Submitting Sequence Data to NCBI using Theiagen's Terra_2_NCBI_PHB Workflow	
Document TG-NCBI-01, Version 1	
Date:	Written for Workflow Version:
2/17/2026	V4.0.0

1. PURPOSE/SCOPE

This SOP outlines the steps for configuring the Terra_2_NCBI_PHB workflow for streamlined submission of sequencing data to NCBI and launching test submissions. This procedure ensures users understand the workflow requirements and how to launch submissions.

The Terra_2_NCBI_PHB workflow will submit sample metadata to the **BioSample database** and sequencing reads and associated metadata to the **Sequencing Read Archive (SRA)**. The Terra_2_NCBI_PHB workflow cannot submit genome assemblies to the GenBank or GISAID databases.

2. REQUIRED RESOURCES

IMPORTANT NOTE: The Terra_2_NCBI_PHB workflow requires several prerequisite credentials and resources. We strongly recommend addressing these prerequisites as described in the Appendices **prior** to attempting production workflow submissions.

- Computer
- Internet connection: at least 10 and 5Mbps for download and upload speeds, respectively
- Internet browser (Google Chrome, Firefox, or Edge)
- Google account
- Terra account, linked to Google account
- Illumina PE raw sequencing read files uploaded to Terra workspace, see TG-TER-03. **Read files must be free of human sequencing data.**
- Access to the NCBI File Transfer Protocol server used for uploading sequence data to NCBI and configuration file providing access credentials, see *Appendix 10.1*
- An accession number for the BioProject to which the data will be submitted
- Required sample and sequencing metadata for submission to NCBI, see *Appendix 10.2*. **Metadata must be free of PII.**
- A public Google bucket accessible by NCBI into which reads can be transferred from Terra, see *Appendix 10.3*
- Completion of a prior successful test submission to NCBI using the Terra_2_NCBI_PHB workflow, see *Appendix 10.4*

3. RELATED DOCUMENTS

Document Number	Document Name
TG-TER-03	Uploading Local or SRA NGS Data & Creating a Results Metadata Table in Terra

4. PROCEDURE

4.1 Verify read files and metadata can be publicly shared

a. *Verify* that read files have been scrubbed of human data sufficiently for submission to the Sequencing Read Archive (SRA). No human read scrubbing or quality control is performed by the Terra_2_NCBI_PHB workflow. **This procedure assumes that read files have been verified ready for submission by the user.** *STOP* and remove human read data if present within the files to be submitted.

b. *Verify* that all metadata, including sample identification numbers, are suitable for public release. **This procedure assumes that all metadata have been verified suitable for public release and do not contain PII.** *STOP* and revise the metadata if they contain PII.

4.2 Verify NCBI Configuration File Prerequisite

a. *Open the Terra workspace* containing the data to be submitted and click the *DATA* tab. Select the *Workspace Data* icon. (Figure 1).

