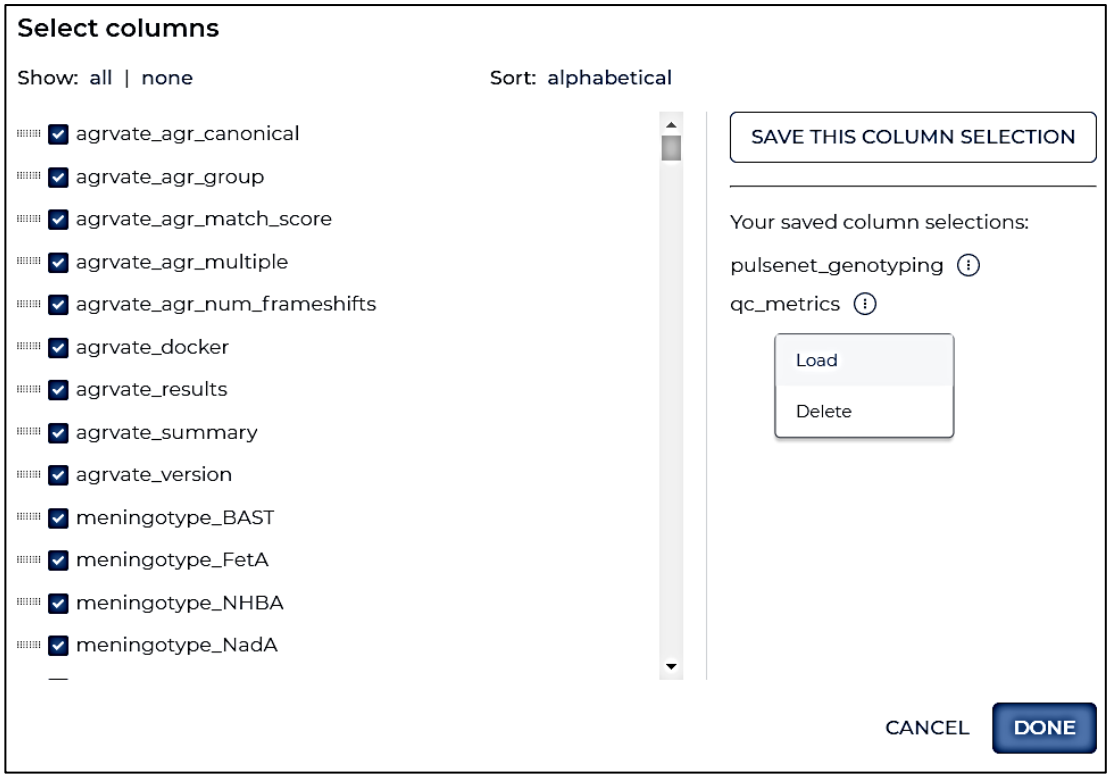


5.5.2 Select "Settings".



5.5.3 On the "Select columns" screen under "Your saved column selections", click on the circle with 3 dots next to the "qc_metrics" and from the drop-down menu select "Load" and then click "Done". This will load the appropriate PulseNet QC metrics into the data table. Refer to the Appendix PNID01-3 for the QC metrics that are supposed to appear on the table and for the instructions on how to add or delete any of the columns (QC metrics) in the QC metrics table.

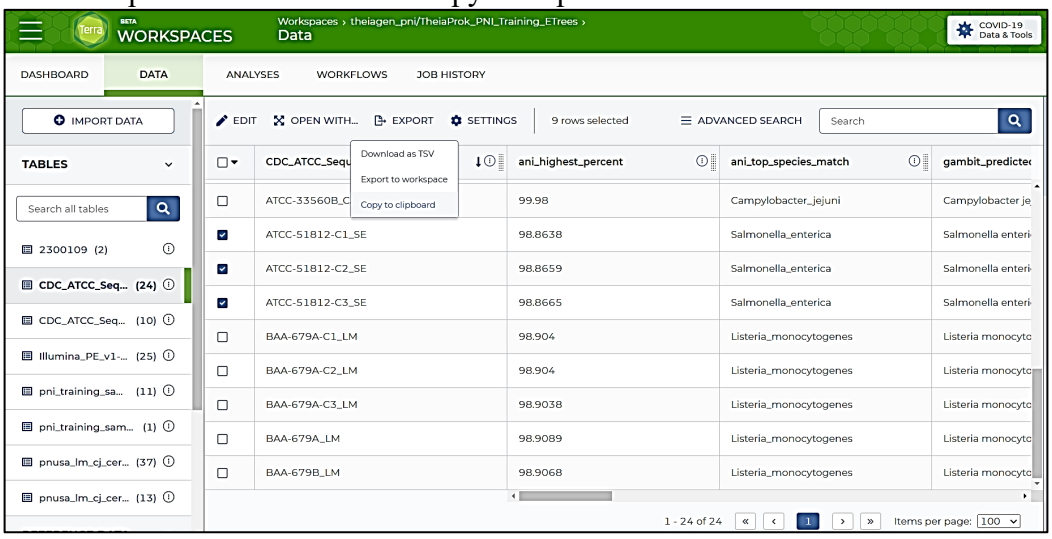




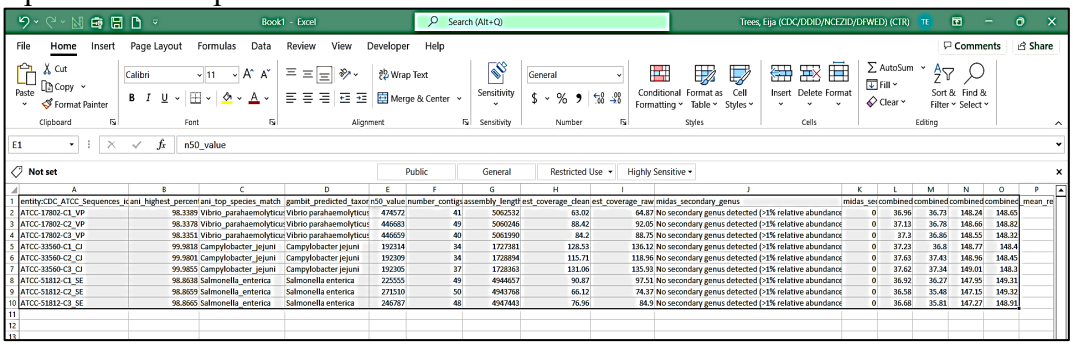
5.5.4 Exporting QC metrics to Excel:

5.5.4.1 Select the desired entries.

5.5.4.2 Click “Export” and then select “Copy to clipboard”.

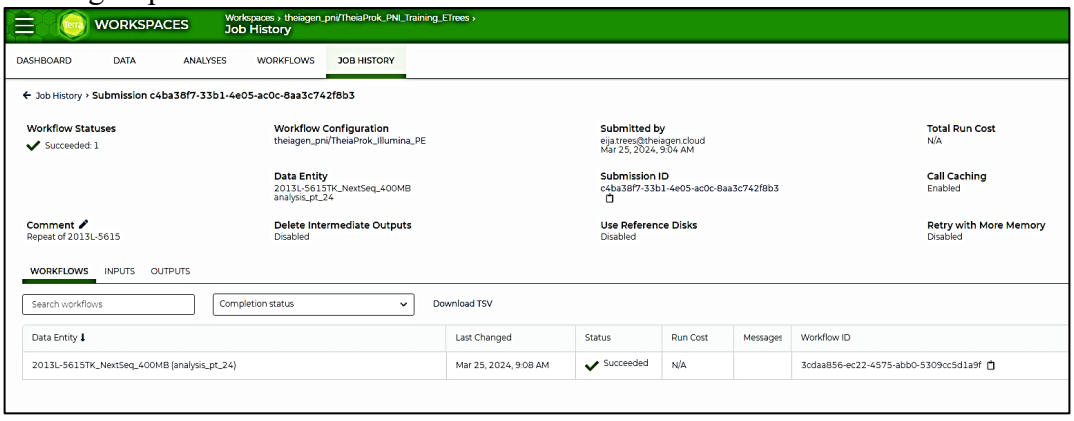


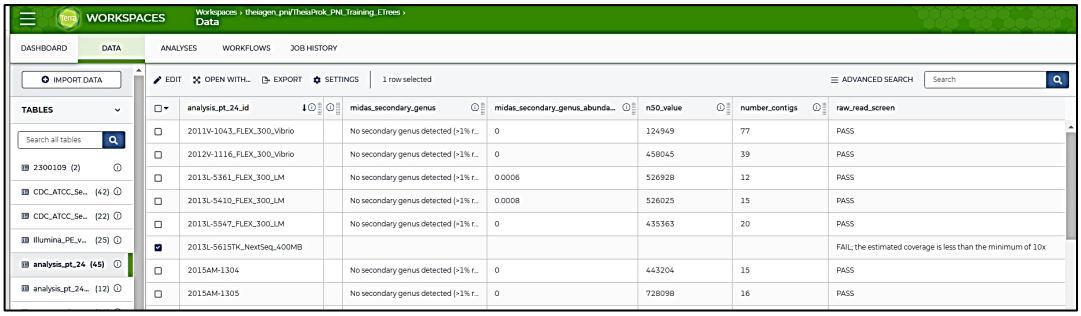
5.5.4.3 Open Excel and paste the data to a worksheet.



5.5.5 Refer to Appendix PNID01-4a for PulseNet critical quality metrics for acceptable Illumina sequences.

5.5.6 TheiaProk employs a read pre-screening step in which it will stop analysis for samples that fall below certain quality thresholds. In this case, the “Job History” tab indicates that the sample was successful, but there are no results in the “Data” tab. The “raw_read_screen” column in qc metrics should indicate the reason for the sequence analysis failure. Refer to Appendix PNID01-4b for thresholds applied in this read pre-screening step.



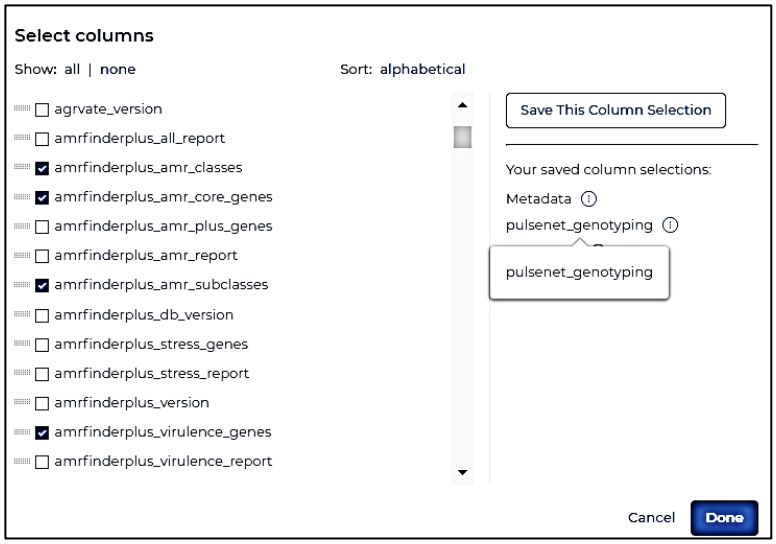


5.6 **View the genotyping results for the sequences:** Genotyping results can be viewed either directly on the data table (5.6.1-5.6.3) or they can be exported to Excel for the selected entries (5.6.4).

5.6.1 Under the "Terra Workspaces", select the "Data" tab, then select the data table of interest, e.g., "CDC_ATCC_Sequences".

5.6.2 In the "Data" tab, select "Settings".

5.6.3 On the "Select columns" screen under "Your saved column selections", click on the circle with 3 dots next to the "pulsenet_genotyping" and from the drop-down menu select "Load" and then click "Done". This will load the genotyping assays appropriate for the PulseNet surveillance into the data table. Refer to the Appendix PNID01-5 for the genotyping assays that are supposed to appear on the table and for the instructions on how to add or delete any of the columns (genotyping assays) in the genotyping results table.



5.6.4 Export the results to Excel for the selected entries; follow the procedure in the step 5.5.4.

5.7 **Upload sequences to NCBI**

NOTE: *Contact Theiagen Genomics (support@theiagen.com) for guidance before starting and to configure your workspace for the NCBI uploads. The configuration process is described in: https://theiagen.notion.site/Terra_2_NCBI-61abcedc066646b3b258f70b561e9f62.*

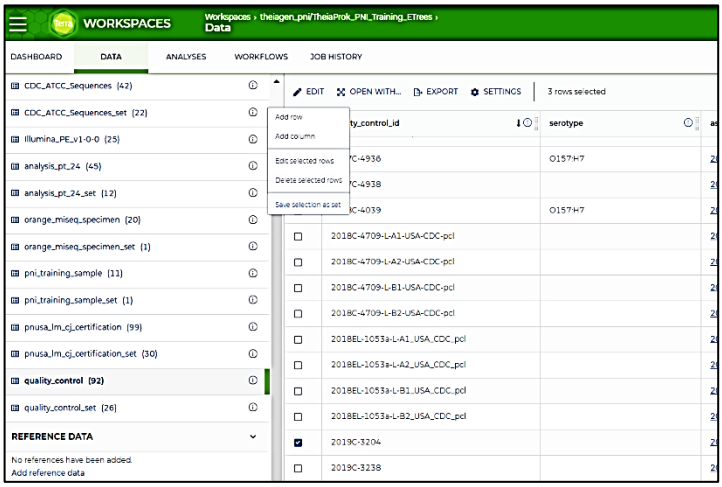
5.7.1 Upload the metadata required for the NCBI submission: refer to the appendix PNID01-6 for the correct metadata formatting and upload.

5.7.2 Create a **set** of the samples to be uploaded to NCBI:

5.7.2.1 Under the "Terra Workspaces", select the "Data" tab, then select the data table of interest, e.g., "quality_control".

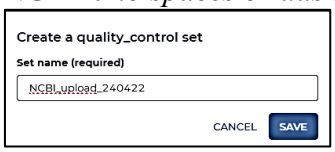
5.7.2.2 In the "Data" tab, select the sequences to be included in the NCBI upload.

5.7.2.3 From the "Edit" drop-down menu, select "Save selection as set".

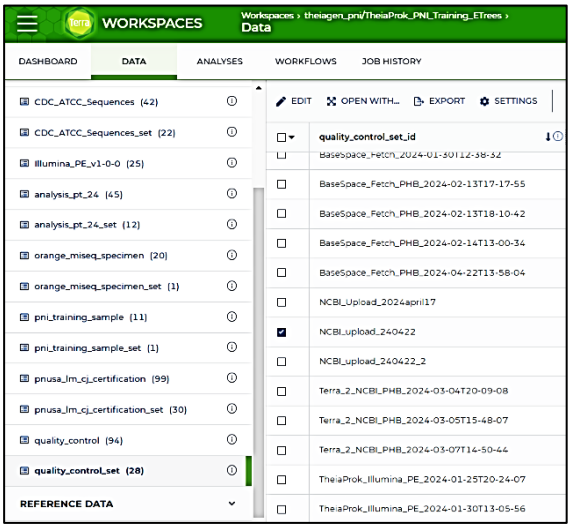


5.7.2.4 In the appearing pop-up window, name the set, e.g., "NCBI_upload_240422" and click "Save".

NOTE: *no spaces or dashes are allowed.*

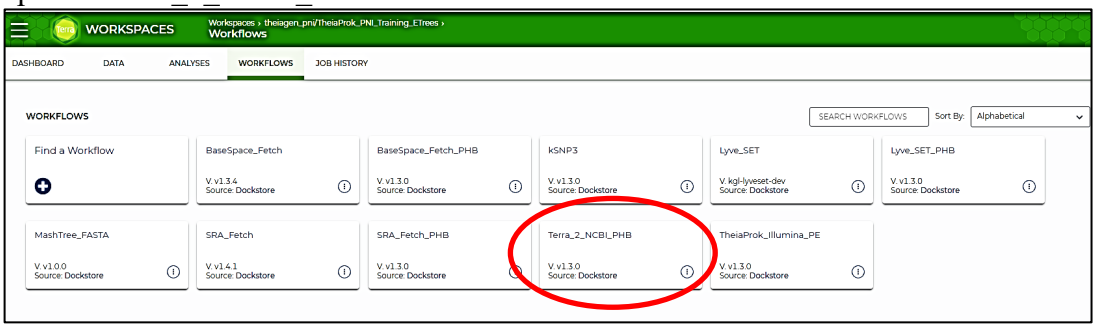


5.7.2.5 The newly created set should now appear in the "set" data table, e.g., "quality_control_set".

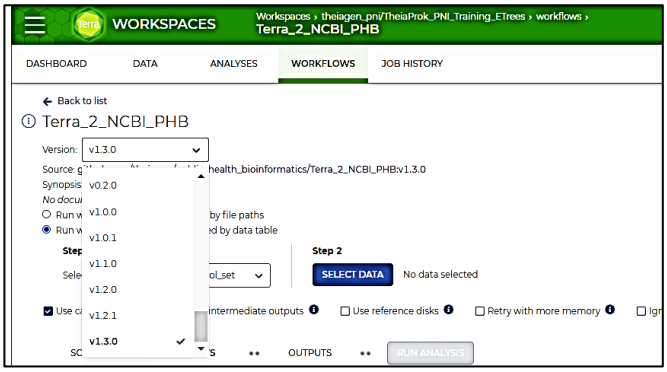


5.7.3 Set up parameters for the NCBI upload workflow:

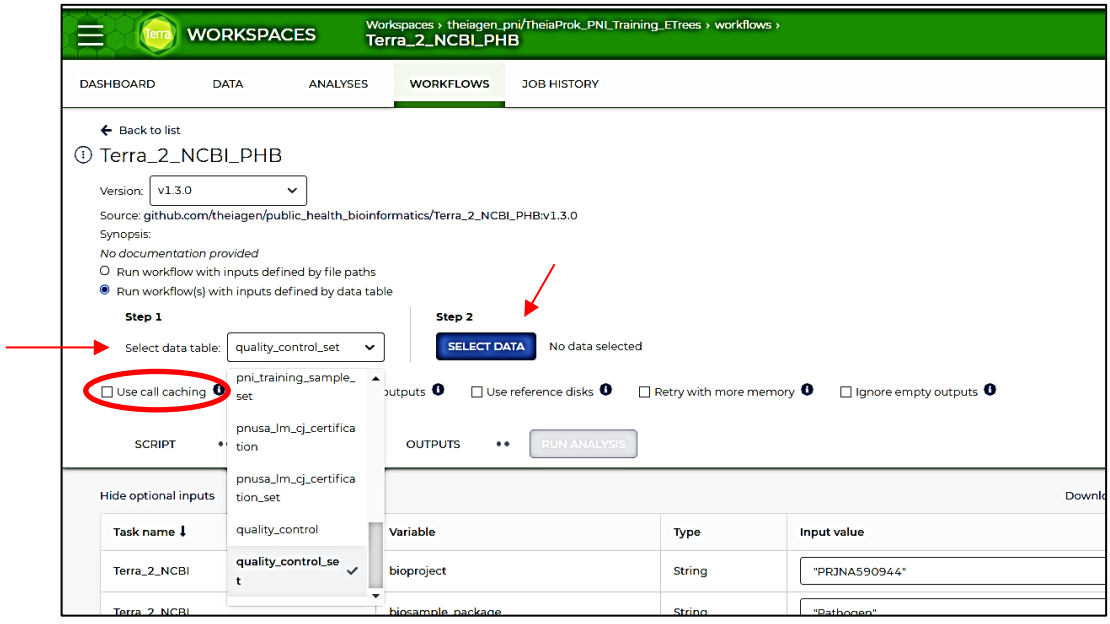
5.7.3.1 In the "Workflows" tab, select the "Terra_2_NCBI_PHB" workflow. This will bring up the "Terra_2_NCBI_PHB" screen.



5.7.3.2 From the "Version" drop-down menu, select the latest version of the "Terra_2_NCBI_PHB".

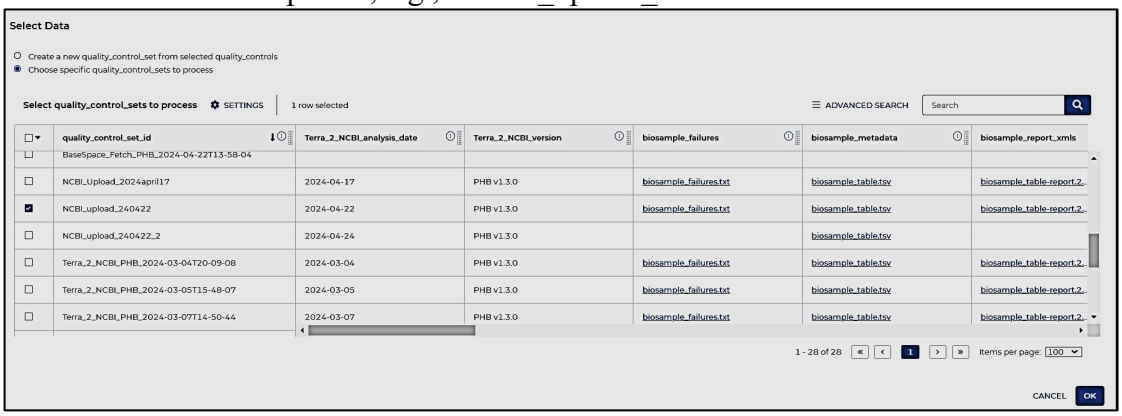


5.7.3.3 Under Step 1, from the "Select root entity type" drop-down menu, select the **set** data table where the sample set created in step 5.7.2. is located, e.g., "quality_control_set". Also de-select the box for "Use call catching".



5.7.3.4 Under Step 2, click "Select data" (screenshot above). This will take you to the **set** data table selected in the previous step.

5.7.3.5 Select the desired sample set, e.g., "NCBI_upload_240422" and click "OK".



5.7.3.6 In the "**Inputs**" tab, the following "Input Values" need to be filled out for the "Variables" listed below:

5.7.3.6.1 bioproject: enter the number for the NCBI BioProject to which you wish to upload in quotation marks, e.g., "PRJNA590944".

5.7.3.6.2 biosample_package in quotation marks: "Pathogen". This is the metadata template/package you use for the metadata upload for sequences belonging to PulseNet surveillance.

5.7.3.6.3 ncbi_config_js: enter the name of the NCBI configuration file created for your workspace, e.g., workspace.ncbi_config_etrees.

5.7.3.6.4 project_name in quotation marks: "theiagen_pni".

5.7.3.6.5 sample_names: enter your data table name in the format:
this.data_table_name**s**.data_table_name_**id**,
e.g., this.quality_controls.quality_control_id.
NOTE: *the double name format is REQUIRED.*

5.7.3.6.6 sra_transfer_gcp_bucket in quotation marks: "gs://theiagen_sra_transfer". This is the temporary public Google storage location for your sequences that NCBI can access.

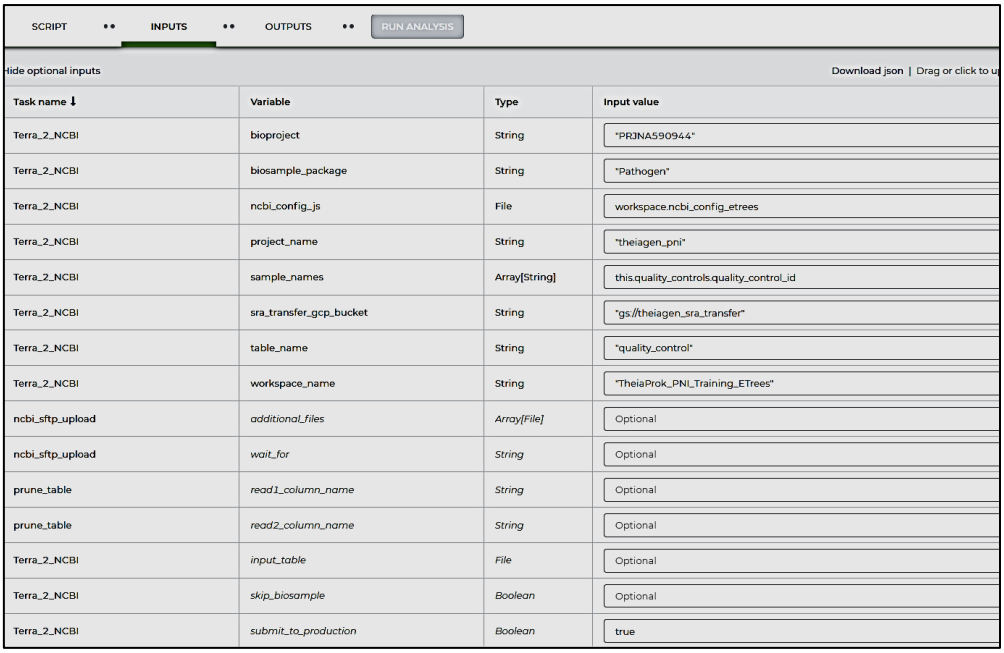
5.7.3.6.7 table_name: enter your data table name in quotation marks, e.g., "quality_control".