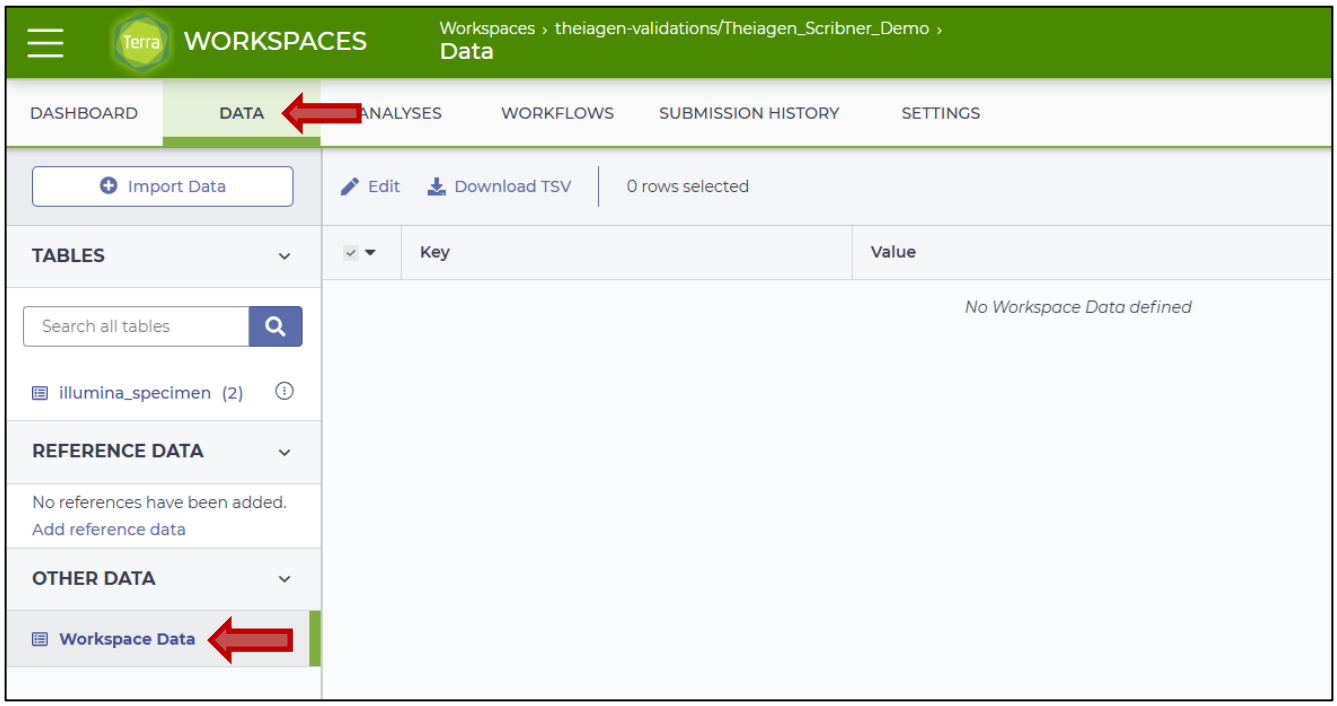

Figure 1

b. *Verify* that an NCBI Configuration File has been added to the workspace as a workspace data element (Figure 2). If no configuration file is present, *STOP* and consult *Appendix 10.1* to establish NCBI submission credentials, generate an NCBI Configuration file, and upload it to your workspace.

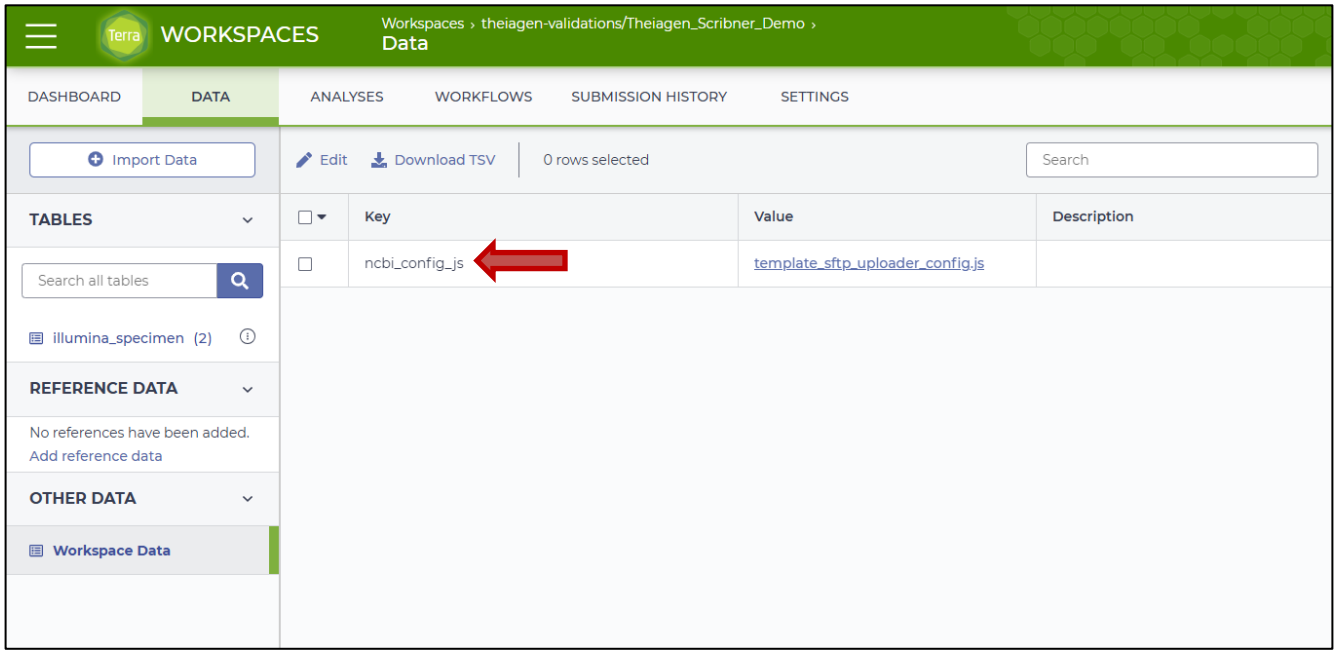

Figure 2

4.3 Verify the BioProject Required for Submission

a. *Verify* that your institution has identified or generated a BioProject to which data will be submitted. Specific BioProjects may be required by certain sequencing programs, for example, ARLN.
   i. If your institution still needs to register a BioProject for the sequences you aim to submit, *STOP* and register a new BioProject with NCBI. Guidance for registering BioProjects is available here.

4.4 Upload Sample and Sequencing Metadata to Data Table

a. *Open the Terra workspace* containing the data to be submitted and click the *DATA* tab (Figure 3).

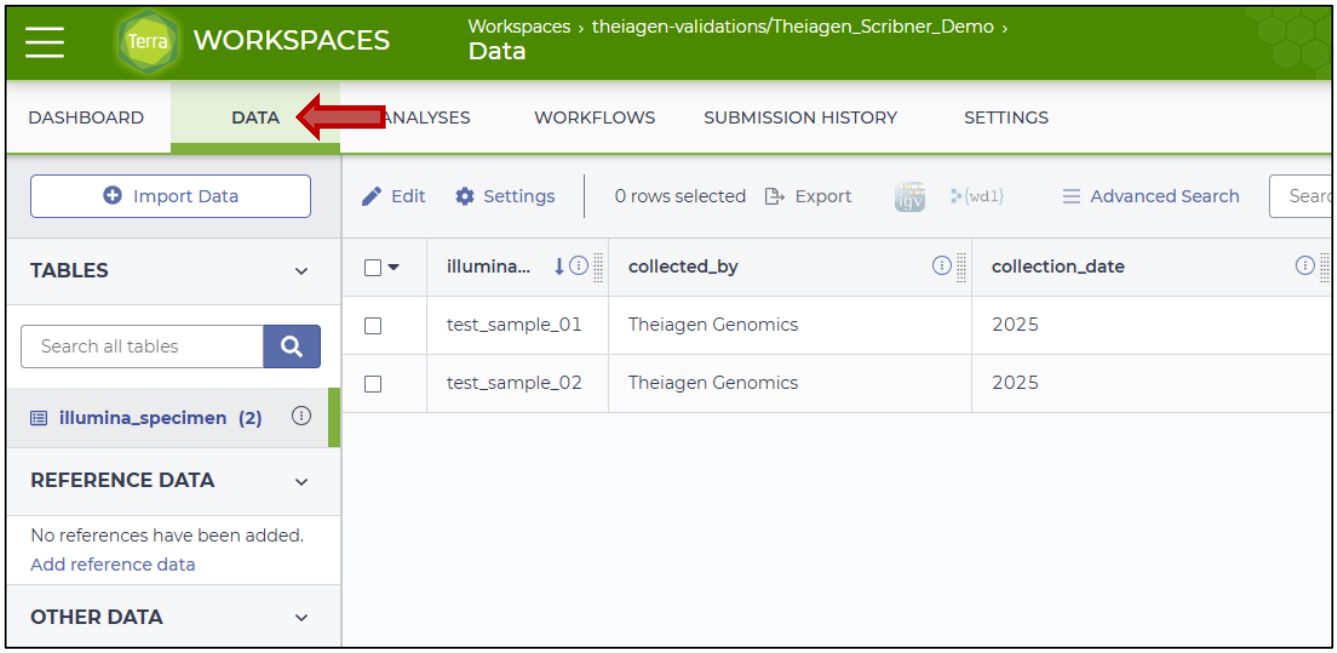

Figure 3

b. *Upload* relevant sample and sequencing metadata to the data table containing the sequencing reads you aim to submit by following the procedure in TG-TER-03. Metadata requirements are dictated by BioSample package utilized for submission. **Metadata must be free of PII.** Further details regarding metadata requirements are available in *Appendix 10.2*.

4.5 Verify public Google bucket prerequisite

a. *Verify* that your institution has identified a public Google bucket where your SRA reads will be temporarily stored before transferring to NCBI.
   i. **NOTE**: *Data stored in Terra workspaces can only be accessed by users to which the workspace owners have granted permission. Therefore, to enable submission, the Terra_2_NCBI_PHB workflow must copy sequencing reads to be submitted to a public Google bucket that is accessible by NCBI. Users may utilize any public Google bucket for this purpose. Theiagen Genomics provides a public Google bucket which may be used if no other option is available. See* Appendix 10.3 *for associated prerequisites.*

4.6 Configure Terra_2_NCBI_PHB Workflow Parameters

a. *Open the Terra workspace* containing the data to be submitted and click the *WORKFLOWS* tab. Select the *Terra_2_NCBI_PHB* workflow (Figure 4).