5.7.3.6.8 workspace_name: enter your workspace name in quotation marks e.g., "TheiaProk_PNI_Training_ETrees".

5.7.3.6.9 submit_to_production: true.

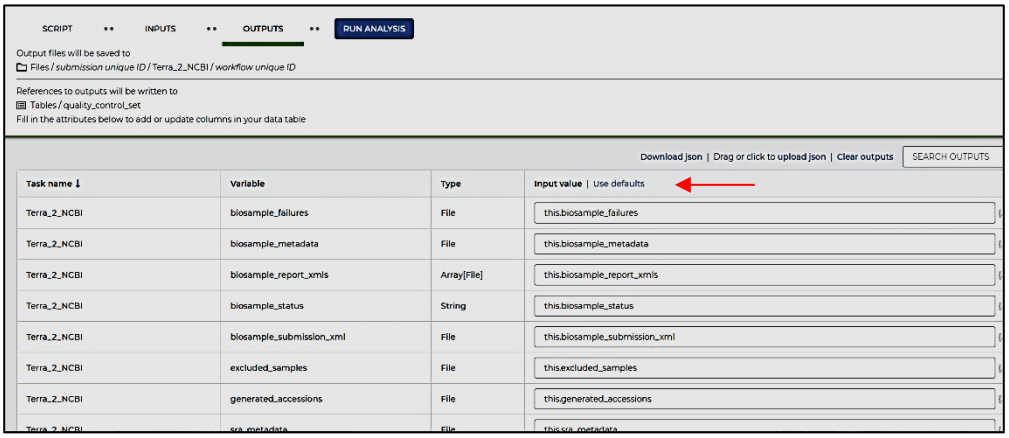
5.7.3.6.10 OPTIONAL: if you want to link an SRA submission to an **existing BioSample** ("biosample_accession" field already populated with the SAMN number in the "Data" tab):

5.7.3.6.10.1 skip_biosample: true.
NOTE: *for using this feature, the required metadata fields need to be populated, i.e., populating only the "biosample_accession" field is not enough and will not pass the Terra pre-upload checks.*



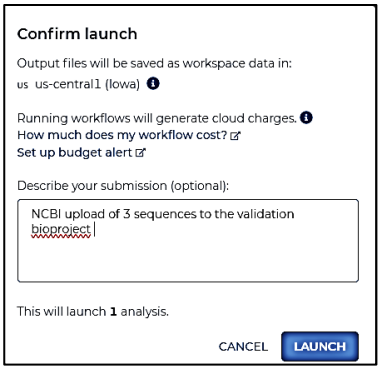
Task name	Variable	Type	Input value
Terra_2_NCBI	bioproject	String	"PRJNA590944"
Terra_2_NCBI	biosample_package	String	"Pathogen"
Terra_2_NCBI	ncbi_config_js	File	workspace.ncbi_config_etrees
Terra_2_NCBI	project_name	String	"theiagen_pni"
Terra_2_NCBI	sample_names	Array[String]	this.quality_controls.quality_control_id
Terra_2_NCBI	sra_transfer_gcp_bucket	String	"gs://theiagen_sra_transfer"
Terra_2_NCBI	table_name	String	"quality_control"
Terra_2_NCBI	workspace_name	String	"TheiaProk_PNI_Training_ETrees"
ncbi_sftp_upload	*additional_files*	*Array[File]*	Optional
ncbi_sftp_upload	*wait_for*	*String*	Optional
prune_table	*read1_column_name*	*String*	Optional
prune_table	*read2_column_name*	*String*	Optional
Terra_2_NCBI	*input_table*	*File*	Optional
Terra_2_NCBI	*skip_biosample*	*Boolean*	Optional
Terra_2_NCBI	*submit_to_production*	*Boolean*	true

5.7.3.7 In the "**Outputs**" tab, click "Use defaults" for the "Attributes" and then click "Save". **NOTE**: *the "Save" button is only visible if you have changed the inputs from the previous submission.*

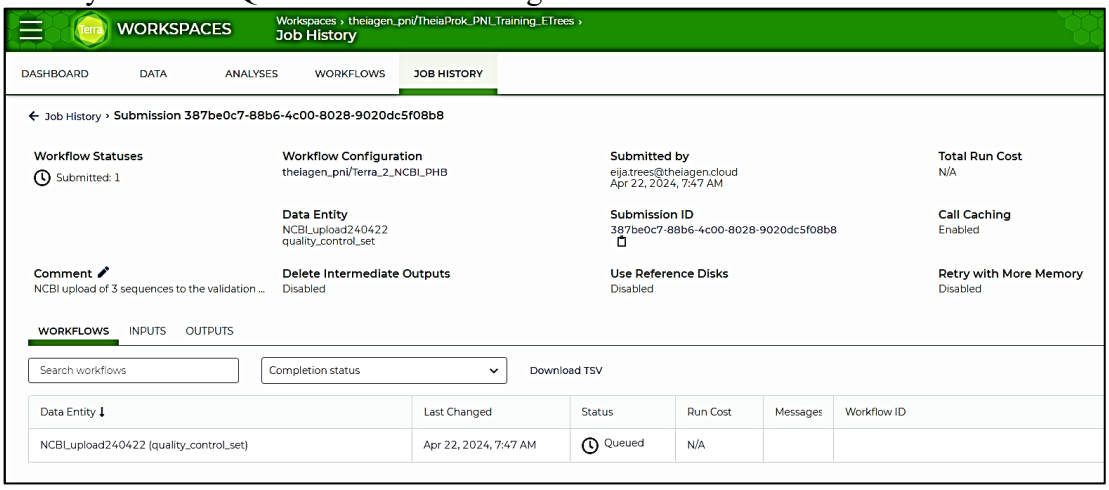


Task name	Variable	Type	Input value | Use defaults
Terra_2_NCBI	biosample_failures	File	this.biosample_failures
Terra_2_NCBI	biosample_metadata	File	this.biosample_metadata
Terra_2_NCBI	biosample_report_xmls	Array[File]	this.biosample_report_xmls
Terra_2_NCBI	biosample_status	String	this.biosample_status
Terra_2_NCBI	biosample_submission_xml	File	this.biosample_submission_xml
Terra_2_NCBI	excluded_samples	File	this.excluded_samples
Terra_2_NCBI	generated_accessions	File	this.generated_accessions
Terra_2_NCBI	sra_metadata	File	this.sra_metadata

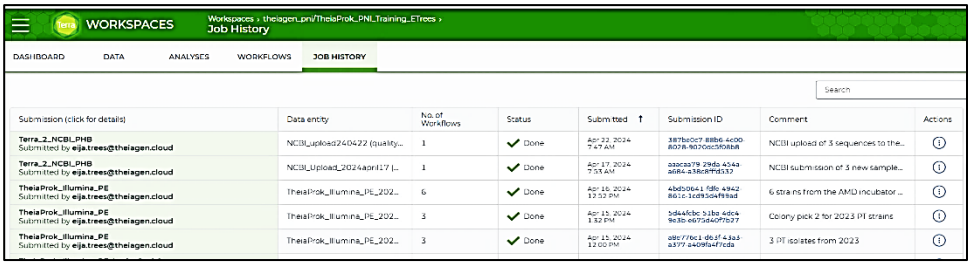
5.7.3.8 Click "Run analysis" (screenshot above). A "Confirm launch" pop-up window appears that allows you to type in an optional description. Click "Launch".



5.7.3.9 A "Workflow statuses" screen will appear where your submitted jobs should be initially listed as "Queued" or "Launching".



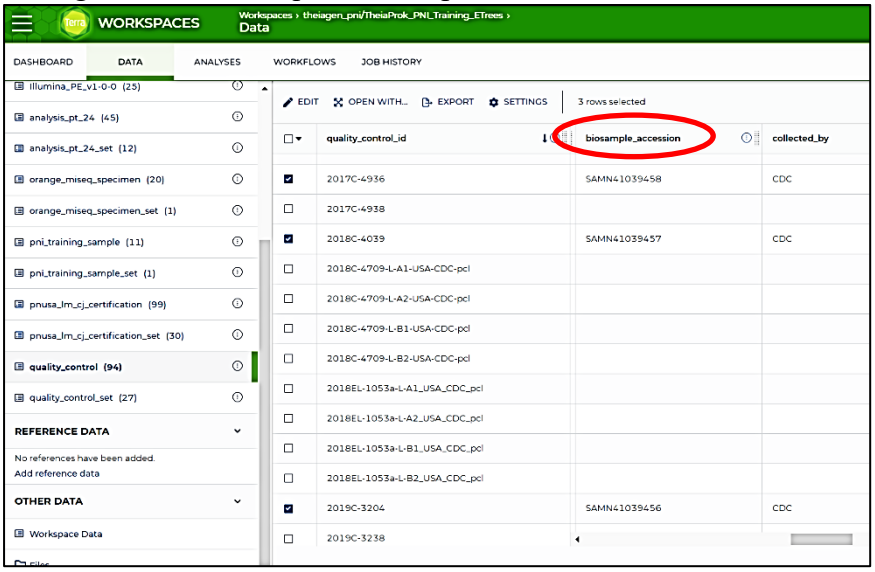
5.7.3.10 Go to the "Job history" tab to check the status of your job submission. A successfully finished job is indicated by a green check mark.



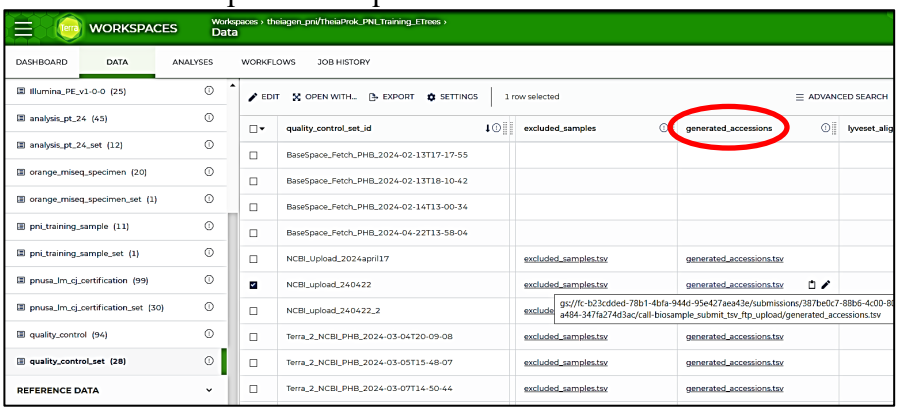
5.7.3.11 Tracking the NCBI accession numbers:

5.7.3.11.1 **BioSample accession** numbers (SAMN numbers)

5.7.3.11.1.1 In the "Data" tab, go to the data table in question, and you should see the "biosample_accession" field populated with the SAMN number that is assigned to each unique BioSample.

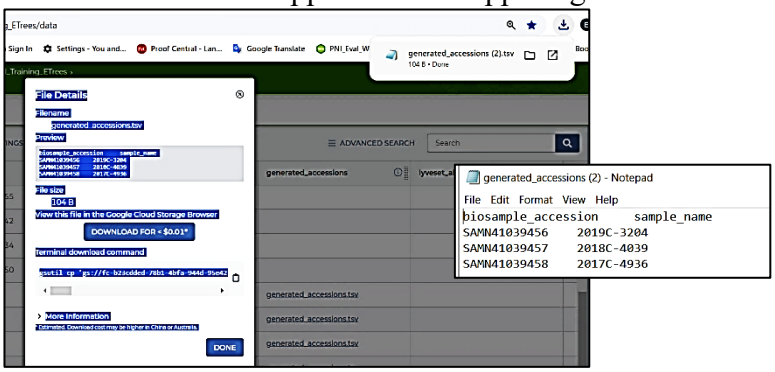


5.7.3.11.1.2 In the "Data" tab, go to the data table **set** in question, find the data set you created for the NCBI upload and then click on the "generated_accessions" link for the tsv file that lists the BioSample numbers for the uploaded sequence set. To download the file:

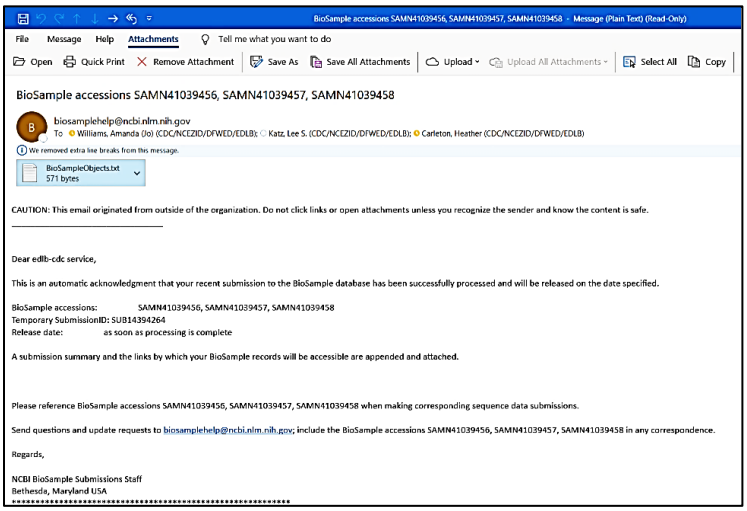


5.7.3.11.1.2.1 Click "Download >$0.01*" and then click "Done".

5.7.3.11.1.2.2 The downloaded file appears on the upper right corner.



5.7.3.11.1.3 The email address associated with the NCBI account used for the upload should receive a confirmation email listing the SAMN numbers in the body of the email and in the "BiosampleObjects" text file attached to the email.



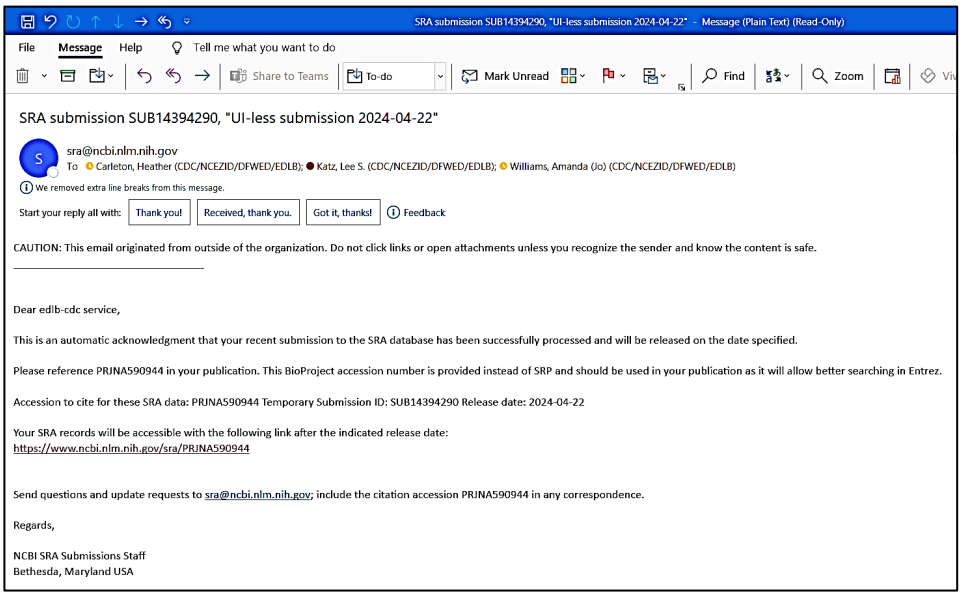
BioSampleObjects - Notepad

File Edit Format View Help

Accession	SPUID	Organism	Tax ID	Strain	BioProject
SAMN41039458	2017C-4936	Escherichia coli	562	2017C-4936	PRJNA590944
SAMN41039457	2018C-4039	Escherichia coli	562	2018C-4039	PRJNA590944
SAMN41039456	2019C-3204	Shigella sonnei 624	2019C-3204		PRJNA590944

5.7.3.11.2 **SRA accession** numbers (SRR numbers)

NOTE: *The email address associated with the NCBI account used for the upload should receive a second confirmation email stating that the submission to the SRA database has been successfully processed and will be released on the date specified in the email. The email does not contain the SRA accession numbers.*



SRA submission SUB14394290, "UI-less submission 2024-04-22" - Message (Plain Text) (Read-Only)

File Message Help Tell me what you want to do

Share to Teams | To-do | Mark Unread | Find | Zoom

SRA submission SUB14394290, "UI-less submission 2024-04-22"

sra@ncbi.nlm.nih.gov

To Carleton, Heather (CDC/NCEZID/DFWED/EDLB); Katz, Lee S. (CDC/NCEZID/DFWED/EDLB); Williams, Amanda (Jo) (CDC/NCEZID/DFWED/EDLB)

We removed extra line breaks from this message.

Start your reply all with: Thank you! | Received, thank you. | Got it, thanks! | Feedback

CAUTION: This email originated from outside of the organization. Do not click links or open attachments unless you recognize the sender and know the content is safe.

Dear edlb-cdc service,

This is an automatic acknowledgment that your recent submission to the SRA database has been successfully processed and will be released on the date specified.

Please reference PRJNA590944 in your publication. This BioProject accession number is provided instead of SRP and should be used in your publication as it will allow better searching in Entrez.

Accession to cite for these SRA data: PRJNA590944 Temporary Submission ID: SUB14394290 Release date: 2024-04-22

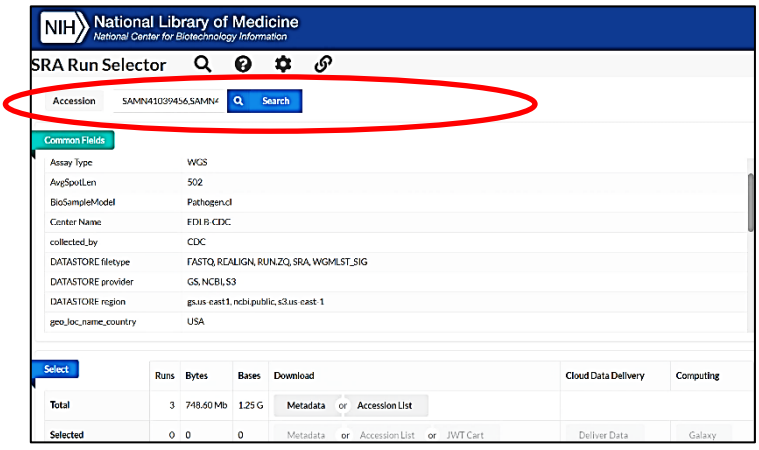
Your SRA records will be accessible with the following link after the indicated release date:
https://www.ncbi.nlm.nih.gov/sra/PRJNA590944

Send questions and update requests to sra@ncbi.nlm.nih.gov; include the citation accession PRJNA590944 in any correspondence.

Regards,

NCBI SRA Submissions Staff
Bethesda, Maryland USA

5.7.3.11.2.1 Copy and paste the SAMN numbers from the tsv file from step 5.7.3.11.1.2. or the txt file from step 5.7.3.11.1.3. to the "NCBI Run Selector" tool at: https://0-www-ncbi-nlm-nih-gov.brum.beds.ac.uk/Traces/study/. Separate the numbers with commas. Click on "Search".



NIH National Library of Medicine
National Center for Biotechnology Information

SRA Run Selector

Accession SAMN41039456,SAMN4 Search

Common Fields

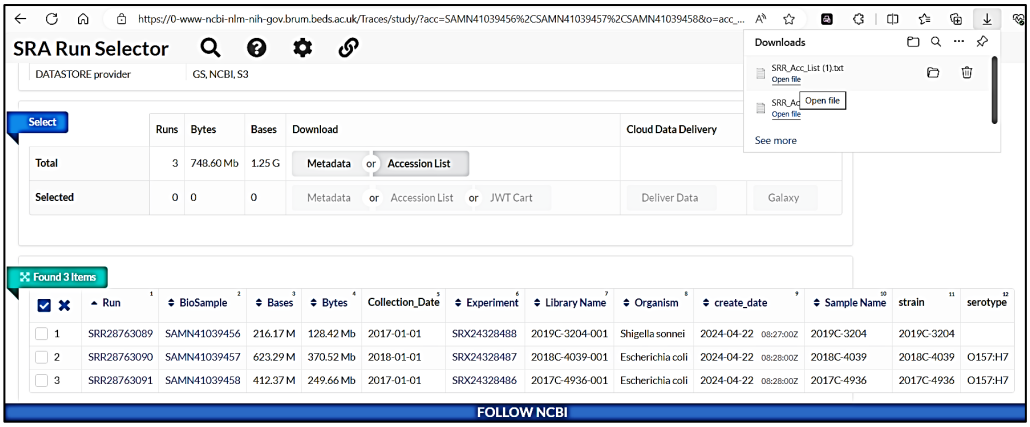
Assay Type	WGS
AvgSpotLen	502
BioSampleModel	Pathogen.cl
Center Name	EDLB-CDC
collected_by	CDC
DATASTORE filetype	FASTQ,REALIGN,RUN.ZQ,SRA,WGMLST_SIG
DATASTORE provider	GS,NCBI,S3
DATASTORE region	gs.us-east1,ncbi.public,s3.us-east-1
geo_loc_name_country	USA

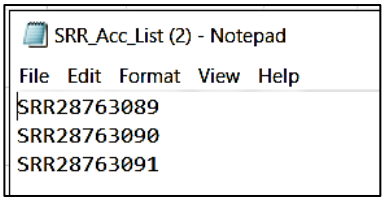
Select

	Runs	Bytes	Bases	Download	Cloud Data Delivery	Computing
Total	3	748.60 Mb	1.25 G	Metadata or Accession List		
Selected	0	0	0	Metadata or Accession List or JWT Cart	Deliver Data	Galaxy

5.7.3.11.2.2 The search will return a table that contains all the information NCBI has about your sequences, including the SRR accession numbers.

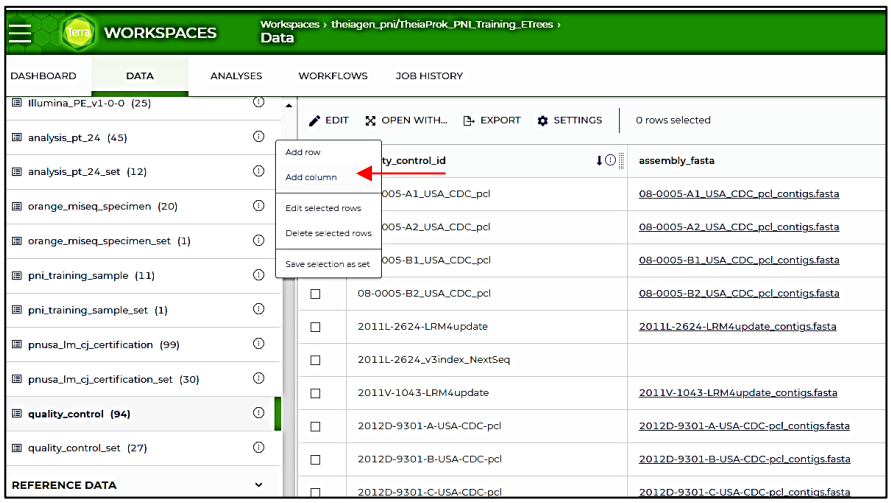
5.7.3.11.2.3 Click on "Accession List" to download the table containing the accession numbers. The downloaded file will appear on the upper right corner.