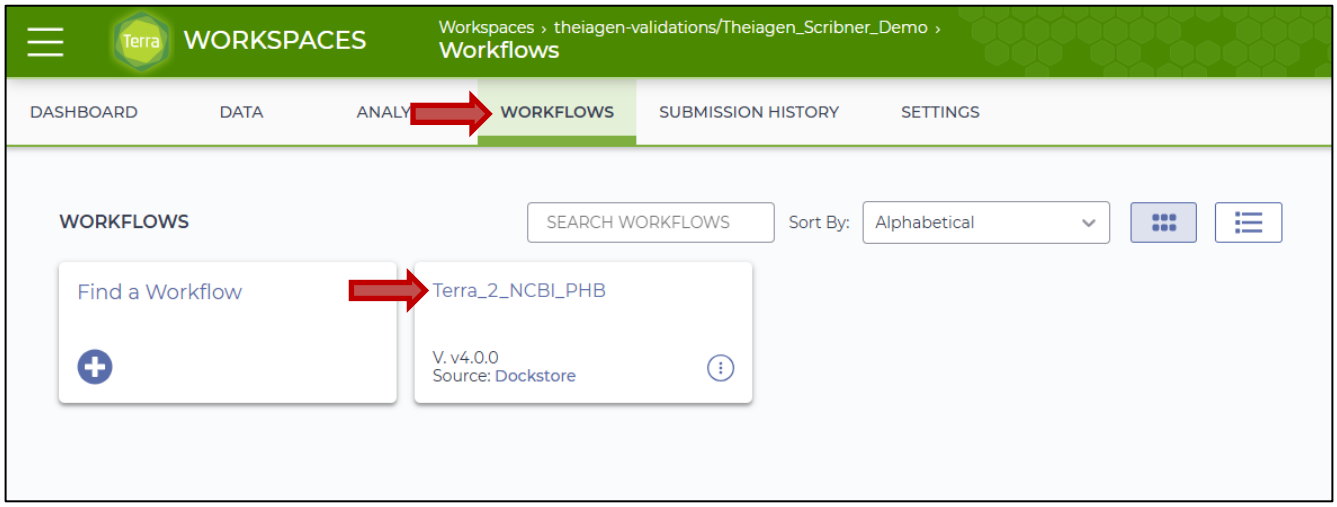

Figure 4

b. From the "Version" dropdown (Figure 5), choose the latest *Version* in the dropdown field or the version internally validated.

   i. ***NOTE:*** *All release versions are numbered in the XX.YY.ZZ format, where XX corresponds to the major release number, YY the minor release number, and ZZ the patch release number. The higher the number, the more recent the release is.*

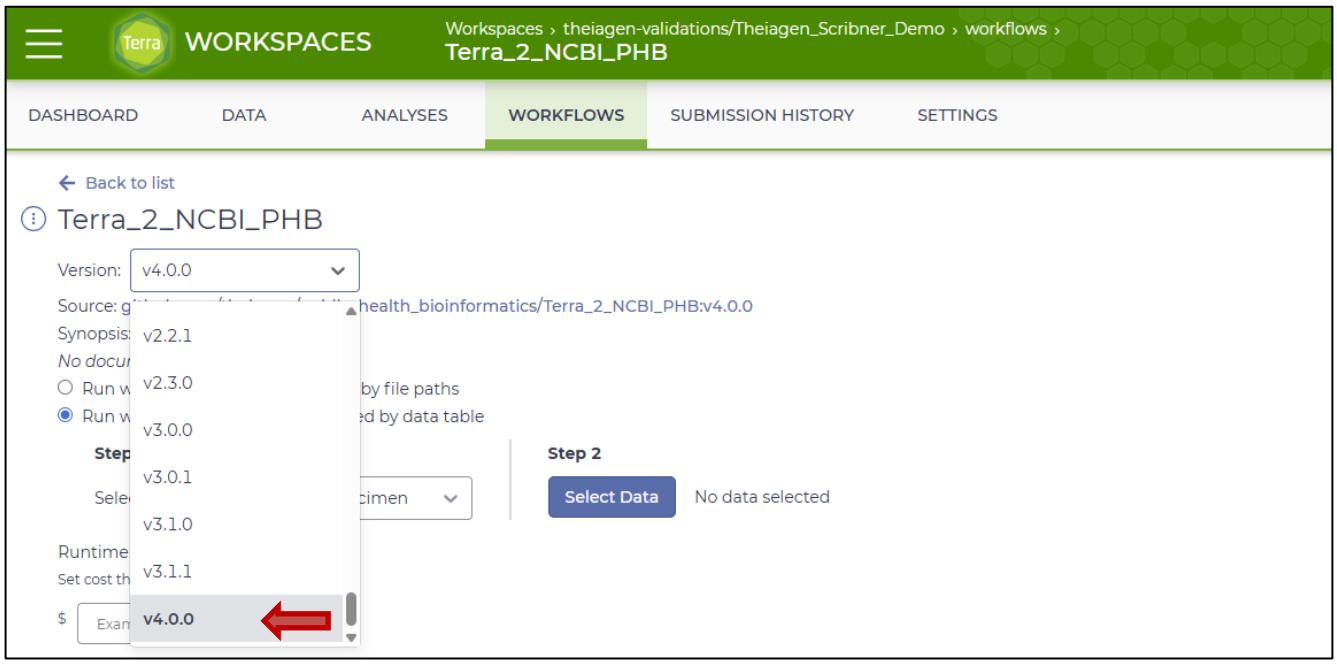

Figure 5

c. Select the second bullet to *Run workflow(s) with inputs defined by data table* (Figure 6).
d. Select the relevant **set** data table under the *Select data table* dropdown. For example, if the sample read data and metadata is stored in the "illumina_specimen" data table, select the "illumina_specimen_set" data table as shown (Figure 6).
e. Click *Select Data* (Figure 6).

Figure 6

f. In the pop-up window, *select the checkbox* for each sample to be included in the analysis (Figure 7).
    i. Click the checkbox dropdown and all to select all samples in the data table; if the checkbox at the top is checked, only the first 100 samples in the data table will be selected
    ii. A subset of samples may be chosen using the search bar to filter before selecting the checkbox dropdown and all to select only samples matching the search criteria
    iii. Optional: name the output set name to differentiate this analysis from others, e.g. *Terra_2_NCBI_PHB_YYYYMMDD*.
    iv. Click *OK*.

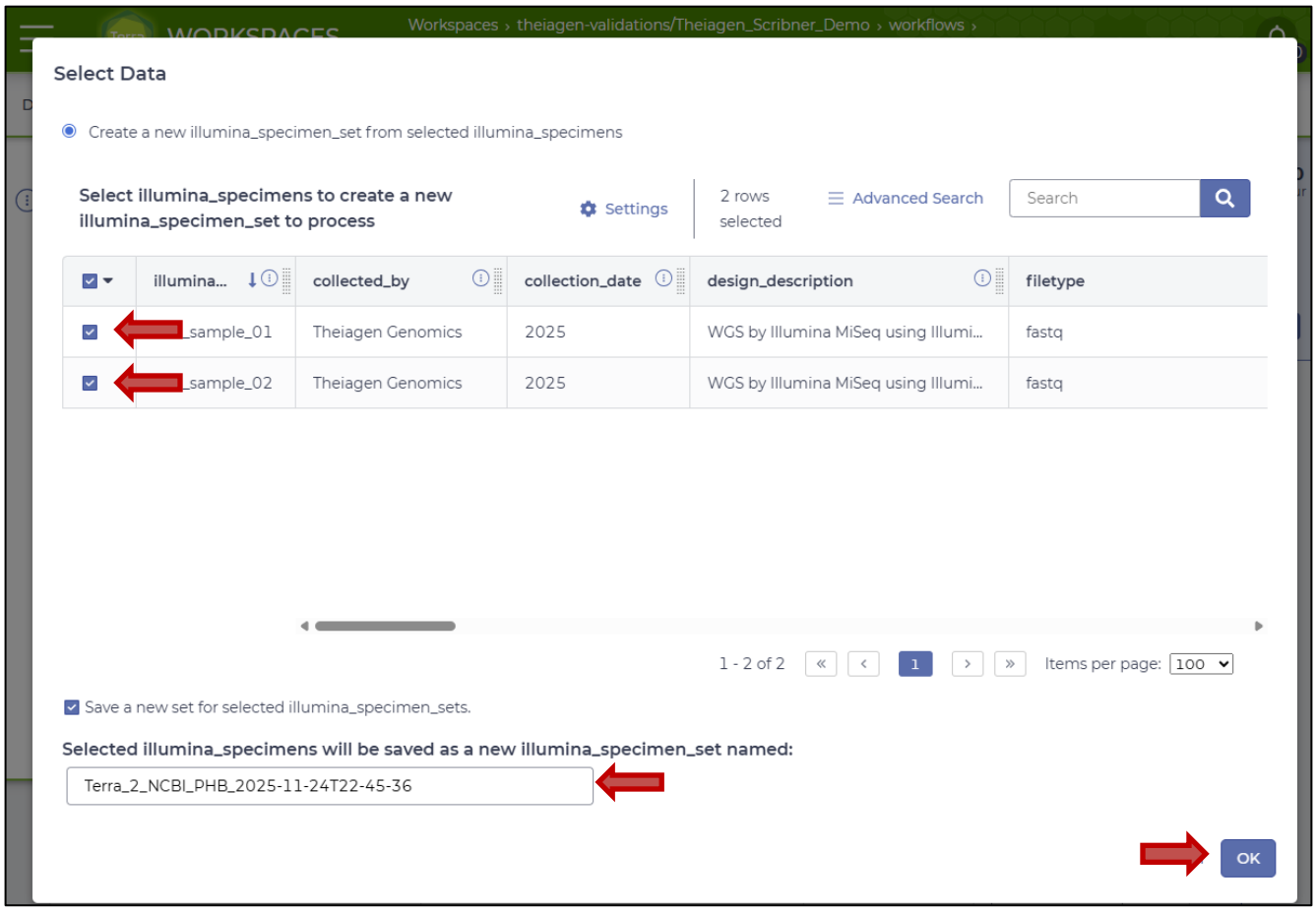

Figure 7

g. In the *INPUTS* tab, specify input values (Figure 8). Input attributes which must be considered are described below:

   i. *bioproject*, with a *String* type. The required input is the BioProject accession to which the submitted data will be associated in NCBI, in double quotes. The BioProject will be a string in the format *"PRJNA######"*.

   ii. *biosample_package*, with a *String* type. The required input is the BioSample package associated with the sample data to be submitted, in double quotes. BioSample packages represent types of samples and specify the attributes that must be described by the submitted metadata. The *biosample_package* input must align with the metadata fields uploaded to the data table. See *Appendix 10.2*. The *biosample_package* options available in Terra_2_NCBI_PHB are:

i. Microbe: *"Microbe"*
ii. Pathogen: clinical or host-associated: *"Pathogen"* or *"Pathogen.cl"*
iii. Pathogen: environmental/food/other: *"Pathogen.env"*
iv. SARS-CoV-2: wastewater surveillance: *"Wastewater"*
v. Virus: *"Virus"*

iii. *ncbi_config_js*, with a *File* type. The required input is the path to a configuration file located in the Terra workspace that contains your institution's username and password for the NCBI FTP and who should be contacted regarding the submission. Instructions for creating this file are available in *Appendix 10.1*.
   i. ***Example**:* workspace.ncbi_config_js

iv. *project_name*, with a String type. The required input is name of your Terra project, in double quotes. You can find this information in the URL of the webpage when you are viewing your Terra workspace. It is the string right after "#workspaces/".
   i. ***Example**: Assuming the URL for the Terra workspace is https://app.terra.bio/#workspaces/theiagen-demo/Theiagen_Scribner_Demo the project_name is "theiagen-demo"*