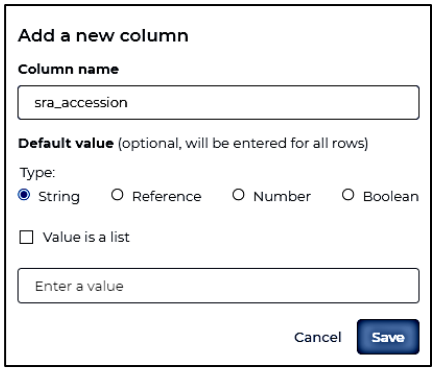


5.7.3.11.2.4 You can track the SRA accession numbers either on a separate Excel spreadsheet, LIMS system or you can create an "sra_accession" field in your Terra data table and copy and paste the accession numbers there for each sample entry. To create a sra_accession column and use it for tracking:

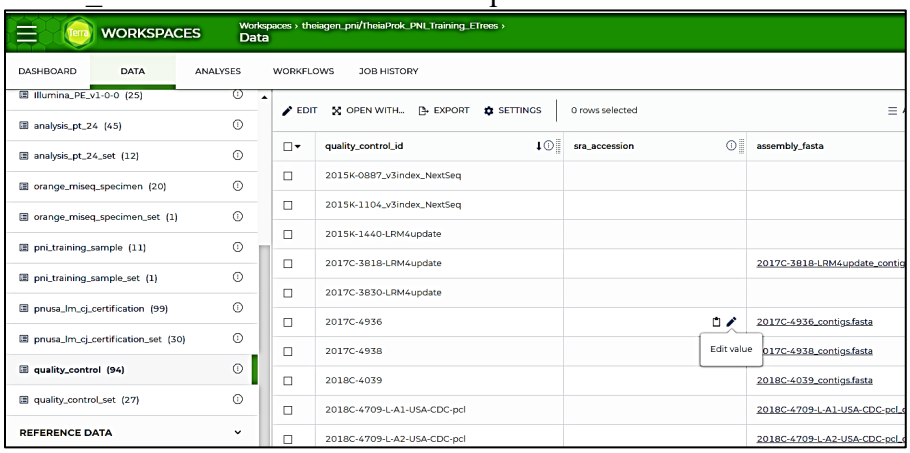
5.7.3.11.2.4.1 From the "Edit" drop-down menu, select "Add column".



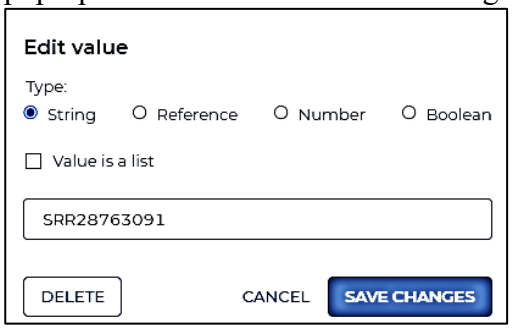
5.7.3.11.2.4.2 In the "Add a new column" pop-up window, name the new column "sra_accession" and click "Save".



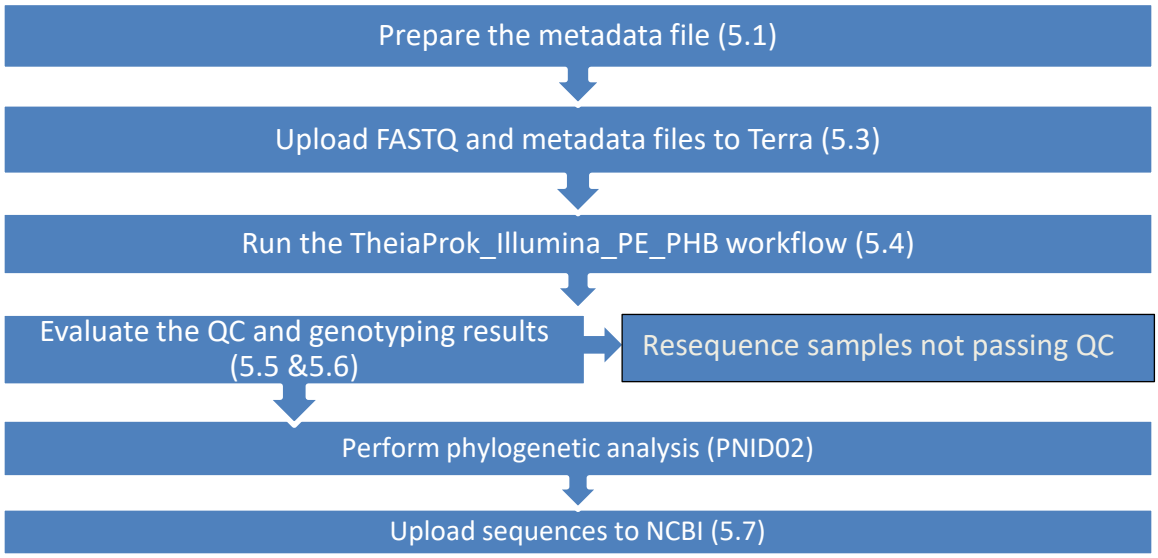
5.7.3.11.2.4.3 The new column should appear in the data table. To enter a SRR accession number for a specific sequence, click on “Edit value” in the sra_accession column for that sample.



5.7.3.11.2.4.4 Paste the SRA accession number to the field in the “Edit value” pop-up window and click “Save changes”.



6. FLOW CHART:



7. RELATED DOCUMENTS:

7.1 **PNID02:** PulseNet International Standard Operating Procedure for Phylogenetic Analysis of WGS Data Using the Terra.Bio Platform.

8. REFERENCES:

8.1 Libuit K.G., Doughty E.L., Otieno J.R., Ambrosio F., Kapsak C.J., Smith E.A., Wright S.M., Scribner M.R., Petit III R.A., Mendes C.I., Huergo M., Legacki G., Loreth C., Park D.J., Sevinsky J.R. (2023) Accelerating bioinformatics implementation in public health. Microbial Genomics 9:001051.

9. CONTACTS:

9.1 CDC USA PulseNet NGS Laboratory: pulsenetngslab@cdc.gov
9.2 PulseNet International Quality Assurance Coordinator Eija Trees: ehyytia-trees@cdc.gov
9.3 Theiagen:
9.3.1 Generic email for support : support@theiagen.com
9.3.2 Michelle Scribner: michelle.scribner@theiagen.com
9.3.3 Frank Ambrosio: frank.ambrosio@theiagen.com

10. AMENDMENTS: NA

11. APPROVAL SIGNATURES:

Approved By: Eija K. Trees Digitally signed by Eija K. Trees Date: 2024.07.25 07:34:38 -04'00' Date: 07/25/2024
PulseNet QA/QC Personnel

Approved By: Peyton A. Smith -S Digitally signed by Peyton A. Smith -S Date: 2024.07.03 09:32:52 -04'00' Date: 07/03/2024
PulseNet WGS Technical Lead

Approved By: KRISTY A. KUBOTA -S Digitally signed by KRISTY A. KUBOTA -S Date: 2024.06.14 14:42:05 -04'00' Date: 06/14/2024
PulseNet International Coordinator

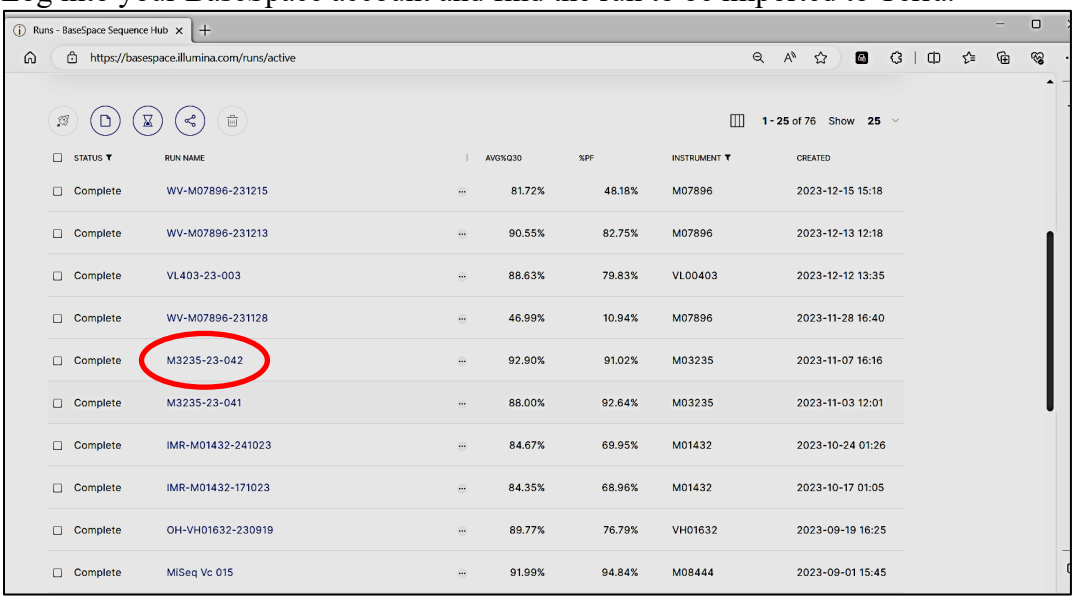
Approved By: Kelley B Hise Digitally signed by Kelley B Hise Date: 2024.06.13 08:45:25 -04'00' Date: 06/13/2024
PulseNet Response and Outbreak Management Team Lead

Approved By: Heather A. Carleton -S Digitally signed by Heather A. Carleton -S Date: 2024.07.03 11:41:30 -04'00' Date: 07/03/2024
Enteric Diseases Laboratory Branch Chief

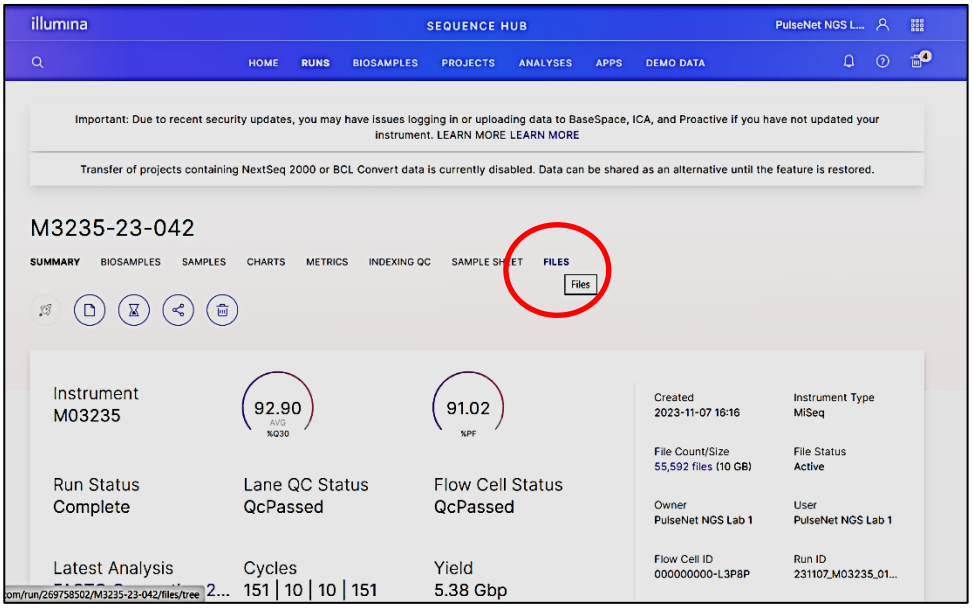
Appendix PNID01-1: Data Import to Terra Directly from the Illumina BaseSpace

NOTE: *In order to configure your Terra workspace to connect to your Illumina BaseSpace account, follow the instructions found in the Theiagen resources site at https://theiagen.notion.site/BaseSpace_Fetch-34978656aa2d46ba82f2059434bd9369. For further assistance, contact Theiagen (see the Contacts section (9) for contact information).*

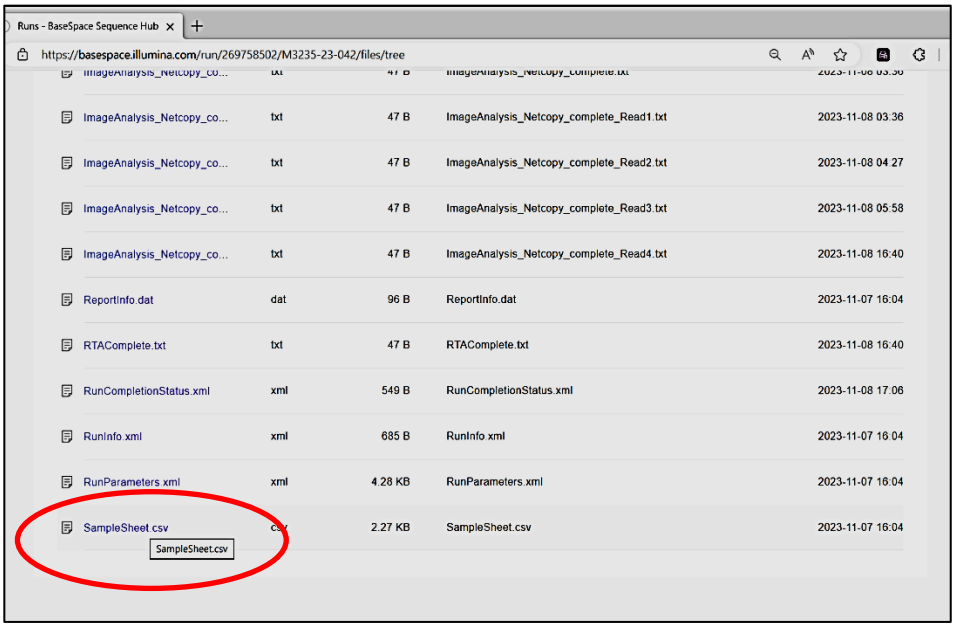
1. Log into your BaseSpace account and find the run to be imported to Terra.



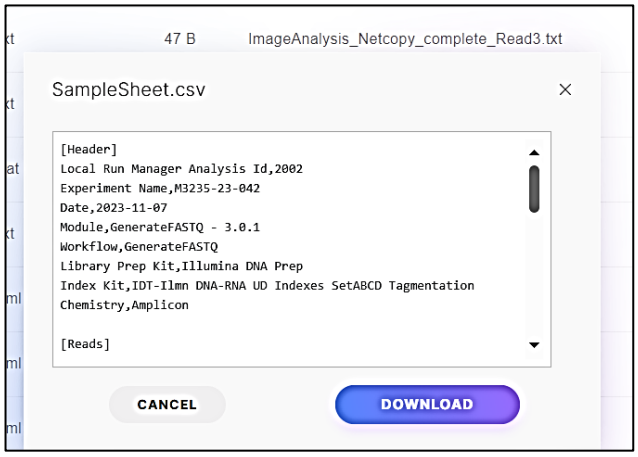
2. Download the SampleSheet for the run:
   a. Switch to the "Files" tab and scroll all the way to the bottom.



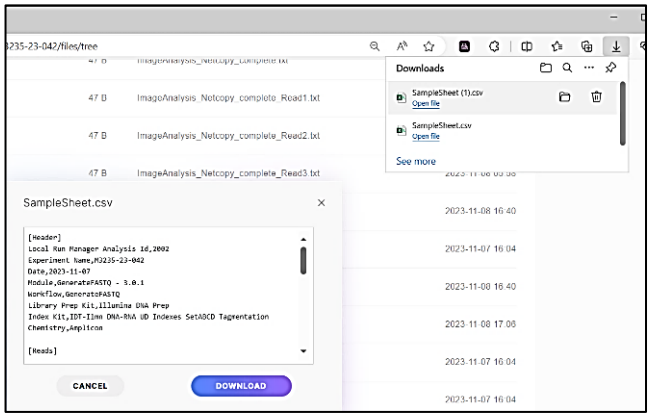
   b. Click on the SampleSheet.csv link.



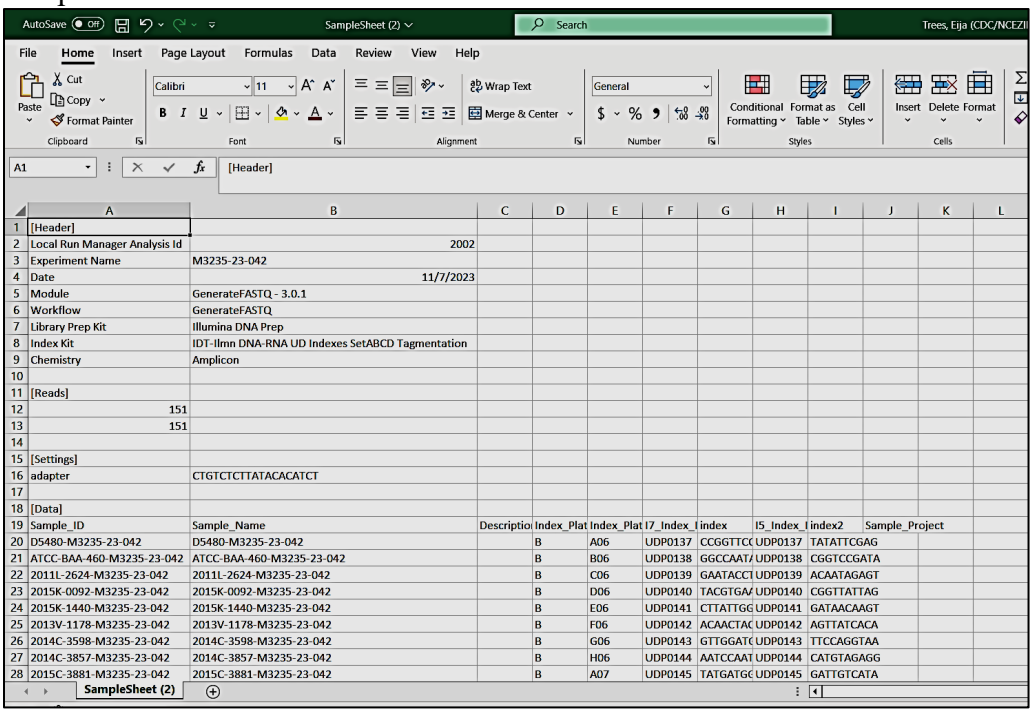
c. In the "SampleSheet.csv" pop-up window, click on "Download".



d. The downloaded csv file will appear on the top right corner under "Downloads".

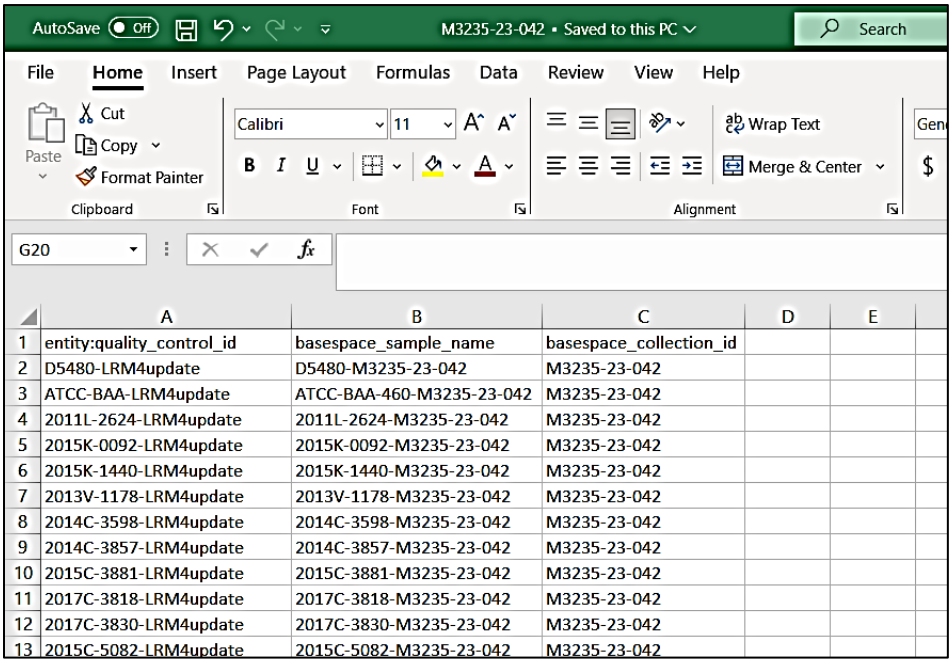


3. Open the SampleSheet. The columns that are needed in the metadata tsv file will depend on the columns and column content present in the SampleSheet.
4. Prepare the metadata tsv file:
    a. Columns needed when the "Sample_Name" and "Sample_ID" columns in the SampleSheet have the exact **same** content:



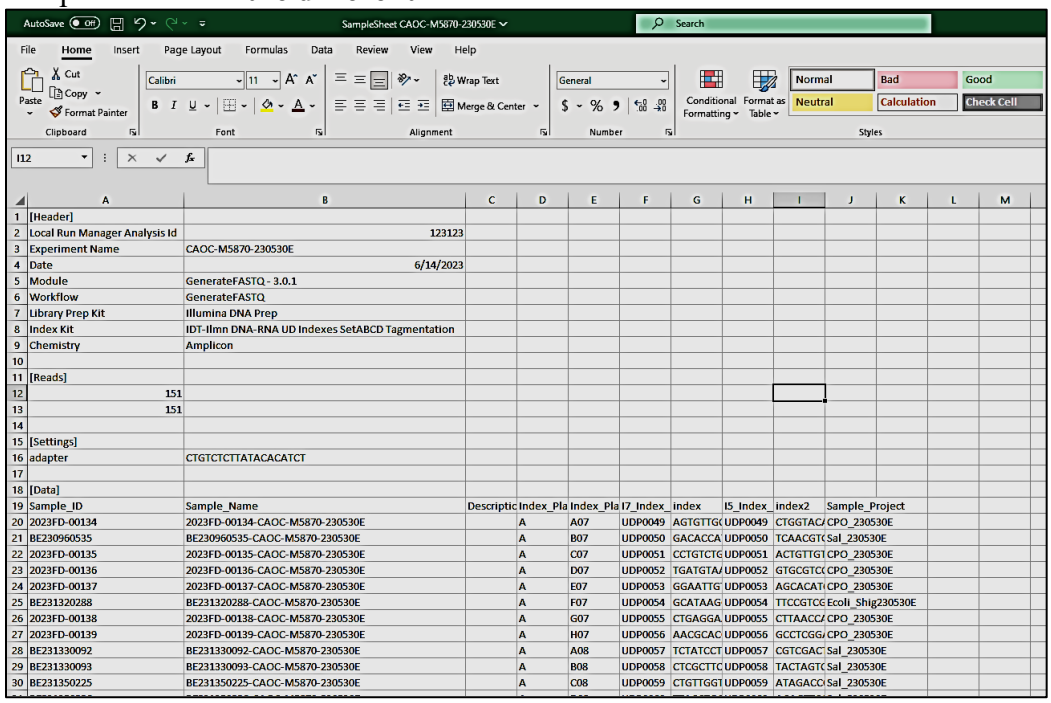
	A	B	C	D	E	F	G	H	I	J	K	L
1	[Header]											
2	Local Run Manager Analysis Id	2002										
3	Experiment Name	M3235-23-042										
4	Date	11/7/2023										
5	Module	GenerateFASTQ - 3.0.1										
6	Workflow	GenerateFASTQ										
7	Library Prep Kit	Illumina DNA Prep										
8	Index Kit	IDT-Ilmn DNA-RNA UD Indexes SetABCD Tagmentation										
9	Chemistry	Amplicon										
10												
11	[Reads]											
12	151											
13	151											
14												
15	[Settings]											
16	adapter	CTGTCTCTTATACACATCT										
17												
18	[Data]											
19	Sample_ID	Sample_Name	Description	Index_Plat	Index_Plat	I7_Index_I	index	I5_Index_I	index2	Sample_Project		
20	D5480-M3235-23-042	D5480-M3235-23-042		B	A06	UDP0137	CCGGTTCC	UDP0137	TATATTCGAG			
21	ATCC-BAA-460-M3235-23-042	ATCC-BAA-460-M3235-23-042		B	B06	UDP0138	GGCCAATA	UDP0138	CGGTCCGATA			
22	2011L-2624-M3235-23-042	2011L-2624-M3235-23-042		B	C06	UDP0139	GAATACCT	UDP0139	ACAATAGAGT			
23	2015K-0092-M3235-23-042	2015K-0092-M3235-23-042		B	D06	UDP0140	TACGTGAA	UDP0140	CGGTTATTAG			
24	2015K-1440-M3235-23-042	2015K-1440-M3235-23-042		B	E06	UDP0141	CTTATTGG	UDP0141	GATAACAAGT			
25	2013V-1178-M3235-23-042	2013V-1178-M3235-23-042		B	F06	UDP0142	ACAACTAC	UDP0142	AGTTATCACA			
26	2014C-3598-M3235-23-042	2014C-3598-M3235-23-042		B	G06	UDP0143	GTTGGATC	UDP0143	TTCCAGGTAA			
27	2014C-3857-M3235-23-042	2014C-3857-M3235-23-042		B	H06	UDP0144	AATCCAAT	UDP0144	CATGTAGAGG			
28	2015C-3881-M3235-23-042	2015C-3881-M3235-23-042		B	A07	UDP0145	TATGATGC	UDP0145	GATTGTCATA			

        i. **entity**_*datatablename*_**id**:
            1. Enter the name of the data table (either new or existing) into the cell A1 between "entity:" and "id".
            2. Enter the sample IDs into the column A the way you want them to appear in the Terra data table.
        ii. basespace_sample_name: copy and paste the content from the SampleSheet "Sample_Name" field.
        iii. basespace_collection_id: enter the Run name the way it appears on BaseSpace.



	A	B	C	D	E
1	entity:quality_control_id	basespace_sample_name	basespace_collection_id		
2	D5480-LRM4update	D5480-M3235-23-042	M3235-23-042		
3	ATCC-BAA-LRM4update	ATCC-BAA-460-M3235-23-042	M3235-23-042		
4	2011L-2624-LRM4update	2011L-2624-M3235-23-042	M3235-23-042		
5	2015K-0092-LRM4update	2015K-0092-M3235-23-042	M3235-23-042		
6	2015K-1440-LRM4update	2015K-1440-M3235-23-042	M3235-23-042		
7	2013V-1178-LRM4update	2013V-1178-M3235-23-042	M3235-23-042		
8	2014C-3598-LRM4update	2014C-3598-M3235-23-042	M3235-23-042		
9	2014C-3857-LRM4update	2014C-3857-M3235-23-042	M3235-23-042		
10	2015C-3881-LRM4update	2015C-3881-M3235-23-042	M3235-23-042		
11	2017C-3818-LRM4update	2017C-3818-M3235-23-042	M3235-23-042		
12	2017C-3830-LRM4update	2017C-3830-M3235-23-042	M3235-23-042		
13	2015C-5082-LRM4update	2015C-5082-M3235-23-042	M3235-23-042		

b. Columns needed when the "Sample_Name" and "Sample_ID" columns in the SampleSheet have **the different** content:

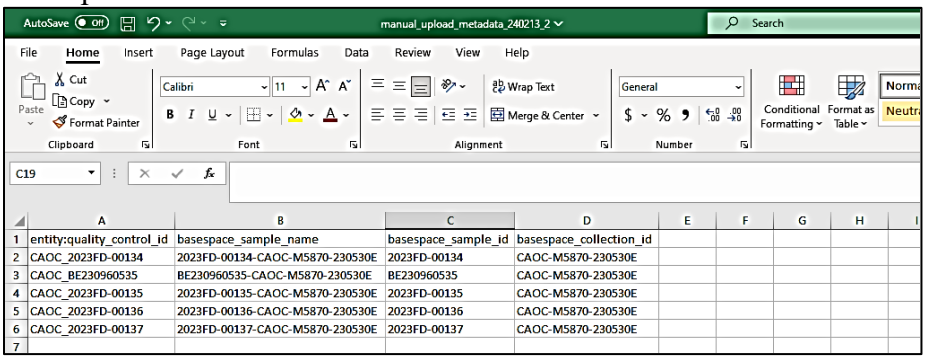


	A	B	C	D	E	F	G	H	I	J	K	L	M
1	[Header]												
2	Local Run Manager Analysis Id	123123											
3	Experiment Name	CAOC-M5870-230530E											
4	Date	6/14/2023											
5	Module	GenerateFASTQ - 3.0.1											
6	Workflow	GenerateFASTQ											
7	Library Prep Kit	Illumina DNA Prep											
8	Index Kit	IDT-Ilmn DNA-RNA UD Indexes SetABCD Tagmentation											
9	Chemistry	Amplicon											
10													
11	[Reads]												
12	151												
13	151												
14													
15	[Settings]												
16	adapter	CTGTCTCTTATACACATCT											
17													
18	[Data]												
19	Sample_ID	Sample_Name	Descriptic	Index_Pla	Index_Pla	I7_Index_	index	I5_Index_	index2	Sample_Project			
20	2023FD-00134	2023FD-00134-CAOC-M5870-230530E		A	A07	UDP0049	AGTGTTG(	UDP0049	CTGGTAC(	CPO_230530E			
21	BE230960535	BE230960535-CAOC-M5870-230530E		A	B07	UDP0050	GACACCA	UDP0050	TCAACGT(	Sal_230530E			
22	2023FD-00135	2023FD-00135-CAOC-M5870-230530E		A	C07	UDP0051	CCTGTCTG	UDP0051	ACTGTTG1	CPO_230530E			
23	2023FD-00136	2023FD-00136-CAOC-M5870-230530E		A	D07	UDP0052	TGATGTA/	UDP0052	GTGCGTC(	CPO_230530E			
24	2023FD-00137	2023FD-00137-CAOC-M5870-230530E		A	E07	UDP0053	GGAATTG	UDP0053	AGCACAT(	CPO_230530E			
25	BE231320288	BE231320288-CAOC-M5870-230530E		A	F07	UDP0054	GCATAAG	UDP0054	TTCCGTCG	Ecoli_Shig230530E			
26	2023FD-00138	2023FD-00138-CAOC-M5870-230530E		A	G07	UDP0055	CTGAGGA	UDP0055	CTTAACC/	CPO_230530E			
27	2023FD-00139	2023FD-00139-CAOC-M5870-230530E		A	H07	UDP0056	AACGCAC	UDP0056	GCCTCGG/	CPO_230530E			
28	BE231330092	BE231330092-CAOC-M5870-230530E		A	A08	UDP0057	TCTATCCT	UDP0057	CGTCGAC(	Sal_230530E			
29	BE231330093	BE231330093-CAOC-M5870-230530E		A	B08	UDP0058	CTCGCTTC	UDP0058	TACTAGT(	Sal_230530E			
30	BE231350225	BE231350225-CAOC-M5870-230530E		A	C08	UDP0059	CTGTTGGT	UDP0059	ATAGACG(	Sal_230530E			

i. **entity**_*datatablename*_**id**:

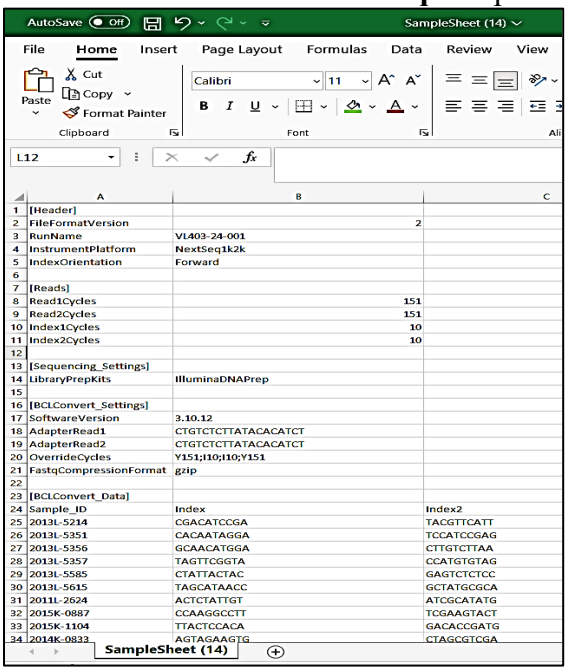
1. Enter the name of the data table (either new or existing) into the cell A1 between "entity:" and "id".
2. Enter the sample IDs into the column A the way you want them to appear in the Terra data table.

ii. basespace_sample_name: copy and paste the content from the SampleSheet "Sample_Name" field.
iii. basespace_sample_id: copy and paste the content from the SampleSheet "Sample_ID" field.
iv. basespace_collection_id: enter the Run name the way it appears on BaseSpace.



	A	B	C	D	E	F	G	H
1	entity:quality_control_id	basespace_sample_name	basespace_sample_id	basespace_collection_id				
2	CAOC_2023FD-00134	2023FD-00134-CAOC-M5870-230530E	2023FD-00134	CAOC-M5870-230530E				
3	CAOC_BE230960535	BE230960535-CAOC-M5870-230530E	BE230960535	CAOC-M5870-230530E				
4	CAOC_2023FD-00135	2023FD-00135-CAOC-M5870-230530E	2023FD-00135	CAOC-M5870-230530E				
5	CAOC_2023FD-00136	2023FD-00136-CAOC-M5870-230530E	2023FD-00136	CAOC-M5870-230530E				
6	CAOC_2023FD-00137	2023FD-00137-CAOC-M5870-230530E	2023FD-00137	CAOC-M5870-230530E				
7								

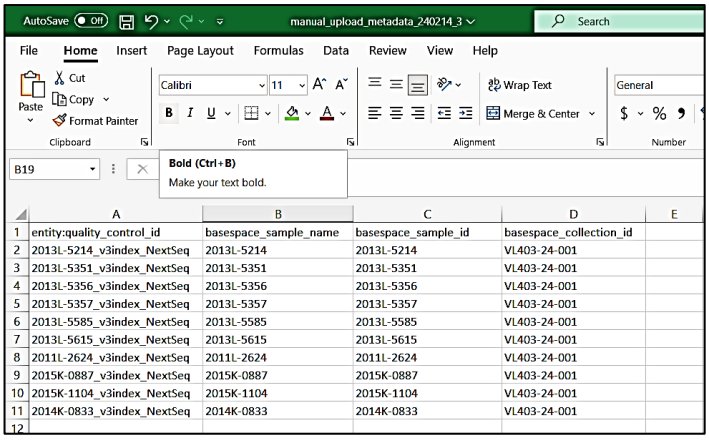
c. Columns needed for the **NextSeq** SampleSheet:



	A	B	C
1	[Header]		
2	FileFormatVersion	2	
3	RunName	VL403-24-001	
4	InstrumentPlatform	NextSeq1k2k	
5	IndexOrientation	Forward	
6			
7	[Reads]		
8	Read1Cycles	151	
9	Read2Cycles	151	
10	Index1Cycles	10	
11	Index2Cycles	10	
12			
13	[Sequencing_Settings]		
14	LibraryPrepKits	IlluminaDNAPrep	
15			
16	[BCLConvert_Settings]		
17	SoftwareVersion	3.10.12	
18	AdapterRead1	CTGTCTCTTATACACATCT	
19	AdapterRead2	CTGTCTCTTATACACATCT	
20	OverrideCycles	Y151;I10;I10;Y151	
21	FastqCompressionFormat	gzip	
22			
23	[BCLConvert_Data]		
24	Sample_ID	Index	Index2
25	2013L-5214	CGACATCCGA	TACGTTCATT
26	2013L-5351	CACAATAGGA	TCCATCCGAG
27	2013L-5356	GCAACATGGA	CTTGTCTTAA
28	2013L-5357	TAGTTCGGTA	CCATGTGTAG
29	2013L-5585	CTATTACTAC	GAGTCTCTCC
30	2013L-5615	TAGCATAACC	GCTATGCGCA
31	2011L-2624	ACTCTATTGT	ATCGCATATG
32	2015K-0887	CCAAGGCCTT	TCGAAGTACT
33	2015K-1104	TTACTCCACA	GACACCGATG
34	2014K-0833	AGTAGAAGTG	CTAGCGTCGA

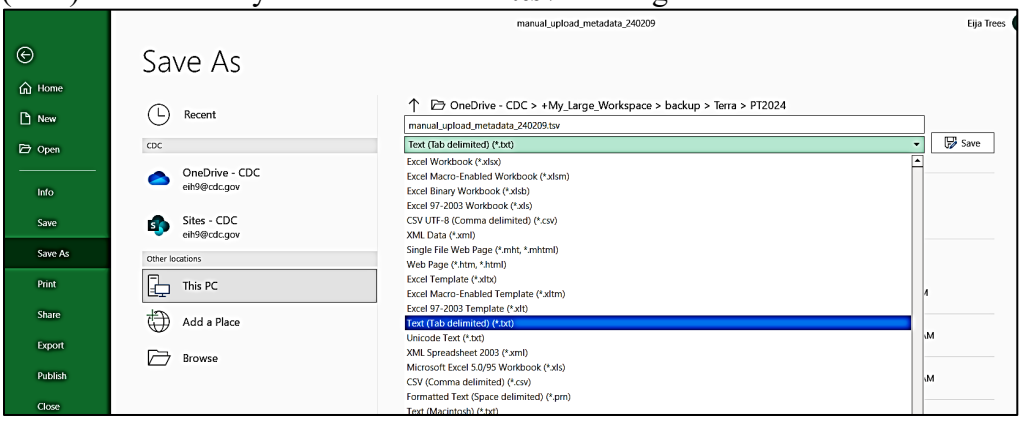
i. **entity**_*datatablename*_**id**:
    1. Enter the name of the data table (either new or existing) into the cell A1 between "entity:" and "id".
    2. Enter the sample IDs into the column A the way you want them to appear in the Terra data table.
ii. basespace_sample_name: copy and paste the content from the SampleSheet "Sample_ID" field.
iii. basespace_sample_id: copy and paste the content from the SampleSheet "Sample_ID" field.

iv. basespace_collection_id: enter the Run name the way it appears on BaseSpace.

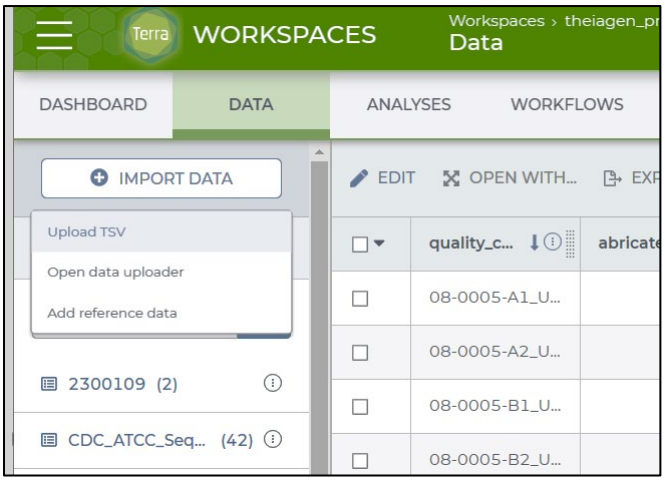


	A	B	C	D	E
1	entity:quality_control_id	basespace_sample_name	basespace_sample_id	basespace_collection_id	
2	2013L-5214_v3index_NextSeq	2013L-5214	2013L-5214	VL403-24-001	
3	2013L-5351_v3index_NextSeq	2013L-5351	2013L-5351	VL403-24-001	
4	2013L-5356_v3index_NextSeq	2013L-5356	2013L-5356	VL403-24-001	
5	2013L-5357_v3index_NextSeq	2013L-5357	2013L-5357	VL403-24-001	
6	2013L-5585_v3index_NextSeq	2013L-5585	2013L-5585	VL403-24-001	
7	2013L-5615_v3index_NextSeq	2013L-5615	2013L-5615	VL403-24-001	
8	2011L-2624_v3index_NextSeq	2011L-2624	2011L-2624	VL403-24-001	
9	2015K-0887_v3index_NextSeq	2015K-0887	2015K-0887	VL403-24-001	
10	2015K-1104_v3index_NextSeq	2015K-1104	2015K-1104	VL403-24-001	
11	2014K-0833_v3index_NextSeq	2014K-0833	2014K-0833	VL403-24-001	
12					

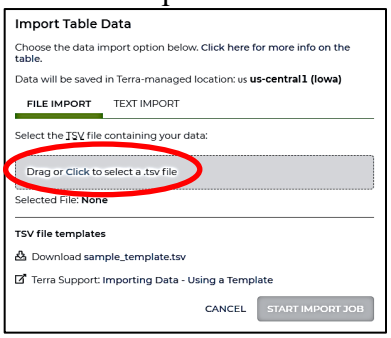
d. Save the file in the **tsv format**: choose "Save As" and "Text (Tab delimited) (*.txt)". Make sure your file name has "**.tsv**" ending.

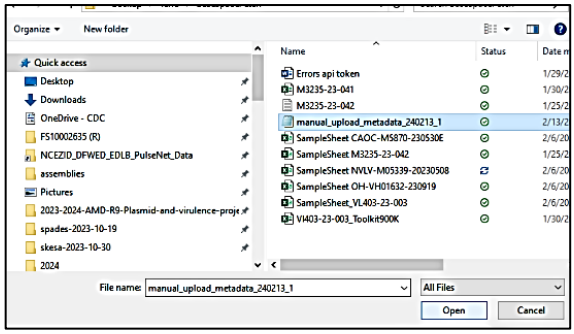


5. Import the metadata tsv file:
   a. In the "Data" tab, click on "Import Data" and select "Upload tsv" from the drop-down menu.

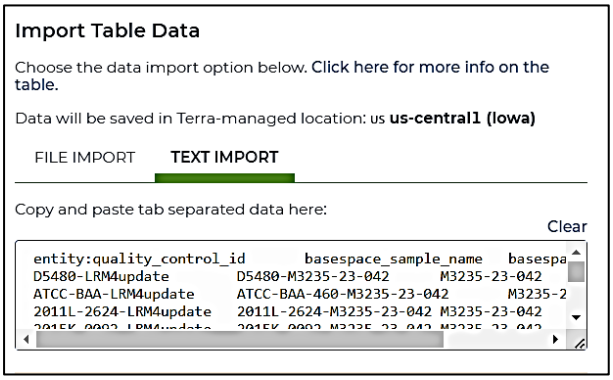


b. From the "Import Table Data" pop-up window, you can import the metadata in two ways:
   i. In the "File Import" tab, click in the middle to select the tsv file and navigate to the location where the metadata tsv file is saved, select the file and click "Open".

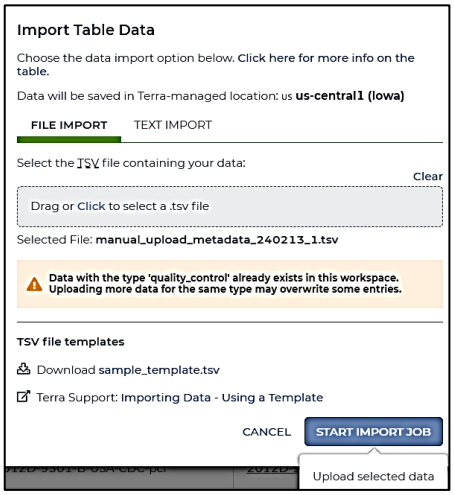




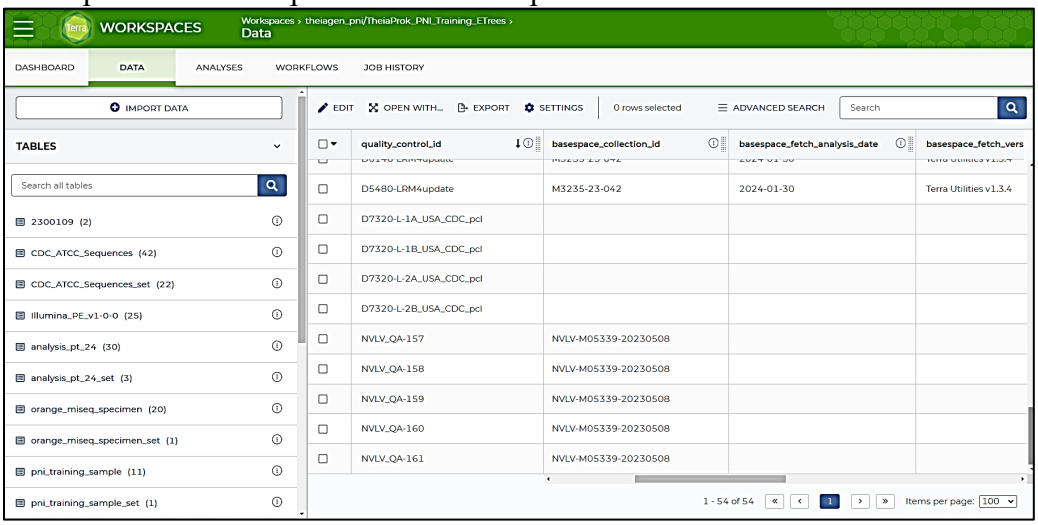
   ii. Alternatively, you can switch to the "text Import" tab and copy and paste the contents of the tsv file into the field in the middle.



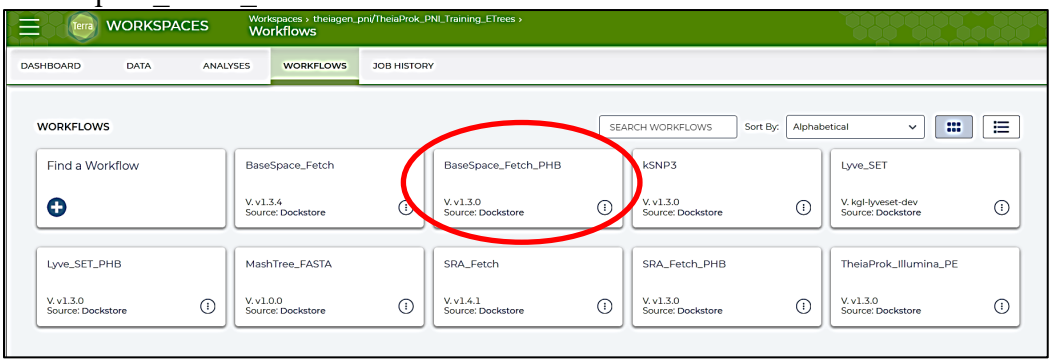
c. In the "Import Table Data" pop-up window, you will get a warning that data already exists in the data table in question (if importing to an existing data table) and uploading more data to it may overwrite existing data. Click on "Start Import Job".



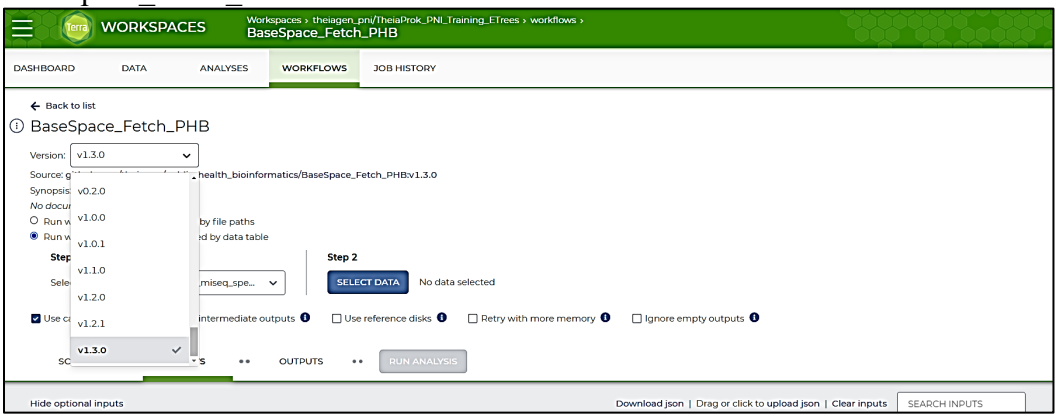
d. After the import is done you should see new entries created to the data table for the sequences to be imported from BaseSpace.