v. *sample_names*, with an *Array[String]* type. The required input is an array of the sample names for the samples you aim to submit.
   i. ***Example**: Assuming the data table is called "unicorn", the required input for the* sample_names *variable will be* this.unicorns.unicorn_id

vi. *sra_transfer_gcp_bucket*, with a String type. The required input is the path to a Google bucket where your SRA reads will be temporarily stored before transferring to SRA, in double quotes. To use the Theiagen SRA Transfer bucket, see *Appendix 10.3*.
   i. Example: *"gs://theiagen_sra_transfer"*

vii. *table_name*, with a String type. The required input is name of the Terra table where your samples are found, in double quotes.
   i. ***Example:** Assuming the data table is called "unicorn", the required input for the* table_name *variable will be* "unicorn"

viii. *workspace_name*, with a String type. The required input is name of your Terra workspace, in double quotes. You can find this information in the url of the webpage when you are viewing your Terra workspace. It is the string right after the Terra project.

i. ***Example**: Assuming the url for the Terra workspace is https://app.terra.bio/#workspaces/theiagen-demo/Theiagen_Scribner_Demo the workspace_name is "Theiagen_Scribner_Demo"*

ix. *submit_to_production*, with a Boolean type. Used to indicate whether the workflow should submit to NCBI's production environment.

i. Provide *true* if you would like to perform a Production submission.

1. **NOTE:** NCBI requires that users perform successful test submissions prior to accepting production submissions. If no successful test submissions have been completed, and the user provides *true* for this input, **the workflow will fail.**

ii. Leave blank or provide *false* if you would like to perform a Test submission.

x. *skip_biosample*, with a Boolean type. Used to indicate whether the workflow will submit metadata to the BioSample database prior to submitting sequencing data to the Sequencing Read Archive (SRA).

i. Leave blank or provide *false* if you would like to generate linked BioSample and SRA submissions.

ii. Provide *true* to generate an SRA submission linked to existing BioSample accessions for each sample. In this condition, **BioSample accessions for each sample must be uploaded to the Terra data table** in a column called *biosample_accession*. If BioSample accessions are not present in the data table, the **workflow will fail**.

1. If this behavior is desired, *STOP* and upload BioSample accessions for each sample to be submitted to the metadata table according to according to TG-TER-03.

xi. *read1_column_name* and *read2_column_name*, with String types. The workflow will assume that sequencing data to be submitted to SRA will be in columns titled "read1" and "read2". If the data to be submitted is stored in alternative column names, these column names must be provided here, in double quotes.

i. ***Example**: "read1_dehosted" and "read2_dehosted", respectively*

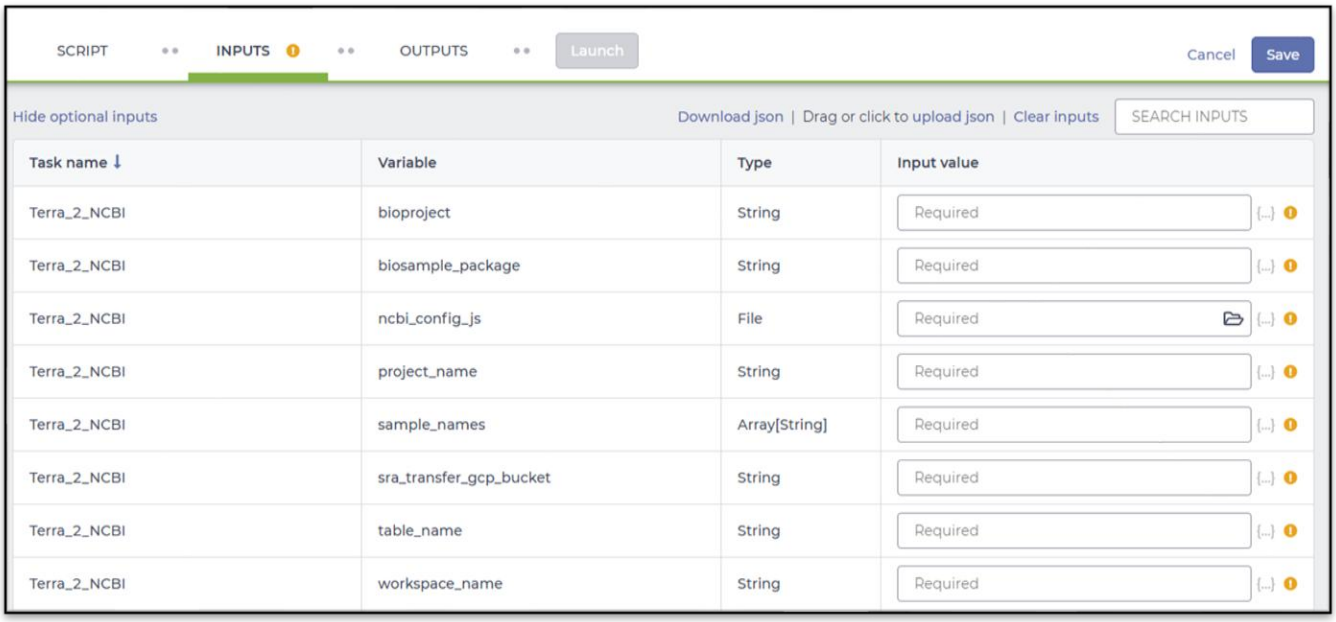

Figure 8

h. Specify outputs in the *OUTPUTS* tab by clicking *use defaults* (Figure 9).
i. In the top header, deselect the "*Use call caching*" option.
j. In the top header, select the "*Ignore empty outputs*" option and Click *save* (Figure 9).