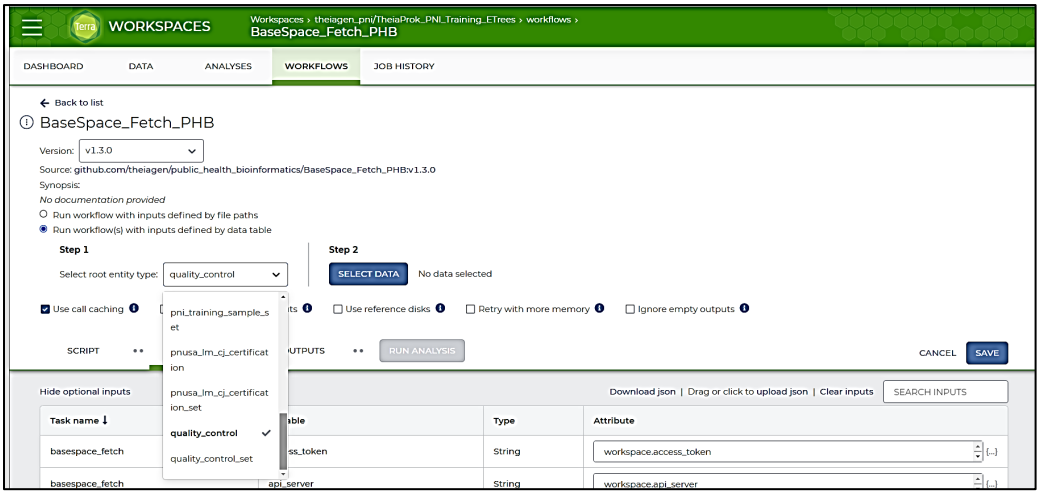
6. Run the "BaseSpace_Fetch_PHB" workflow:
   a. In the "Workflows" tab, click on "BaseSpace_Fetch_PHB". This will open the "BaseSpace_Fetch_PHB" screen.



   b. From the "Version" drop-down menu, select the latest version of BaseSpace_Fetch_PHB.

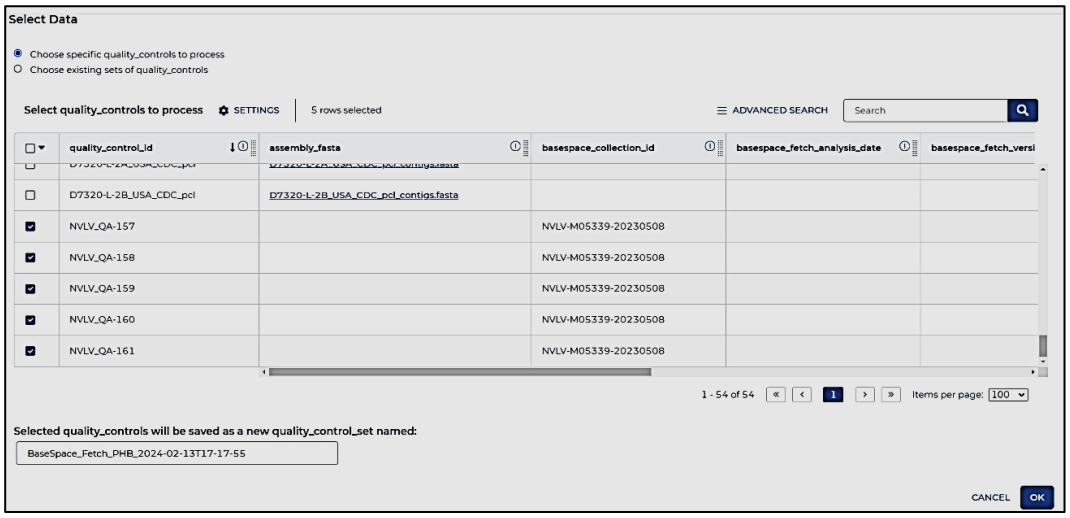


   c. Under "Step 1", click on the "Select root entity type" drop-down menu and select the data table into which you uploaded (steps 4-5) the tsv file containing the metadata for the run to be imported from BaseSpace.

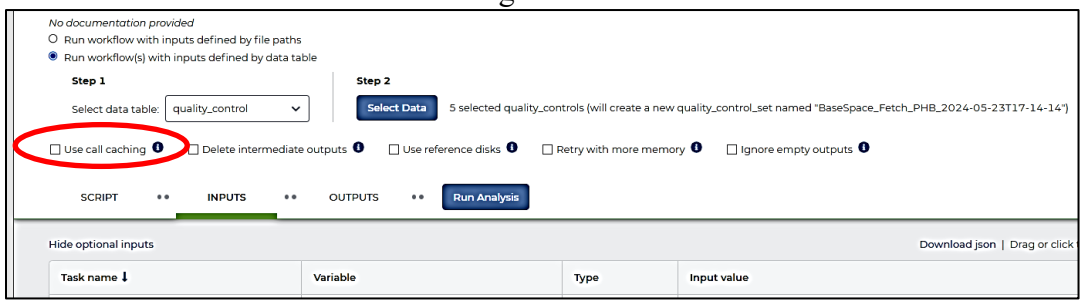


d. Under “Step 2”, click on “Select data” (screenshot above). This will take you to the sample selection screen.

e. Check the boxes next to the samples to be imported from BaseSpace and click “OK”. This will take you back to the “BaseSpace_Fetch_PHB” screen.

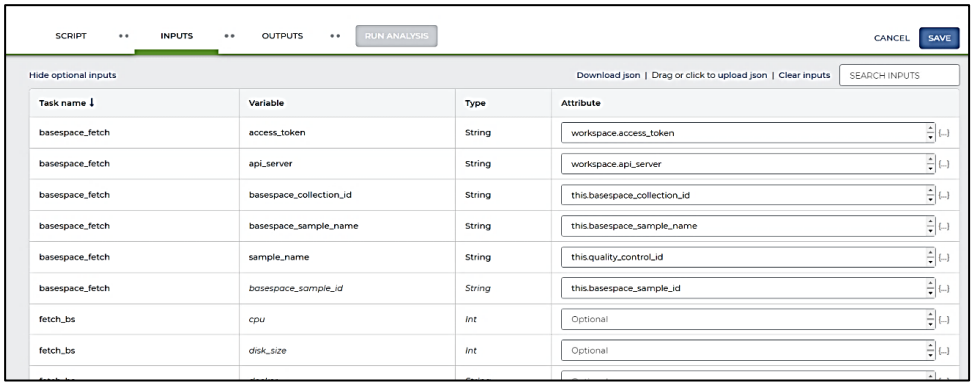


f. De-select the box for “Use call catching”.

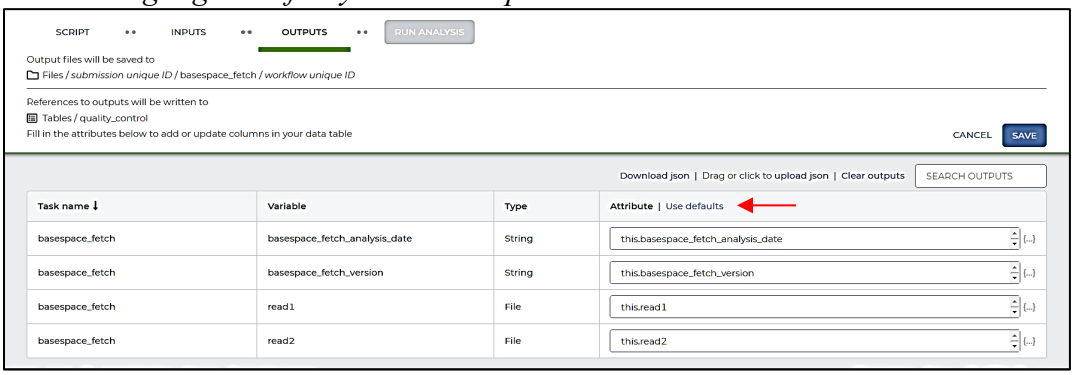


g. In the “**Inputs**” tab, define the following variables in the “Attribute” field:
NOTE: *When you fill in the Attribute column, clicking inside the cell will bring up a drop-down menu of attributes that you can selectk to avoid typos*
    i. access_token: “workspace.access_token”.

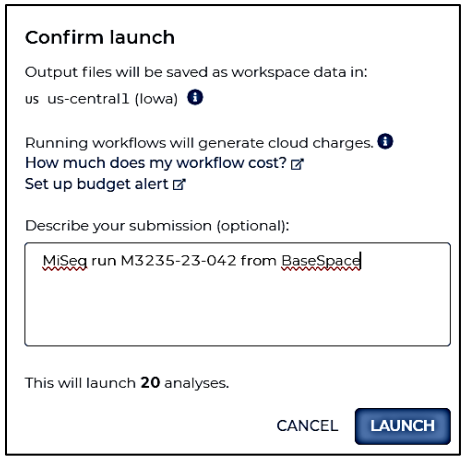
ii. api_server: "workspace.api_server".
iii. basespace_collection_id: "this.basespace_collection_id".
iv. basespace_sample_name: "this.basespace_sample_name".
v. sample_name: "**this**.*datatablename*_**id**", e.g. "this.quality_control_id".
vi. basespace_sample_id: "this.basespace_sample_id".
NOTE: *needs to be filled out **only** if the contents in the "Sample_Name" and "Sample_ID" fields are different in the run SampleSheet or you are importing data from a NextSeq run.*



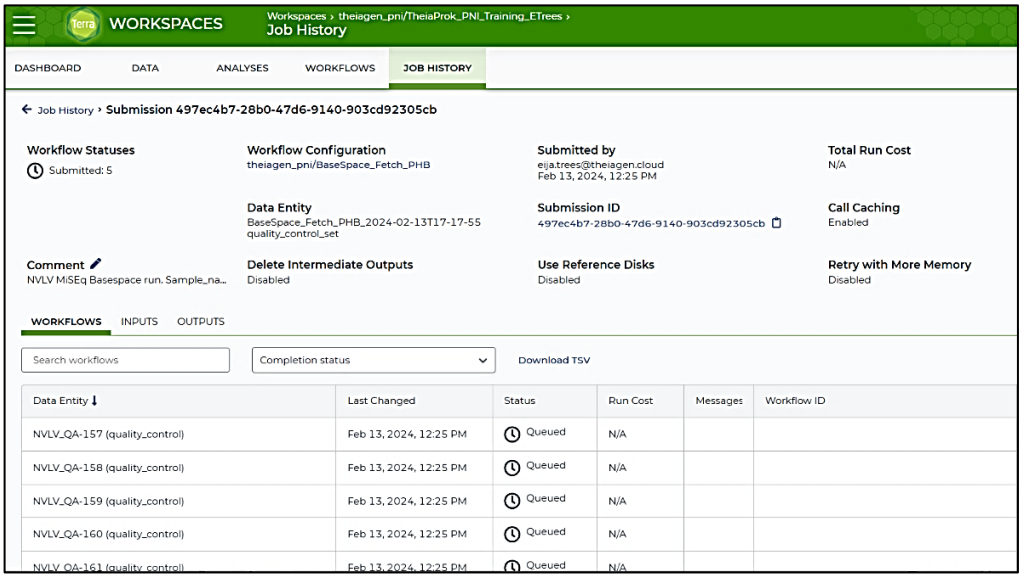
h. In the "**Outputs**" tab, click "Use defaults", then click "Save", and then "Run Analysis".
NOTE: *The "Save" button is only visible if parameters (other than sample IDs) have changed from the previous job submission. The "Run Analysis" button becomes highlighted after you save the parameters.*



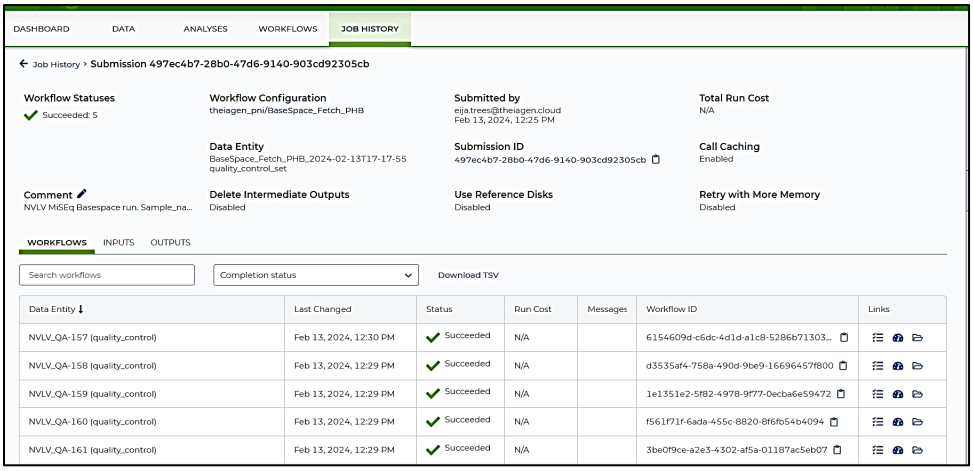
i. In the "Confirm launch" pop-up window, describe your job submission (optional) and click "Launch".

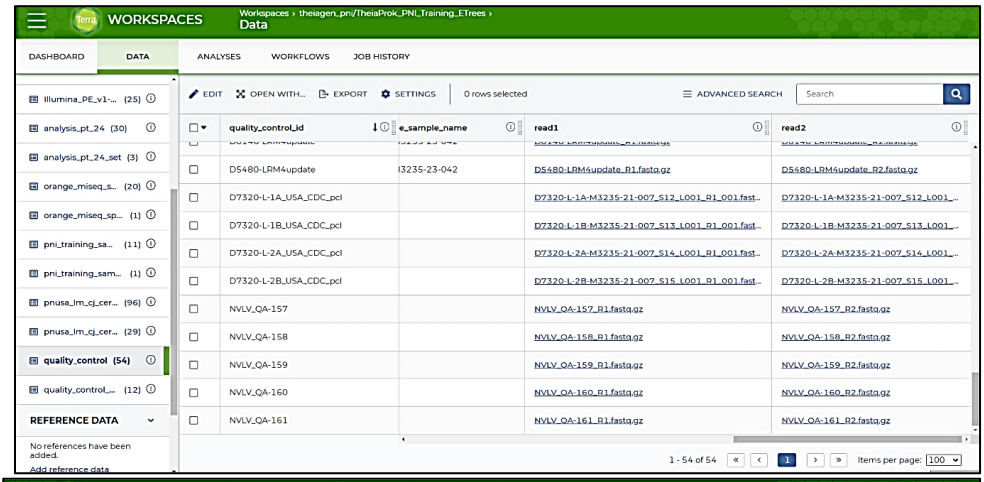


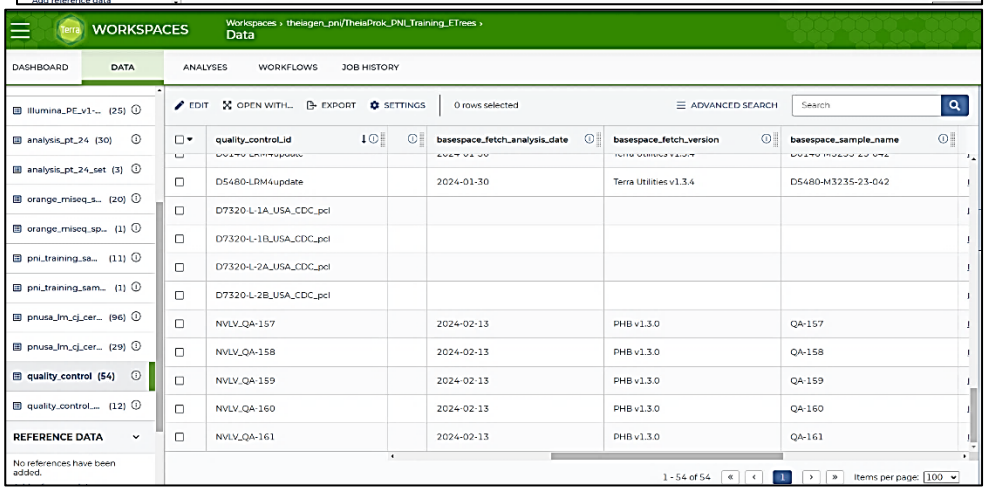
j. The "Job History" tab opens where the status of the job submissions should initially be "Queued".



k. The status will be "Succeeded" or "Done" once the job has finished. In the "Data" tab, you should now see the FASTQ file names in "Read1" and "Read2" columns and also information in columns "basespace_fetch_analysis_date" and "basespace_fetch_version".

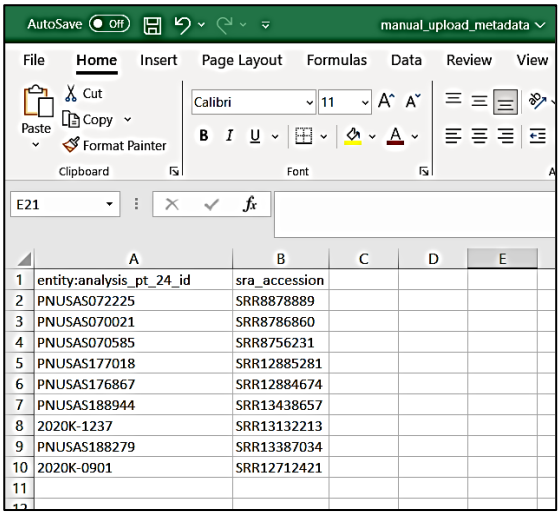

DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
Job History › Submission 497ec4b7-28b0-47d6-9140-903cd92305cb
Workflow Statuses
Succeeded: 5
Workflow Configuration
theiagen_pni/BaseSpace_Fetch_PHB
Submitted by
eija.trees@theiagen.cloud
Feb 13, 2024, 12:25 PM
Total Run Cost
N/A
Data Entity
BaseSpace_Fetch_PHB_2024-02-13T17-17-55
quality_control_set
Submission ID
497ec4b7-28b0-47d6-9140-903cd92305cb
Call Caching
Enabled
Comment
NVLV MiSEq Basespace run, Sample_na...
Delete Intermediate Outputs
Disabled
Use Reference Disks
Disabled
Retry with More Memory
Disabled
WORKFLOWS
INPUTS
OUTPUTS
Search workflows
Completion status
Download TSV
Data Entity
Last Changed
Status
Run Cost
Messages
Workflow ID
Links
NVLV_QA-157 (quality_control)
Feb 13, 2024, 12:30 PM
Succeeded
N/A
6154609d-c6dc-4d1d-a1c8-5286b71303...
NVLV_QA-158 (quality_control)
Feb 13, 2024, 12:29 PM
Succeeded
N/A
d3535af4-758a-490d-9be9-16696457f800
NVLV_QA-159 (quality_control)
Feb 13, 2024, 12:29 PM
Succeeded
N/A
1e1351e2-5f82-4978-9f77-0ecba6e59472
NVLV_QA-160 (quality_control)
Feb 13, 2024, 12:29 PM
Succeeded
N/A
f561f71f-6ada-455c-8820-8f6fb54b4094
NVLV_QA-161 (quality_control)
Feb 13, 2024, 12:29 PM
Succeeded
N/A
3be0f9ce-a2e3-4302-af5a-01187ac5eb07


WORKSPACES
Workspaces › theiagen_pni/TheiaProk_PNI_Training_ETrees ›
Data
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
EDIT
OPEN WITH...
EXPORT
SETTINGS
0 rows selected
ADVANCED SEARCH
Search
quality_control_id
e_sample_name
read1
read2
D5480-LRM4update
D5480-LRM4update_R1.fasta.gz
D5480-LRM4update_R2.fasta.gz
D7320-L-1A_USA_CDC_pcl
D7320-L-1B_USA_CDC_pcl
D7320-L-2A_USA_CDC_pcl
D7320-L-2B_USA_CDC_pcl
NVLV_QA-157
NVLV_QA-157_R1.fasta.gz
NVLV_QA-157_R2.fasta.gz
NVLV_QA-158
NVLV_QA-158_R1.fasta.gz
NVLV_QA-158_R2.fasta.gz
NVLV_QA-159
NVLV_QA-159_R1.fasta.gz
NVLV_QA-159_R2.fasta.gz
NVLV_QA-160
NVLV_QA-160_R1.fasta.gz
NVLV_QA-160_R2.fasta.gz
NVLV_QA-161
NVLV_QA-161_R1.fasta.gz
NVLV_QA-161_R2.fasta.gz
quality_control (54)
REFERENCE DATA
No references have been added.
1 - 54 of 54
Items per page: 100


WORKSPACES
Workspaces › theiagen_pni/TheiaProk_PNI_Training_ETrees ›
Data
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
EDIT
OPEN WITH...
EXPORT
SETTINGS
0 rows selected
ADVANCED SEARCH
Search
quality_control_id
basespace_fetch_analysis_date
basespace_fetch_version
basespace_sample_name
D5480-LRM4update
2024-01-30
Terra Utilities v1.3.4
D5480-M3235-23-042
D7320-L-1A_USA_CDC_pcl
D7320-L-1B_USA_CDC_pcl
D7320-L-2A_USA_CDC_pcl
D7320-L-2B_USA_CDC_pcl
NVLV_QA-157
2024-02-13
PHB v1.3.0
QA-157
NVLV_QA-158
QA-158
NVLV_QA-159
QA-159
NVLV_QA-160
QA-160
NVLV_QA-161
QA-161
quality_control (54)
REFERENCE DATA
No references have been added.
1 - 54 of 54
Items per page: 100

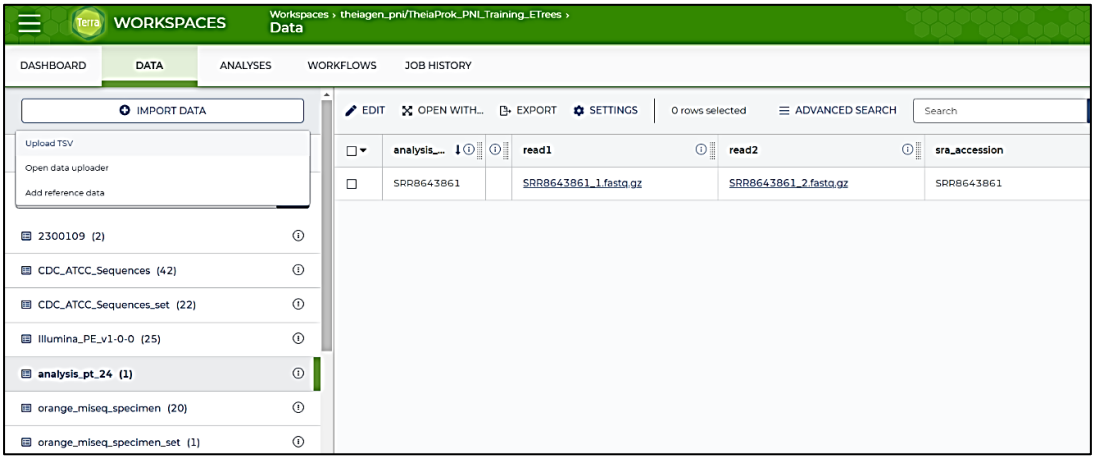
Appendix PNID01-2: Data Download from NCBI SRA

1. Prepare the metadata tsv file:
    a. Enter the name of the data table (either new or existing) into the cell A1 between "entity:" and "id".
    b. Enter the sample IDs into the column A the way you want them to appear in the Terra data table.
    c. sra_accession: enter the SRA accession numbers for the sequences to be downloaded from SRA.

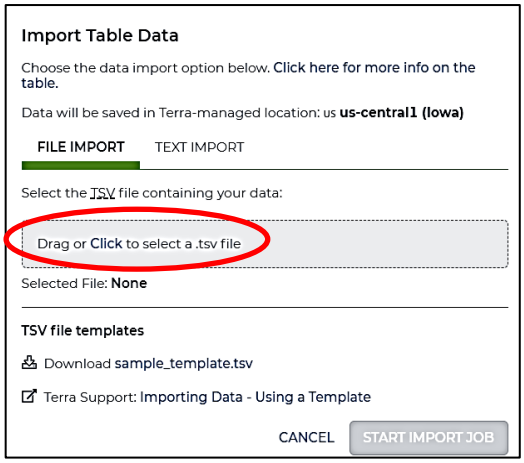


	A	B	C	D	E
1	entity:analysis_pt_24_id	sra_accession			
2	PNUSAS072225	SRR8878889			
3	PNUSAS070021	SRR8786860			
4	PNUSAS070585	SRR8756231			
5	PNUSAS177018	SRR12885281			
6	PNUSAS176867	SRR12884674			
7	PNUSAS188944	SRR13438657			
8	2020K-1237	SRR13132213			
9	PNUSAS188279	SRR13387034			
10	2020K-0901	SRR12712421			
11					

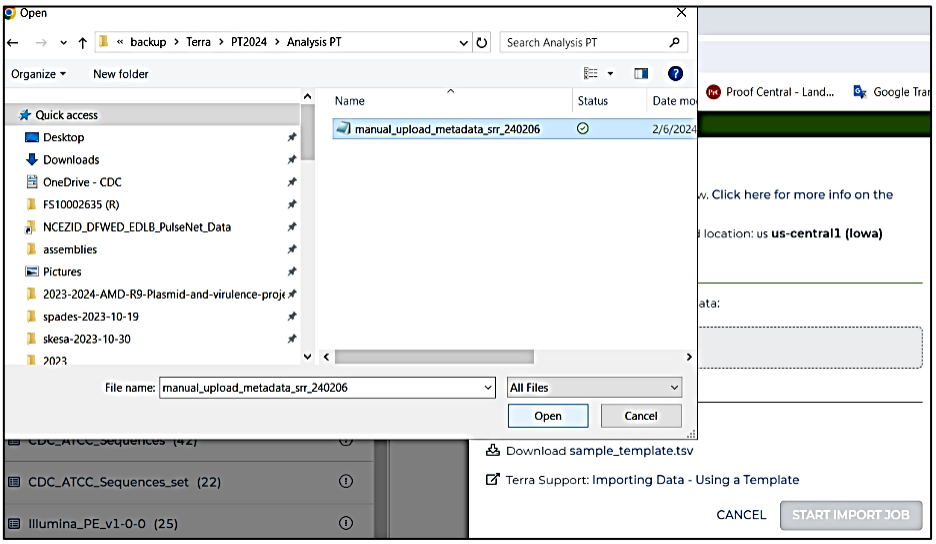
    d. Save the file in the **tsv format**: choose "Save As" and "Text (Tab delimited) (*.txt)". Make sure your file name has "**.tsv**" ending.
2. In the "Data" tab, click on "Import Data" and select "Upload tsv" from the drop-down menu.



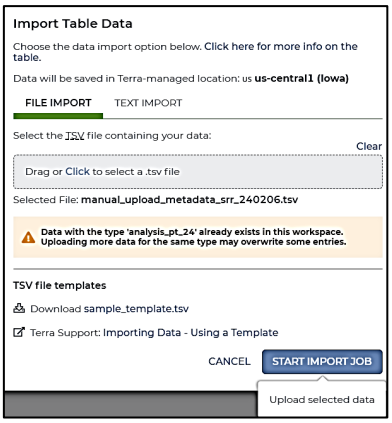
3. In the "Import Table Data" pop-up window "File Import" tab, click in the middle to select the tsv file.



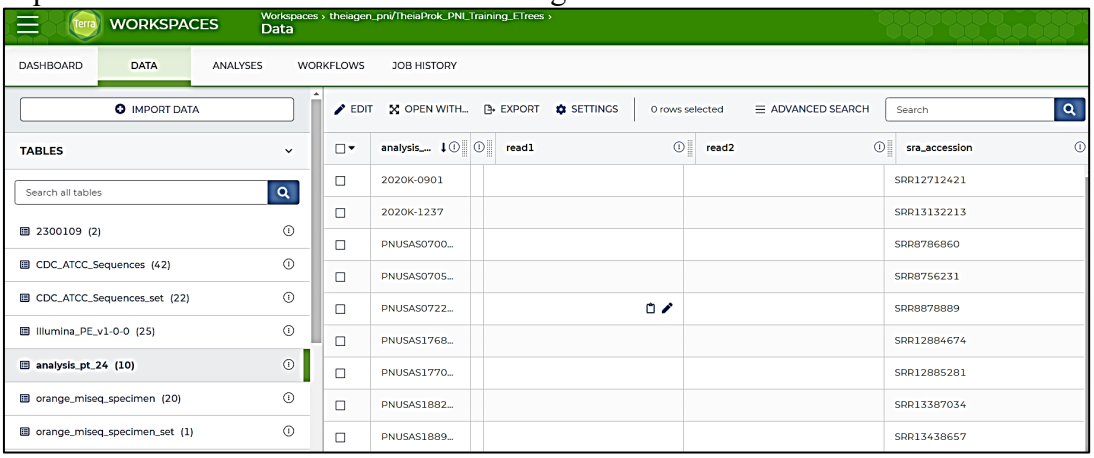
4. Navigate to the location where the metadata tsv file is saved, select the file and click "Open".



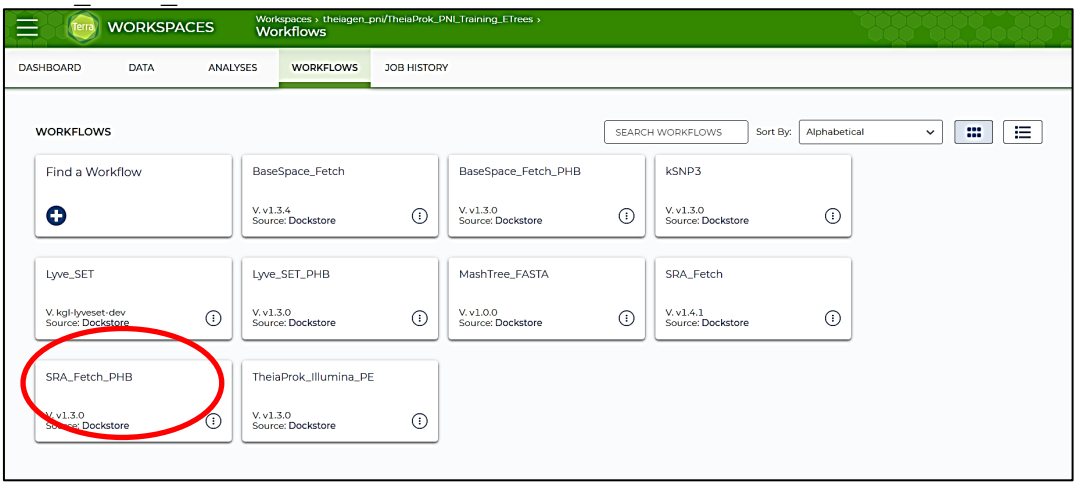
5. In the "Import Table Data" pop-up window, you will get a warning that data already exists in the data table in question (if importing to an existing data table) and uploading more data to it may overwrite existing data. Click on "Start Import Job".



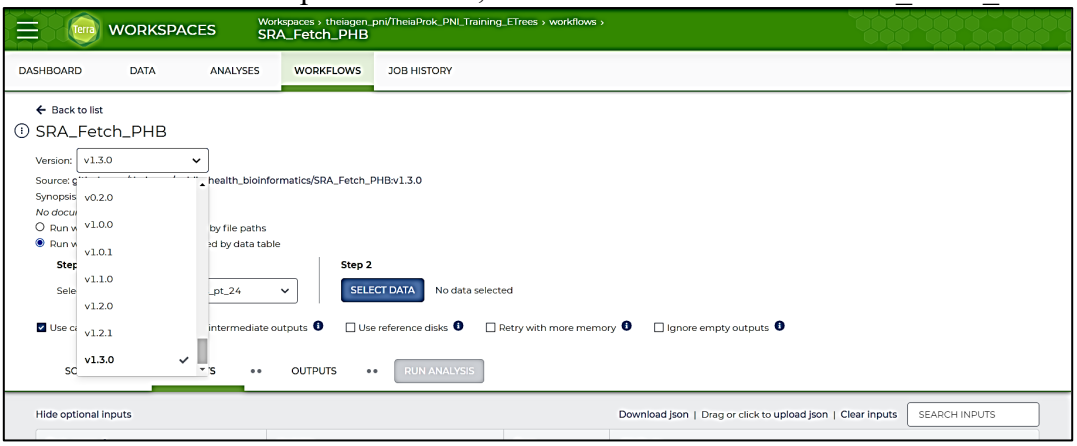
6. After the import is done you should see new entries created to the data table for the sequences to be downloaded from NCBI together with their SRA accession numbers



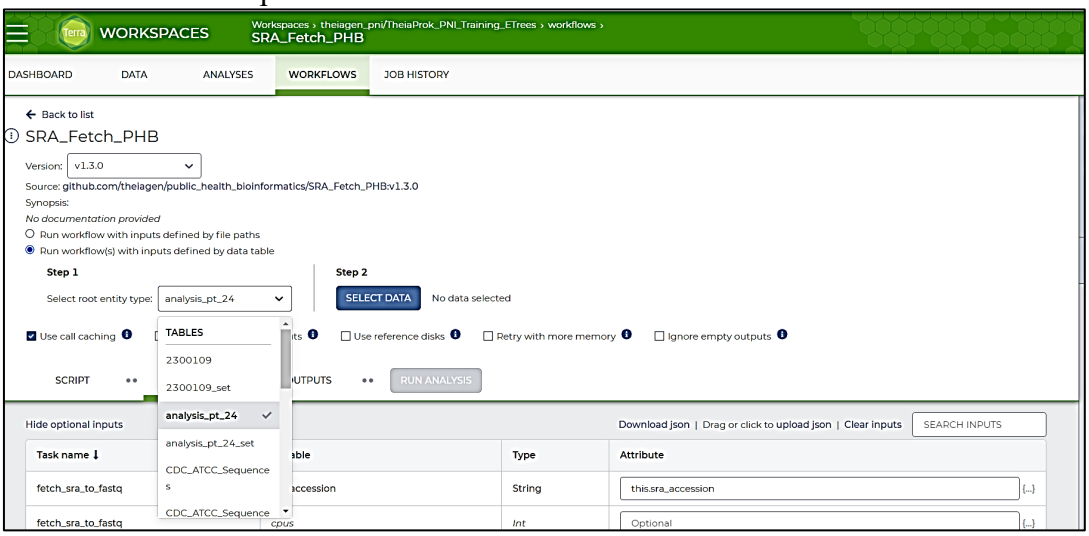
7. In the "Workflows" tab, click on the "SRA_Fetch_PHB" workflow. This will open the "SRA_Fetch_PHB" screen.



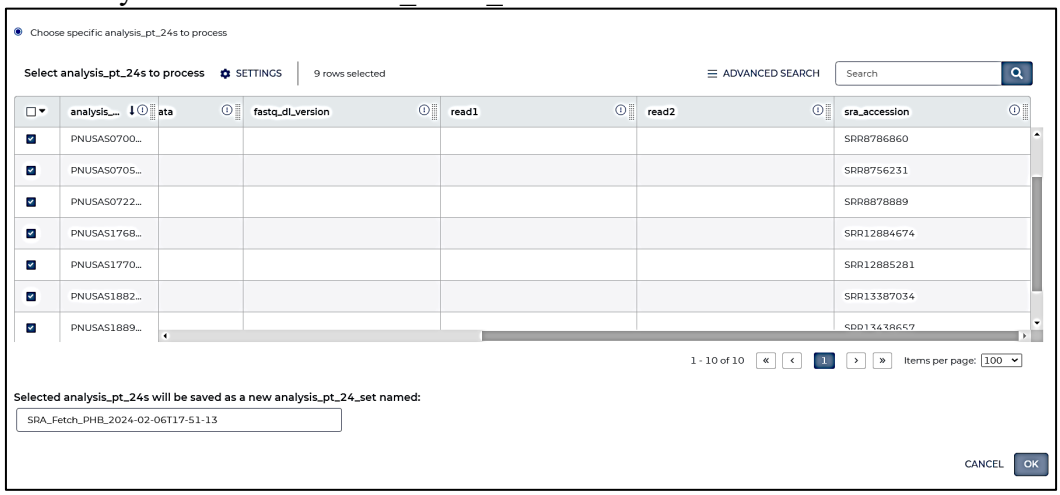
8. From the "Version" drop-down menu, select the latest version of SRA_Fetch_PHB.



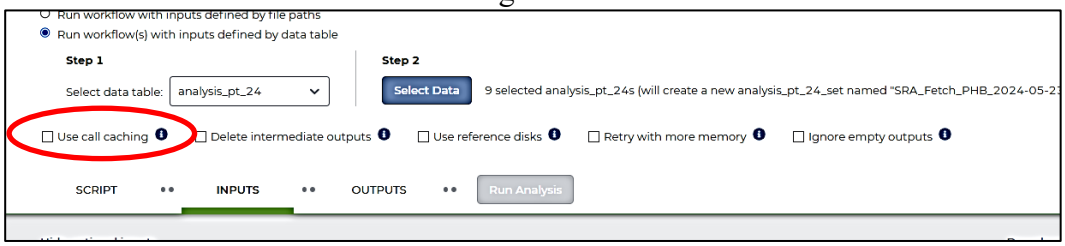
9. Under "Step 1", click on the "Select root entity type" drop-down menu and select the data table into which you imported (steps 1-6) the tsv file containing the SRA accession numbers for the samples to be downloaded from SRA.



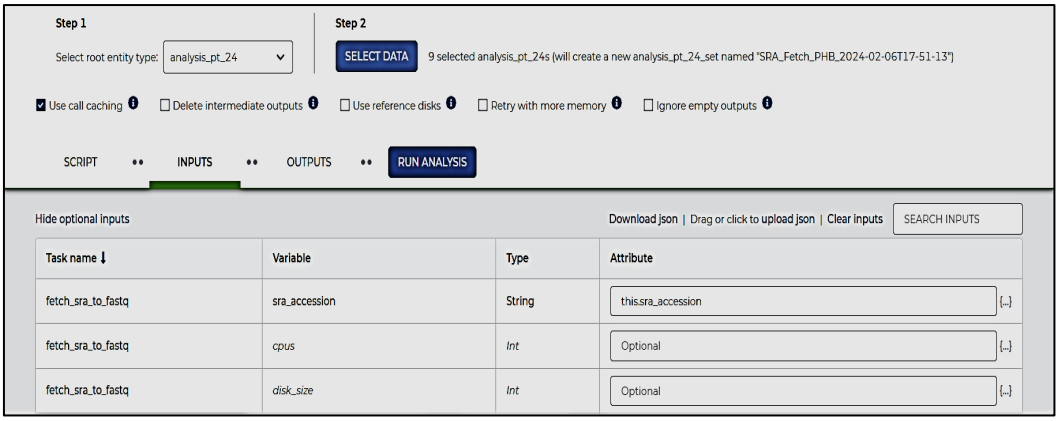
10. Under "Step 2", click on "Select data" (screenshot above). This will take you to the sample selection screen.
11. Check the boxes next to the samples to be downloaded from NCBI and click "OK". This will take you back to the "SRA_Fetch_PHB" screen.



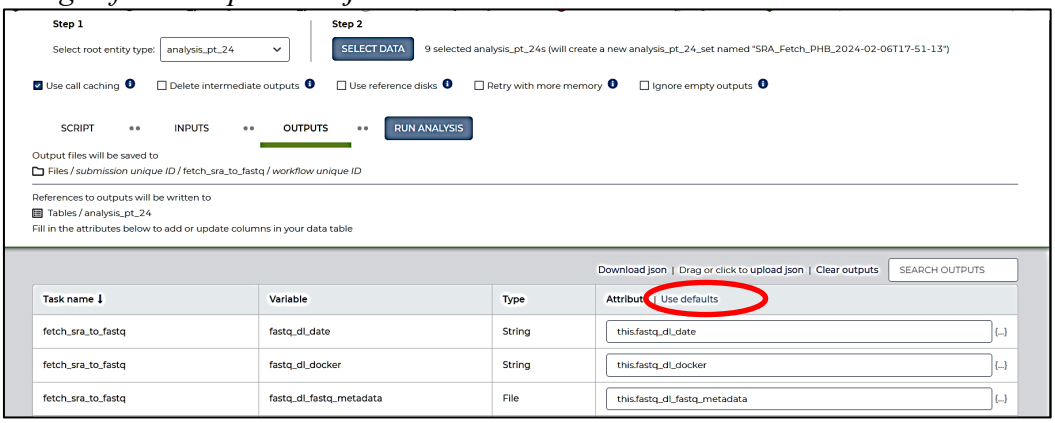
12. De-select the box for "Use call catching".



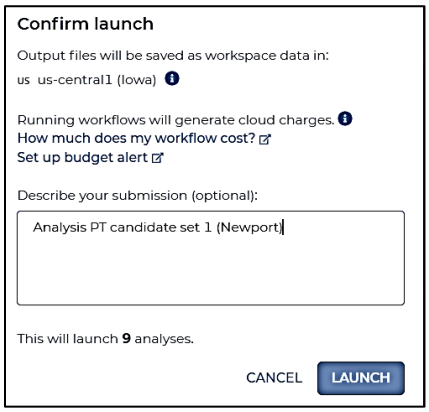
13. In the "**Inputs**" tab, define the "sra_accession" variable in the "Attribute" field: "this.sra_accession".



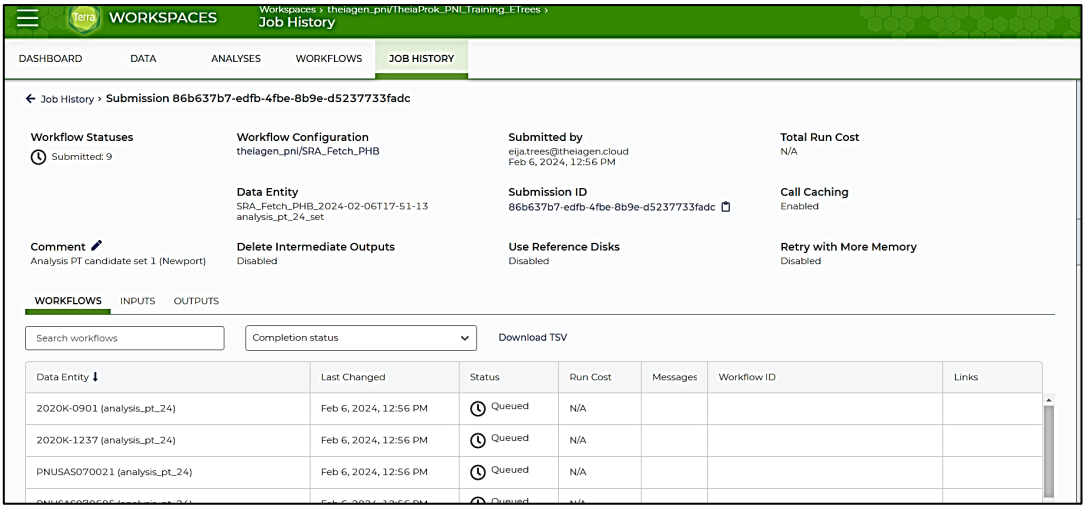
14. In the "**Outputs**" tab, click "Use defaults", then click "Save", and then "Run Analysis". **NOTE:** *The "Save" button is only visible if parameters (other than sample IDs) have changed from the previous job submission.*



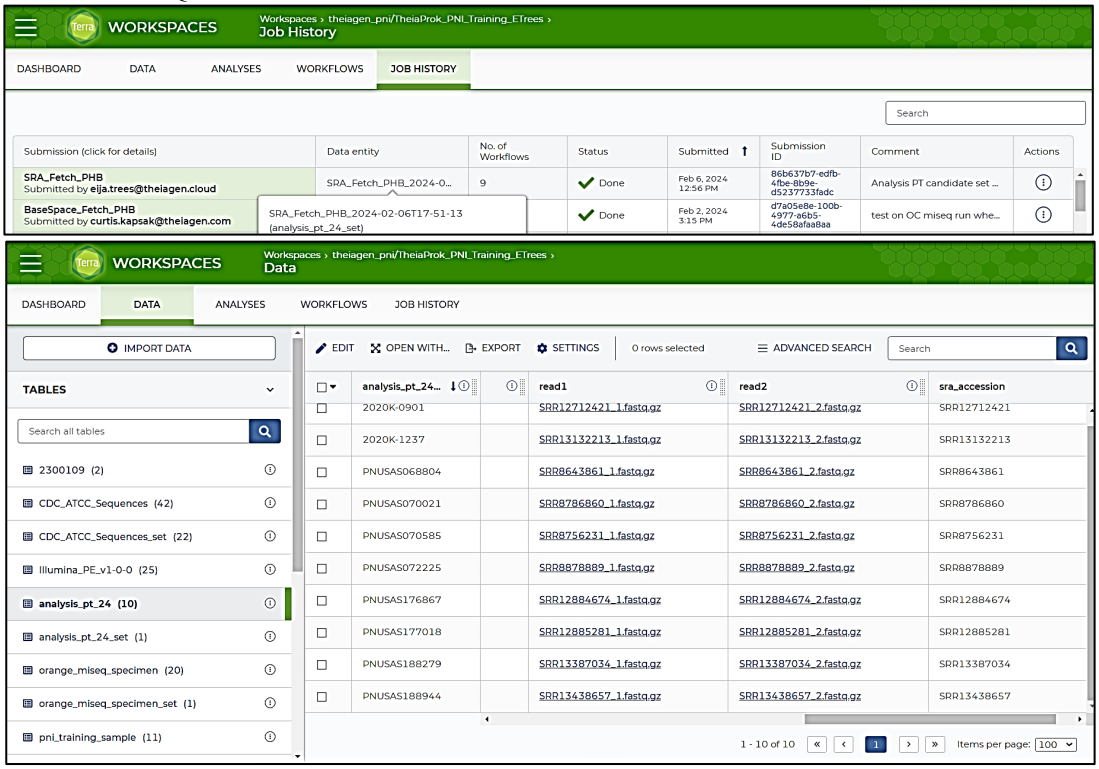
15. In the "Confirm launch" pop-up window, describe your job submission (optional) and click "Launch".



16. The "Job History" tab opens where the status of the job submissions should initially be "Queued".



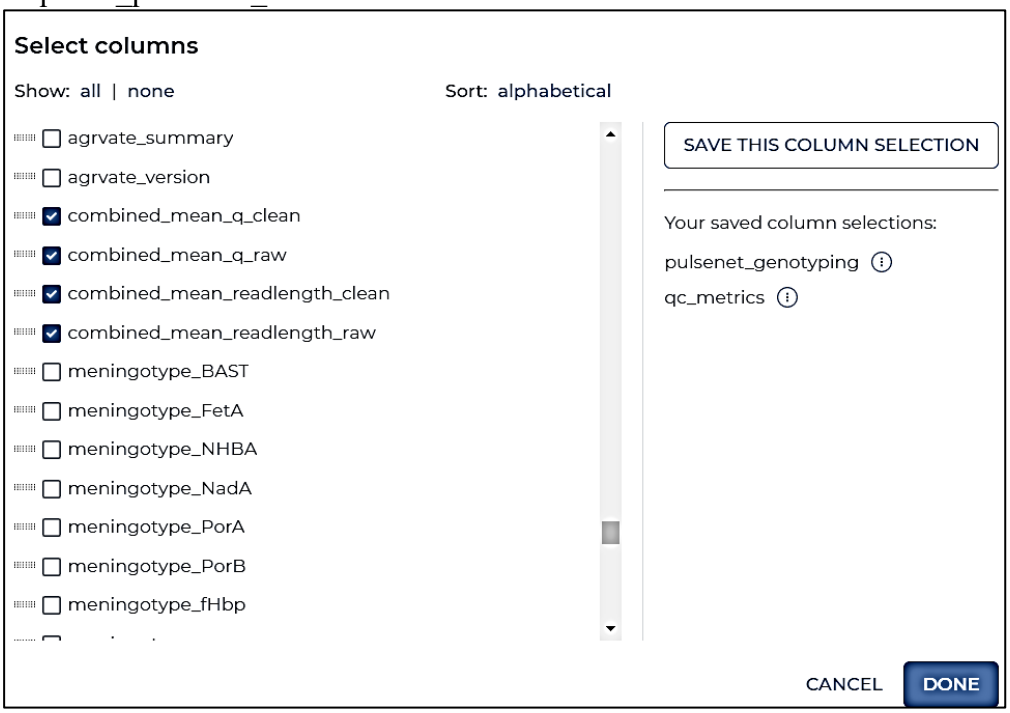
17. The status will be “Done” once the job has finished. In the “Data” tab, you should now see the FASTQ file names in “Read1” and “Read2” columns.