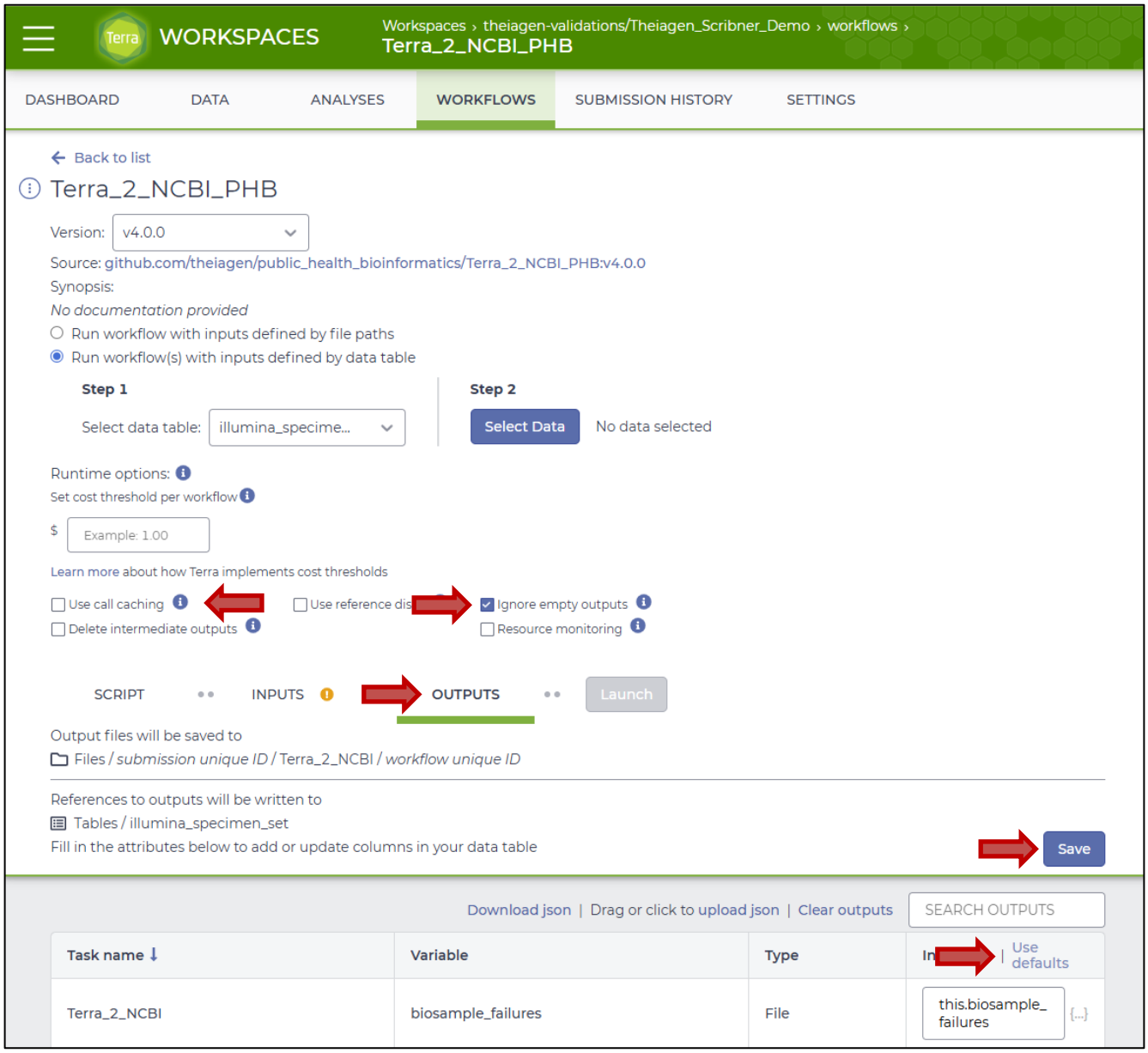

Figure 9

4.7 Launch Terra_2_NCBI_PHB workflow

a. Launch the workflow by clicking *Launch*; enter desired comments and click *Launch*.
    i. Not mandatory but highly encouraged to add a short description about the submission in the "Describe your submission" pop-up box.

b. The *SUBMISSION HISTORY* tab will populate and the status of the launched job will be displayed.