Appendix PNID01-3: Customization of a Data Table View for PulseNet QC Metrics

NOTE: *this customization can only be done after you have run the TheiaProk workflow once.*

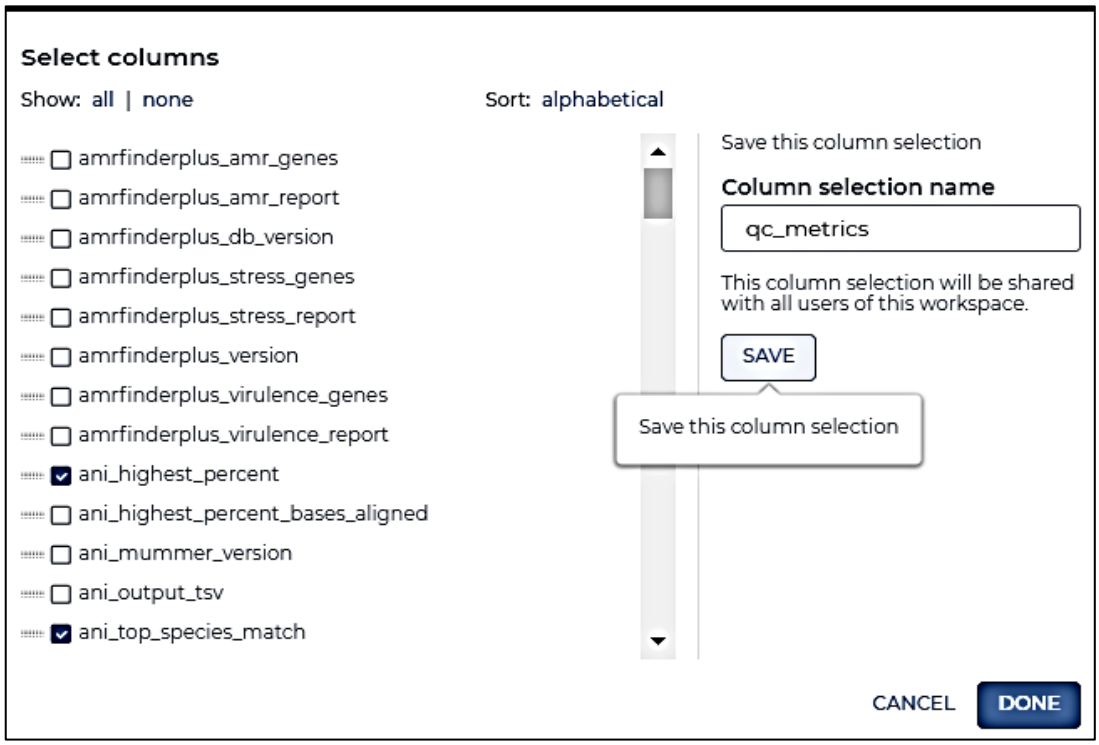
1. In the "Data" tab, select the data table of interest, e.g., "CDC_ATCC_Sequences", then select "Settings".
2. Under "Select columns" the following QC metrics should be checked:
    a. Ani_highest_percent
    b. Ani_top_species_match
    c. Assembly_length
    d. Combined_mean_q_clean
    e. Combined_mean_q_raw
    f. Combined_mean_readlength_clean
    g. Combined_mean_readlength_raw
    h. Est_coverage_clean
    i. Est_coverage_raw
    j. Gambit_predicted_taxon
    k. Midas_secondary_genus
    l. Midas_secondary_genus_abundance
    m. N50_value
    n. Number_contigs
    o. Raw_read_screen
    p. Seqsero2_predicted_contamination



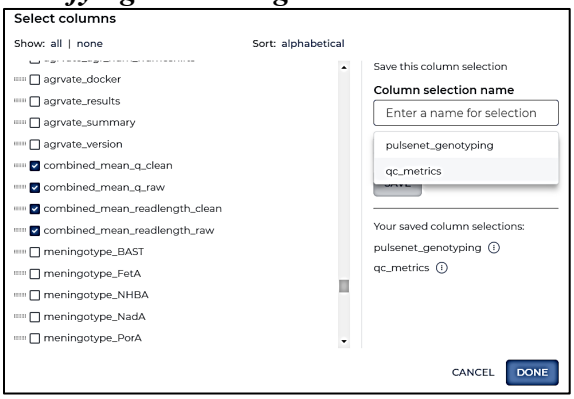
3. Click "Save this column selection", name the column selection "qc_metrics" and click "Save" and then click "Done".

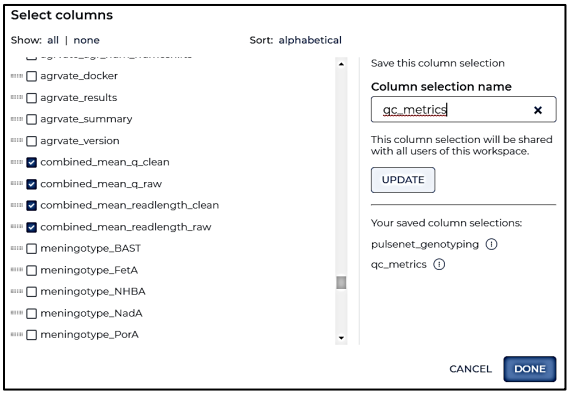
NOTE: *If you are adding or deleting columns from an existing column selection, click "Save this column selection", select the name from the drop-down menu and click "Update".*

Creating a new column selection



Modifying an existing column selection





Appendix PNID01-4a. PulseNet Critical (Pass/Fail) Quality Metrics for Routine Sequence Submissions

Organism	Average denovo coverage[1]	Average quality (Q score)[2]	Assembly length (MB)	Secondary species abundance
Listeria monocytogenes	≥ 20x	≥ 30	2.8-3.2	≤ 0.01
***E. coli* (most serotypes)**	≥ 40x	≥ 30	4.9-6.0	≤ 0.01
Shigella spp./*Rare *E. coli	≥ 40x	≥ 30	4.2-4.9	≤ 0.01
Salmonella spp.	≥ 30x	≥ 30	4.4-5.7	≤ 0.01
Campylobacter spp.	≥ 20x	≥ 30	1.4-2.2	≤ 0.01
Vibrio cholerae	≥ 40x	≥ 30	3.8-4.3	≤ 0.01
Vibrio parahaemolyticus	≥ 40x	≥ 30	4.9-5.5	≤ 0.01
Vibrio vulnificus	≥ 40x	≥ 30	4.7-5.3	≤ 0.01

[1]After quality-based trimming (est_coverage_clean)
[2] Before trimming (combined_mean_q_raw)

Appendix PNID01-4b. TheiaProk Read Pre-Screening Step to Exclude Poor Quality Sequences to Conserve Computational Resources

The screen task ensures the quantity of sequence data is sufficient to undertake genomic analysis. It uses bash commands for quantification of reads and base pairs, and mash sketching to estimate the genome size and its coverage. At each step, the results are assessed relative to pass/fail criteria and thresholds that may be defined by optional user inputs.

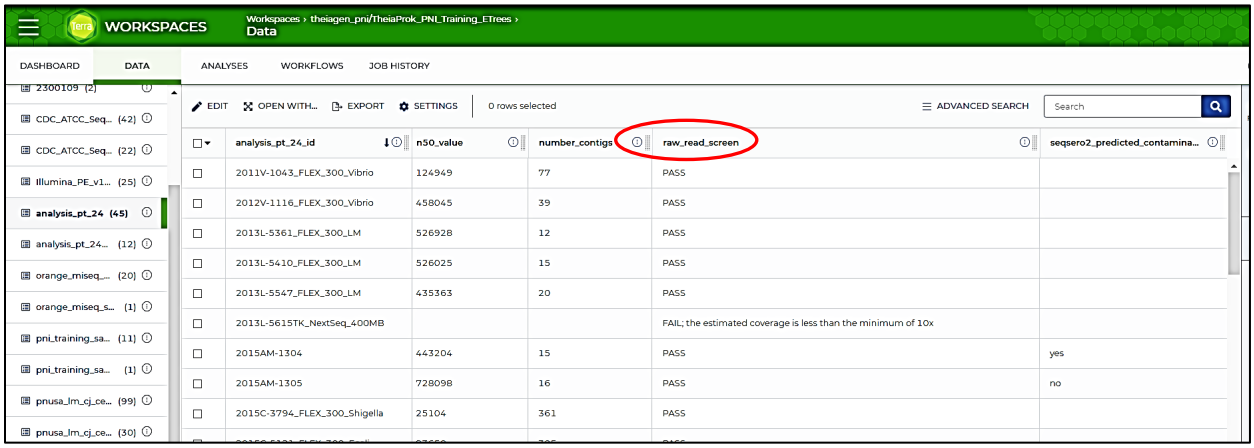
Samples that do not meet these criteria will not be processed further by the workflow:

1. Total number of reads: A sample will fail the read screening task if its total number of reads is less than or equal to min_reads.
2. The proportion of basepairs reads in the forward and reverse read files: A sample will fail the read screening if fewer than min_proportion basepairs are in either the reads1 or read2 files.
3. Number of basepairs: A sample will fail the read screening if there are fewer than min_basepairs basepairs
4. Estimated genome size: A sample will fail the read screening if the estimated genome size is smaller than min_genome_size or bigger than max_genome_size.
5. Estimated genome coverage: A sample will fail the read screening if the estimated genome coverage is less than the min_coverage.

Default values:

Int min_reads = 7472
Int min_basepairs = 2241820
Int min_genome_length = 100000
Int max_genome_length = 18040666
Int min_coverage = 10
Int min_proportion = 40

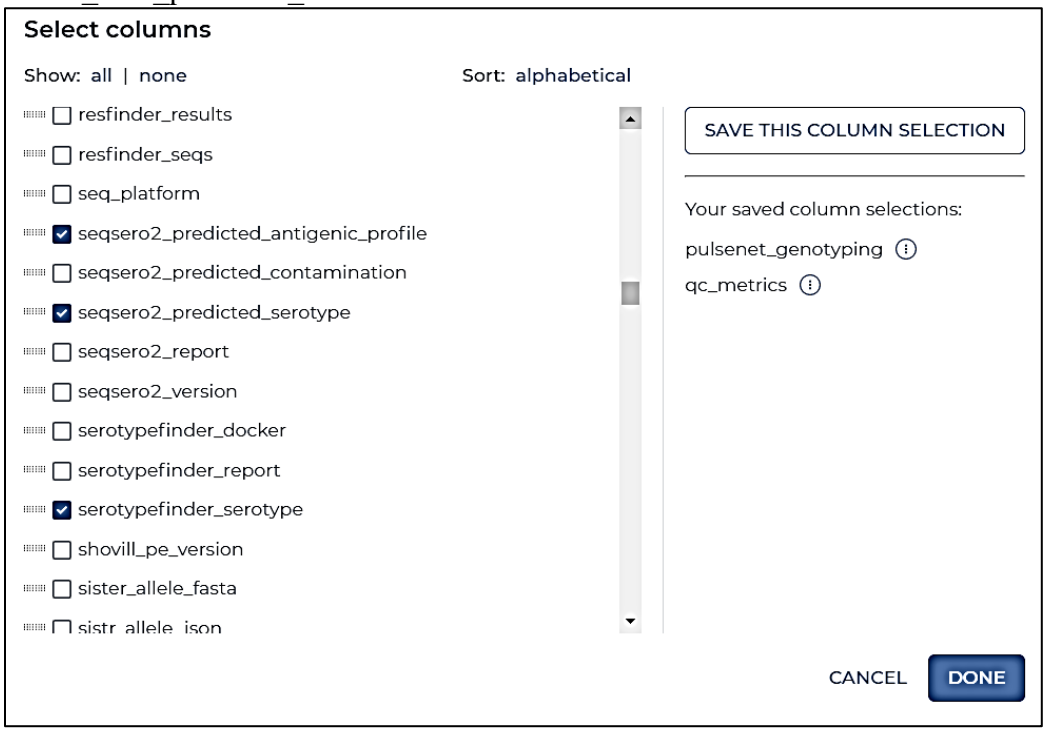
The “raw_read_screen” column under “QC_metrics” will give details if the sample fails the read pre-screen:



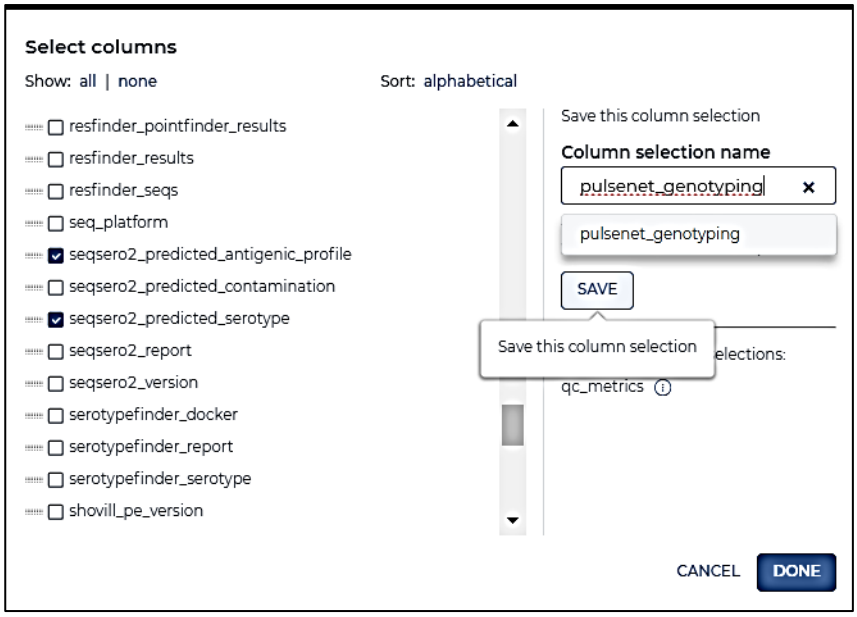
Appendix PNID01-5. Customization of a Data Table View for PulseNet Genotyping Assays

NOTE: *this customization can only be done after you have run the TheiaProk workflow once.*

1. In the "Data" tab, select the data table of interest, e.g., "CDC_ATCC_Sequences", then select "Settings".
2. Under "Select columns" the following genotyping assays should be checked:
    a. Amrfinderplus_amr_classes
    b. Amrfinderplus_amr_core_genes
    c. Amrfinderplus_amr_subclasses
    d. Amrfinderplus_virulence_genes
    e. Plasmidfinder_plasmids
    f. Seqsero2_predicted_antigenic profile
    g. Seqsero2_predicted_serotype
    h. Serotypefinder_serotype
    i. Ts_mlst_predicted_st

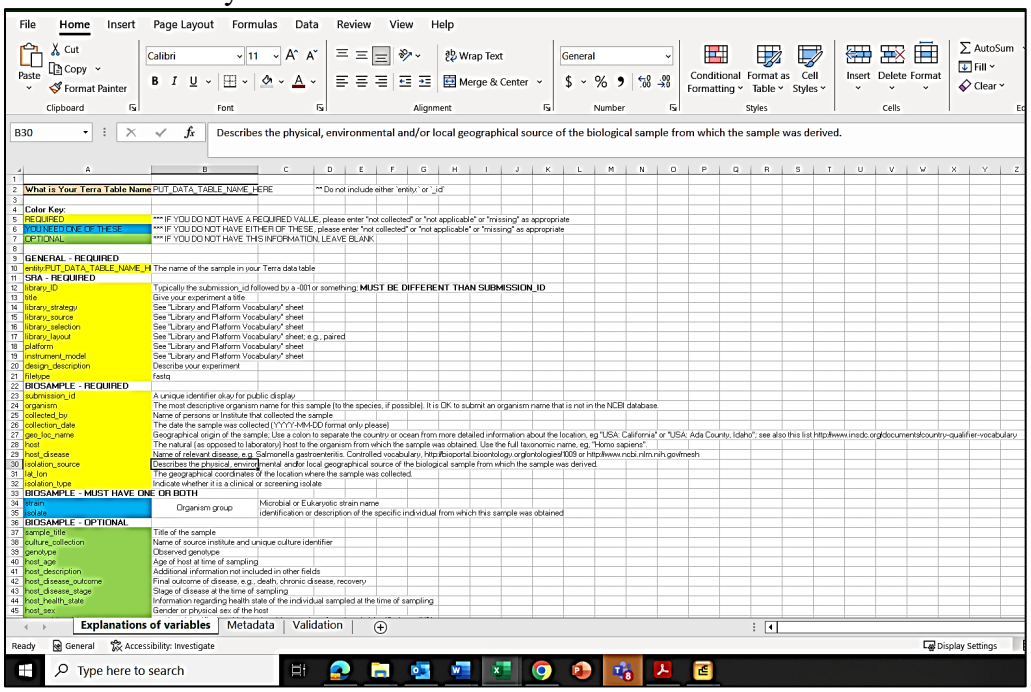


3. Click "Save this column selection", name the column selection "pulsenet_genotyping" and click "Save" and then click "Done".
   NOTE: *If you are adding or deleting columns from an existing column selection, click "Save this column selection", select the name from the drop-down menu and click "Update".*


Select columns
Show: all | none
Sort: alphabetical
resfinder_pointfinder_results
resfinder_results
resfinder_seqs
seq_platform
seqsero2_predicted_antigenic_profile
seqsero2_predicted_contamination
seqsero2_predicted_serotype
seqsero2_report
seqsero2_version
serotypefinder_docker
serotypefinder_report
serotypefinder_serotype
shovill_pe_version
Save this column selection
Column selection name
pulsenet_genotyping
pulsenet_genotyping
SAVE
Save this column selection
elections:
qc_metrics
CANCEL
DONE

Appendix PNID01-6. Uploading Additional Metadata to Terra for NCBI Submissions and Customization of a Data Table View for Metadata

- NCBI requires minimal metadata to be uploaded to NCBI to create BioSamples for the sequences to be uploaded to SRA.
- Because the NCBI metadata needs to be formatted in a certain way, please use the **Pathogen** metadata template spreadsheet provided by Theiagen to upload metadata to Terra: https://theiagen.notion.site/Terra_2_NCBI-8f014c73acc44465a3d69cf4df93adfe.
- The metadata template has three tabs:
    - The first tab, called "Explanation of variables", contains descriptions for the required and optional fields.
    - The metadata to be uploaded is entered in the second "Metadata" tab.
    - The third "Validation" tab can used to ascertain that the required metadata is filled out correctly.

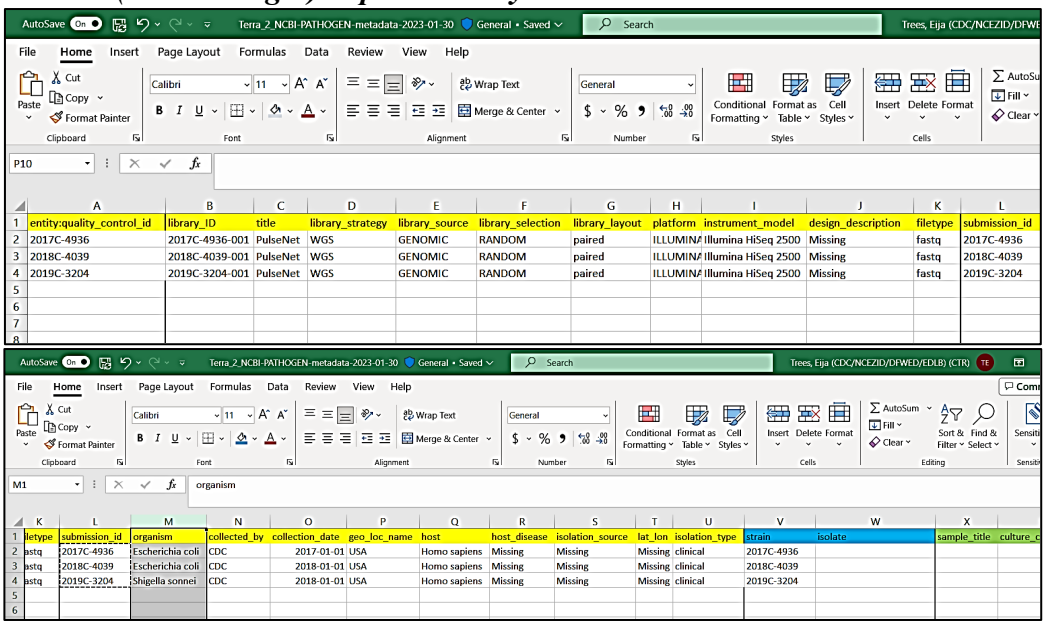


To upload the Pathogen metadata file to Terra:

1. Fill out the required and optional (if applicable) fields in the metadata tab of the Pathogen metadata template spreadsheet:
   NOTE1: *you can enter "Missing" for any required information you don't have or you don't wish to publicly disclose.*
   NOTE2: *The metadata below are the PulseNet USA minimum metadata requirements for NCBI uploads designed to protect patient privacy and the integrity of on-going outbreak investigations. On the other hand, enough epidemiologically useful information is provided particularly on non-clinical samples to facilitate attribution.*

a. In cell A1, enter the data table name into which metadata will be uploaded: **entity**:*data_table_name*_**id**, e.g., entity:quality_control_id.
b. Entity: enter the sample IDs that match the IDs for the sequences in Terra.
c. Submission_id: enter a unique ID for the sample. This is the ID that will be displayed on NCBI.
    i. Enter the submission_id also in the "strain" column.
    ii. Enter the "submission_id-001" to the "library_id" column.
d. Title: "PulseNet".
e. For "library_strategy", "library_source", "library_selection", "platform", "instrument_model" and "filetype", pick the correct option from the drop-down menus.
f. Library layout: "Paired".
g. Organism: genus and species.
h. Collected_by: the laboratory submitting the sequence.
i. Collection_date: **year for clinical** isolates, **year and month for non-clinical** isolates. Required format: YYYY:MM:DD, e.g., for a clinical sample isolated in 2024: 2024-01-01. For non-clinical sample isolated in February 2024: 2024-02-01.
j. Geo_loc_name: **source country for clinical** isolates, **source country and state (or other more detailed location) for non-clinical** isolates. Use a colon to separate the country and a more detailed location, e.g., USA:CA.
k. Isolation_source: **"Missing" for clinical** isolates, the **exact source for non-clinical** isolates, e.g., lettuce, chicken breast, swab, etc.
l. Isolation_type: clinical, environmental, food or animal.
m. Serotype: *E. coli* serotype.
n. Serovar: *Salmonella* serovar.

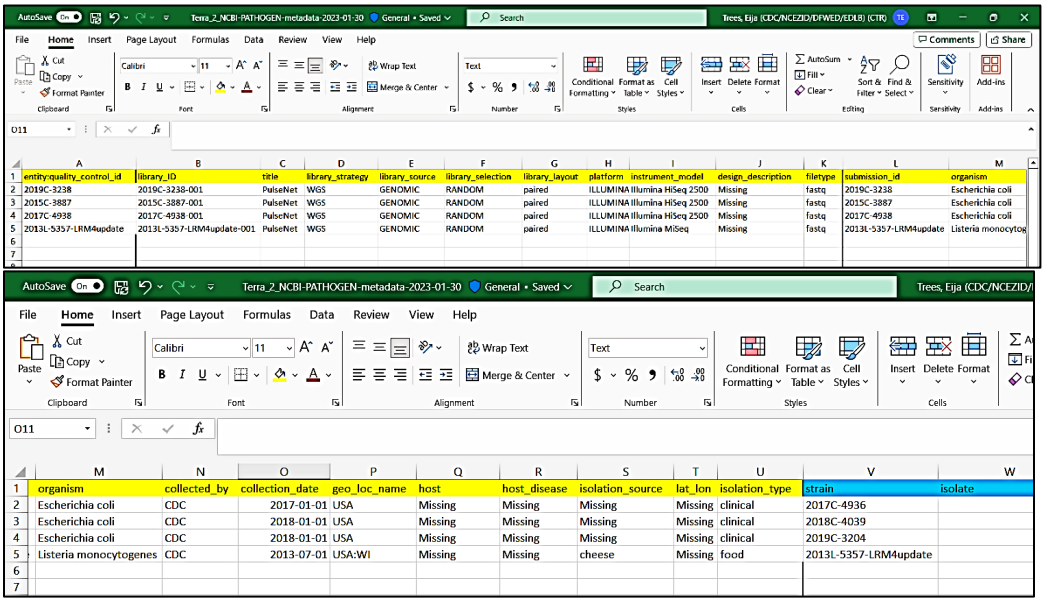
Clinical (human origin) sequences only



	A	B	C	D	E	F	G	H	I	J	K	L
1	entity:quality_control_id	library_ID	title	library_strategy	library_source	library_selection	library_layout	platform	instrument_model	design_description	filetype	submission_id
2	2017C-4936	2017C-4936-001	PulseNet	WGS	GENOMIC	RANDOM	paired	ILLUMINA	Illumina HiSeq 2500	Missing	fastq	2017C-4936
3	2018C-4039	2018C-4039-001	PulseNet	WGS	GENOMIC	RANDOM	paired	ILLUMINA	Illumina HiSeq 2500	Missing	fastq	2018C-4039
4	2019C-3204	2019C-3204-001	PulseNet	WGS	GENOMIC	RANDOM	paired	ILLUMINA	Illumina HiSeq 2500	Missing	fastq	2019C-3204

	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	iletype	submission_id	organism	collected_by	collection_date	geo_loc_name	host	host_disease	isolation_source	lat_lon	isolation_type	strain	isolate	sample_title
2	astq	2017C-4936	Escherichia coli	CDC	2017-01-01	USA	Homo sapiens	Missing	Missing	Missing	clinical	2017C-4936		
3	astq	2018C-4039	Escherichia coli	CDC	2018-01-01	USA	Homo sapiens	Missing	Missing	Missing	clinical	2018C-4039		
4	astq	2019C-3204	Shigella sonnei	CDC	2018-01-01	USA	Homo sapiens	Missing	Missing	Missing	clinical	2019C-3204		

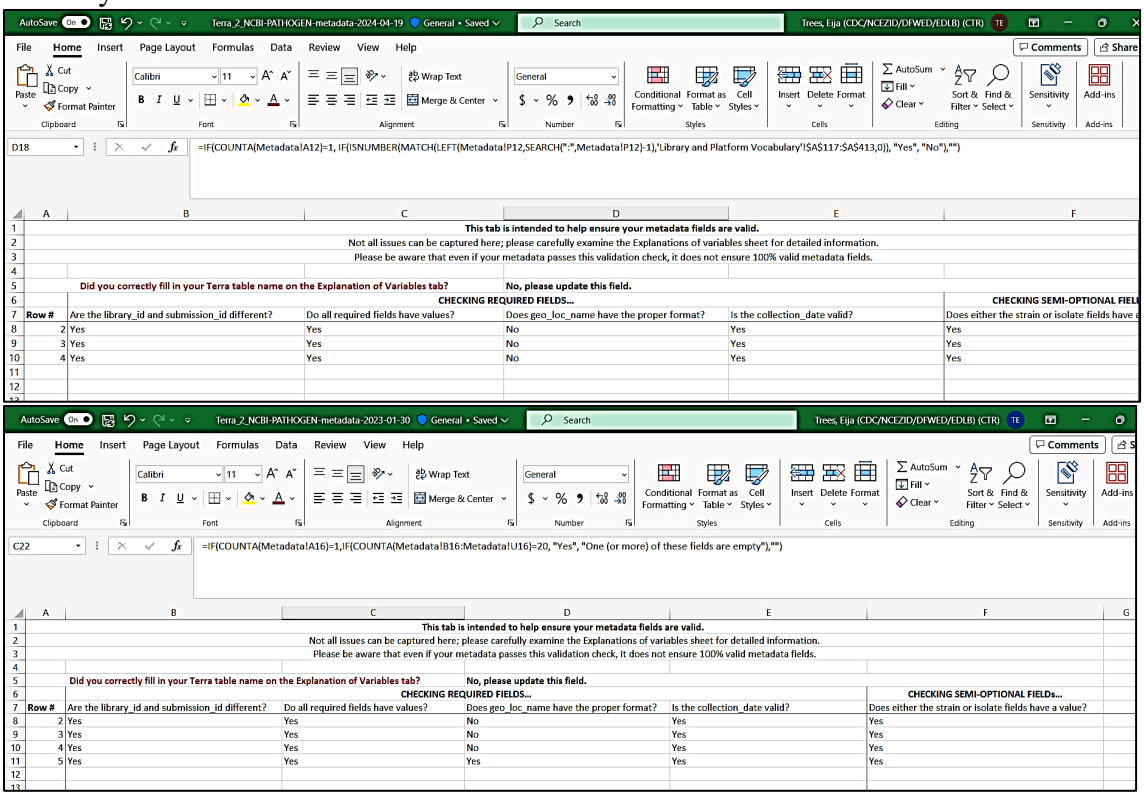
Clinical (human origin) and non-clinical sequences



entity:quality_control_id	library_ID	title	library_strategy	library_source	library_selection	library_layout	platform	instrument_model	design_description	filetype	submission_id	organism
2019C-3238	2019C-3238-001	PulseNet	WGS	GENOMIC	RANDOM	paired	ILLUMINA	Illumina HiSeq 2500	Missing	fastq	2019C-3238	Escherichia coli
2015C-3887	2015C-3887-001	PulseNet	WGS	GENOMIC	RANDOM	paired	ILLUMINA	Illumina HiSeq 2500	Missing	fastq	2015C-3887	Escherichia coli
2017C-4938	2017C-4938-001	PulseNet	WGS	GENOMIC	RANDOM	paired	ILLUMINA	Illumina HiSeq 2500	Missing	fastq	2017C-4938	Escherichia coli
2013L-5357-LRM4update	2013L-5357-LRM4update-001	PulseNet	WGS	GENOMIC	RANDOM	paired	ILLUMINA	Illumina MiSeq	Missing	fastq	2013L-5357-LRM4update	Listeria monocytog

organism	collected_by	collection_date	geo_loc_name	host	host_disease	isolation_source	lat_lon	isolation_type	strain	isolate
Escherichia coli	CDC	2017-01-01	USA	Missing	Missing	Missing	Missing	clinical	2017C-4936	
Escherichia coli	CDC	2018-01-01	USA	Missing	Missing	Missing	Missing	clinical	2018C-4039	
Escherichia coli	CDC	2018-01-01	USA	Missing	Missing	Missing	Missing	clinical	2019C-3204	
Listeria monocytogenes	CDC	2013-07-01	USA:WI	Missing	Missing	cheese	Missing	food	2013L-5357-LRM4update	

2. Switch to the "Validation" tab to confirm that the metadata template has been filled out correctly.



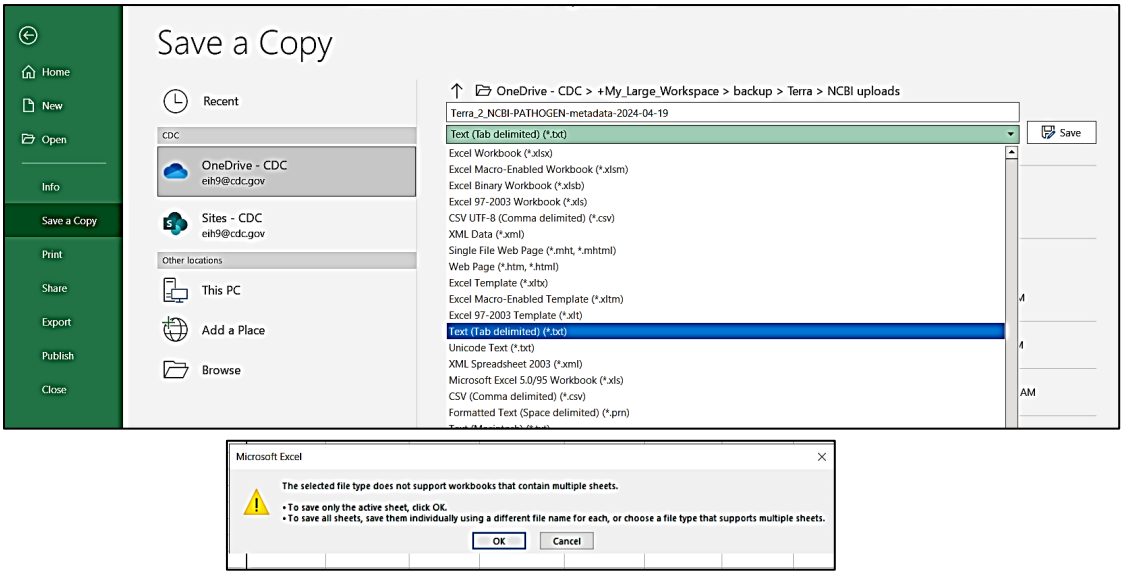
This tab is intended to help ensure your metadata fields are valid.				
Not all issues can be captured here; please carefully examine the Explanations of variables sheet for detailed information.				
Please be aware that even if your metadata passes this validation check, it does not ensure 100% valid metadata fields.				
Did you correctly fill in your Terra table name on the Explanation of Variables tab?	**No, please update this field.**			
	CHECKING REQUIRED FIELDS...			**CHECKING SEMI-OPTIONAL FIEL**
Row #	Are the library_id and submission_id different?	Do all required fields have values?	Does geo_loc_name have the proper format?	Is the collection_date valid?	Does either the strain or isolate fields have a
2	Yes	Yes	No	Yes	Yes
3	Yes	Yes	No	Yes	Yes
4	Yes	Yes	No	Yes	Yes

This tab is intended to help ensure your metadata fields are valid.					
Not all issues can be captured here; please carefully examine the Explanations of variables sheet for detailed information.					
Please be aware that even if your metadata passes this validation check, it does not ensure 100% valid metadata fields.					
Did you correctly fill in your Terra table name on the Explanation of Variables tab?	**No, please update this field.**				
	CHECKING REQUIRED FIELDS...				**CHECKING SEMI-OPTIONAL FIELDS...**
Row #	Are the library_id and submission_id different?	Do all required fields have values?	Does geo_loc_name have the proper format?	Is the collection_date valid?	Does either the strain or isolate fields have a value?
2	Yes	Yes	No	Yes	Yes
3	Yes	Yes	No	Yes	Yes
4	Yes	Yes	No	Yes	Yes
5	Yes	Yes	Yes	Yes	Yes

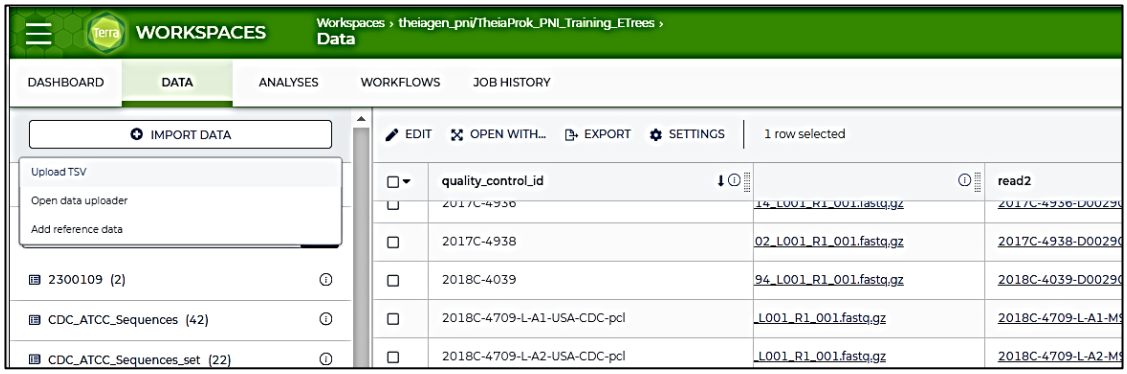
NOTE: *the validation of the proper format of the geological location will only pass if the source country and a more detailed location are both specified. Therefore, validation will*

not pass for clinical isolates for which only the source country is specified. However, NCBI will accept source country as the only geological location.

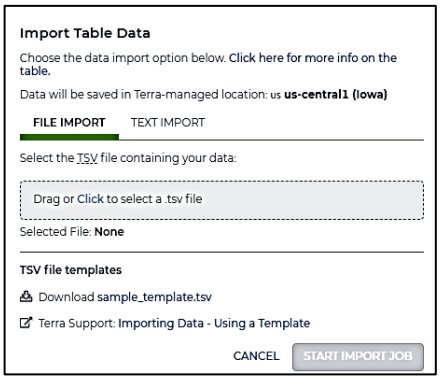
3. Save the filled-out metadata template as a tsv (tab delimited) file. Click "OK" on the pop-up window stating that the selected file type does not support workbooks with multiple sheets.



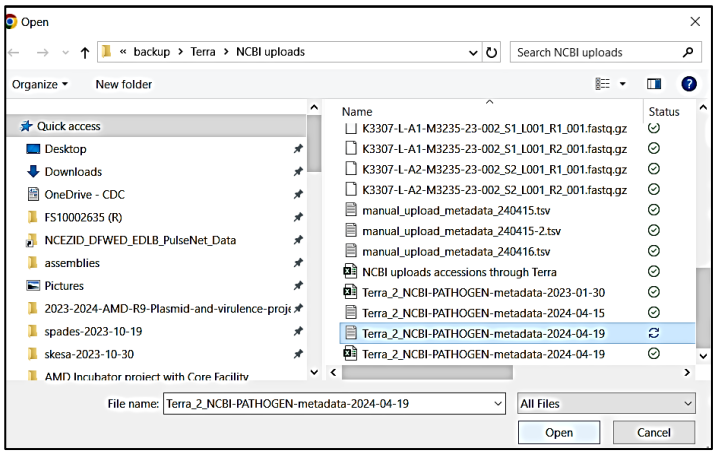
4. In the Terra Workspaces "Data" tab, click on "Import Data" and select "Upload TSV" from the drop-down menu.



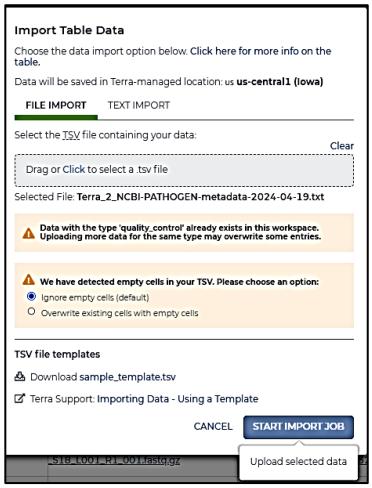
5. In the "Import Table Data" pop-up window "File import" tab, click in the middle to select the tsv file.



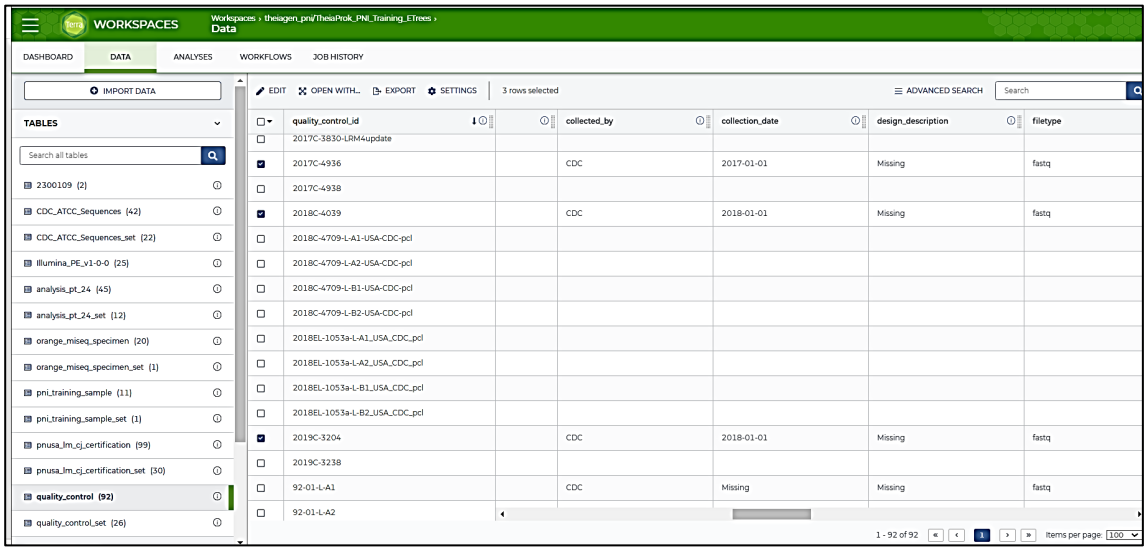
6. Navigate to the location where the metadata tsv file is saved, select the file and click "Open".



7. In the "Import Table Data" pop-up window, you will get a warning that data already exists in the data table in question and uploading more data to it may overwrite existing data. Also, there will be a warning if the metadata tsv file does not have the same information (some data missing for some sequences) for all entries. Click on "Start Import Job".



8. After the upload is done you should see the desired metadata fields populated in the data table.



Create a separate metadata view for the data table

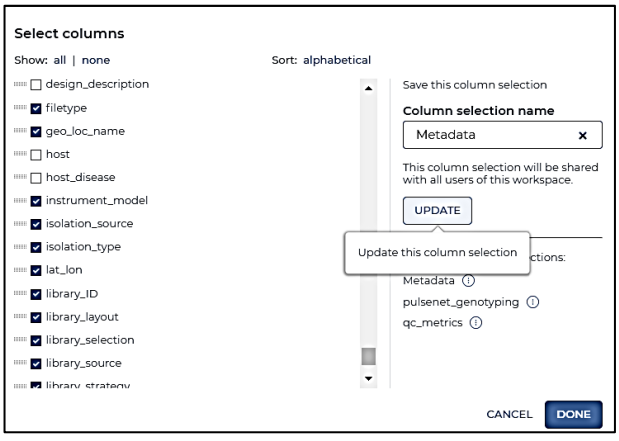
1. In the "Data" tab, select the data table of interest, then select "Settings".
2. Under "Select columns", check all the desired metadata columns.
    a. Recommended metadata for NCBI submissions:
        i. collected_by
        ii. collection _date
        iii. filetype
        iv. geo_loc_name
        v. instrument_model
        vi. isolation_source
        vii. isolation_type
        viii. library_id
        ix. library_layout
        x. library_selection
        xi. library_source
        xii. library_strategy
        xiii. organism
        xiv. platform
        xv. strain
        xvi. submission_id
        xvii. title
        xviii. serotype
        xix. serovar
    b. Additional useful information about the sequence
        i. read1 (R1 FASTQ file name)
        ii. read2 (R2 FASTQ file name)
        iii. assembly_fasta (location for the assembly generated by Terra)

iv. If uploading data directly from the Illumina BaseSpace:
    1. basespace_collection_id
    2. basespace_fetch_analysis_date
    3. basespace_fetch_version
    4. basespace_sample_id
    5. basespace_sample_name

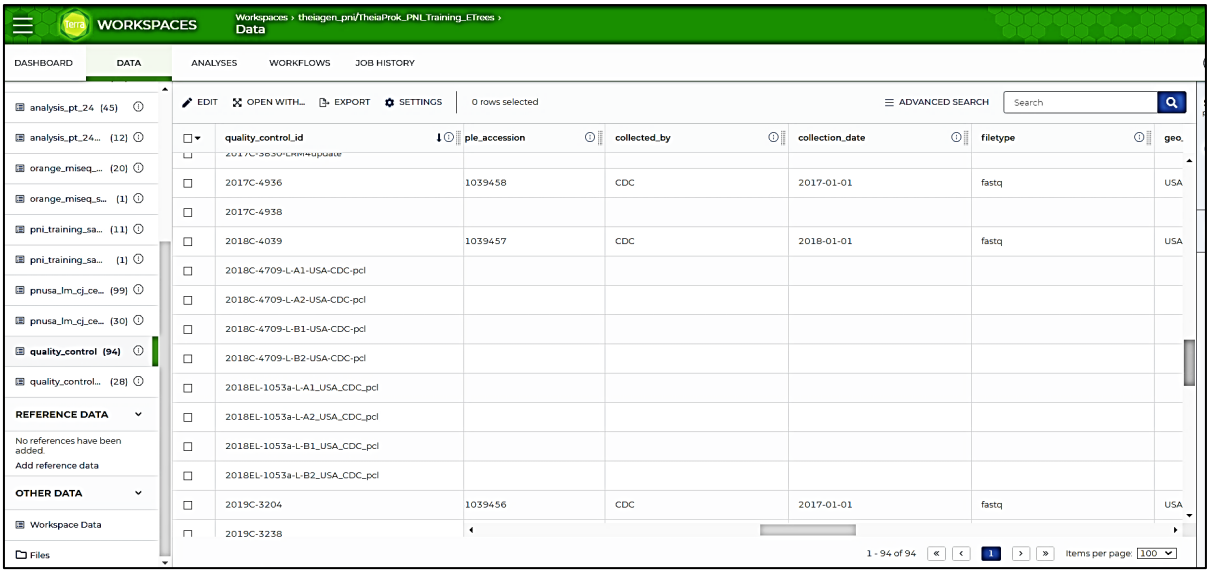
v. biosample_accession

vi. sra_accession

3. Click on “Save this column selection”.
4. Name the column selection “Metadata” and click on “Save” and “Done”.
   NOTE: *If you are adding or deleting columns from an existing column selection, click “Save this column selection”, select the name from the drop-down menu and click “Update”.*



5. Only the desired metadata columns should now be visible in the data table


WORKSPACES
Workspaces › theiagen_pni/TheiaProk_PNI_Training_ETrees ›
Data
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
EDIT
OPEN WITH...
EXPORT
SETTINGS
0 rows selected
ADVANCED SEARCH
Search
analysis_pt_24 (45)
analysis_pt_24... (12)
orange_miseq_... (20)
orange_miseq_s... (1)
pni_training_sa... (11)
pni_training_sa... (1)
pnusa_lm_cj_ce... (99)
pnusa_lm_cj_ce... (30)
quality_control (94)
quality_control... (28)
REFERENCE DATA
No references have been added.
Add reference data
OTHER DATA
Workspace Data
Files
quality_control_id
ple_accession
collected_by
collection_date
filetype
geo.
2017C-4936 1039458 CDC 2017-01-01 fastq USA
2017C-4938
2018C-4039 1039457 CDC 2018-01-01 fastq USA
2018C-4709-L-A1-USA-CDC-pcl
2018C-4709-L-A2-USA-CDC-pcl
2018C-4709-L-B1-USA-CDC-pcl
2018C-4709-L-B2-USA-CDC-pcl
2018EL-1053a-L-A1_USA_CDC_pcl
2018EL-1053a-L-A2_USA_CDC_pcl
2018EL-1053a-L-B1_USA_CDC_pcl
2018EL-1053a-L-B2_USA_CDC_pcl
2019C-3204 1039456 CDC 2017-01-01 fastq USA
2019C-3238
1 - 94 of 94
Items per page: 100

Appendix PNID01-7: Overview of the TheiaProk Workflow for Bacterial Characterization