4.8 Verify submission success

a. **Failed** Terra_2_NCBI_PHB submissions result from improperly configured workflow inputs, permissions issues, or disrupted connections between Terra and NCBI during submission. This will result in a submission failed message in the *SUBMISSION HISTORY* tab. If observed, contact support@theiagen.com for support.
b. **Successful** Terra_2_NCBI_PHB submissions can reflect either successful submission of the associated data to NCBI, OR an **arrested submission due to metadata issues detected by the workflow or NCBI**.
    a. If the Terra_2_NCBI_PHB workflow is indicated to be successful in the *SUBMISSION HISTORY* tab, verify the submission was processed by NCBI through the steps below:
        i. Verify that BioSample accessions have populated for all submitted samples in a column called *biosample_accession* in the sample-level Terra data table. This reflects successful submission to the BioSample database.
        ii. Verify that the point of contact listed in the NCBI Config File received an email indicating successful submission to the SRA database.
    b. If either BioSample or SRA submission was unsuccessful, contact support@theiagen.com for support.

5. QUALITY RECORDS

- Raw read files
- Metadata uploaded to Terra workspace
- Submission logs populated to the set-level data table

6. TROUBLESHOOTING

- Consult with internal staff familiar with this procedure or contact support@theiagen.com for troubleshooting inquiries
- For document edit requests, contact support@theiagen.com

7. LIMITATIONS

- The Terra_2_NCBI_PHB workflow submits genomic data and metadata to NCBI via NCBI's FTP server. As such, all limitations on NCBI FTP submissions apply. For example, **all data and metadata must meet NCBI's requirements for formatting and content**, or the submission will not complete successfully.
- Terra_2_NCBI_PHB can only perform submissions for the **BioSample packages** listed in Appendix 10.2.
- Terra_2_NCBI_PHB can **only submit data to the BioSample and SRA databases**. It cannot submit data to other NCBI databases, including BioProject and Genbank.

8. REFERENCES

1. Libuit, Kevin G., Emma L. Doughty, James R. Otieno, Frank Ambrosio, Curtis J. Kapsak, Emily A. Smith, Sage M. Wright, et al. 2023. "Accelerating Bioinformatics Implementation in Public Health." Microbial Genomics 9 (7). https://doi.org/10.1099/mgen.0.001051
2. Theiagen Genomics. "Terra_2_NCBI." *Theiagen Genomics Public Health* Bioinformatics, 11 Aug. 2025, https://theiagen.github.io/public_health_bioinformatics/latest/workflows/public_data_sharing/terra_2_ncbi/
3. National Center for Biotechnology Information. "BioSample Packages." *BioSample*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/biosample/docs/packages/. Accessed 17 Feb. 2026.

9. REVISION HISTORY

Revision	Version	Release Date
Document creation	1	2/2026

10. APPENDICES

10.1 Establishing NCBI FTP Credentials and NCBI Configuration File

IMPORTANT NOTE: Uploading the Terra_2_NCBI Config to a workspace **enables all users with access to the workspace to submit data to NCBI using these credentials**. Only upload a Terra_2_NCBI Config file to a workspace if all users of that workspace should have the ability to launch submissions.

1. To use the Terra_2_NCBI_PHB workflow for public data submissions, the submitting institution **must** have access to the NCBI FTP. **NCBI FTP credentials and the associated NCBI configuration file only need to be established once and can be reused in all Terra workspaces for your institution.** To gain these credentials, we recommend emailing *sra@ncbi.nlm.nih.gov* a variation of the following example email, including all the information:

   Hello,

   We would like to automate submissions to the Submission Portal using XML metadata to accompany our cloud-hosted data files. We would like to upload via FTP and need to create a submission group.

   Here is the relevant information:

   a. Suggested group abbreviation:
   b. Full group name:
   c. Institution and department:
   d. Contact person (someone likely to remain at the location for an extended time):
   e. Contact email:
   f. Mailing address (including country and postcode):

   We will be using an existing submission pipeline that is known to work and would like to request that the production folder be activated. Thank you for your assistance!

2. From NCBI, you will need to get in response:
   a. an FTP address (it will likely be ftp-private.ncbi.nih.gov)
   b. Username (typically the suggested group abbreviation)
   c. Password
   d. an acknowledgment that the Production area of the submit folder has been activated.

***NOTE**: Please confirm that the response from NCBI states that a Production area of the submit folder has been created for this account, or else the submission pipeline will either fail or only run test submissions and not actually submit to NCBI.*

3. Before you can run the workflow for the first time, your institution will also need to create an NCBI Configuration File with these credentials.
    a. The configuration file contains the NCBI FTP username and password and is utilized by the workflow to submit data on Terra to the NCBI FTP. It also provides important information about who should be contacted regarding the submission. This file will be provided as input to the *ncbi_config_js* input of the Terra_2_NCBI workflow.
    b. We **strongly recommend** scheduling a meeting with Theiagen to ensure this file is properly configured and uploaded.
    c. Instructions for preparing this file and a template are available in the Terra_2_NCBI Config Instructions.

4. As indicated in the Terra_2_NCBI Config Instructions, upload the file to your Terra workspace and link the file in the workspace data elements according to TG-TER-03.

10.2 Collating Sample and Sequencing Metadata

1. Identify appropriate BioSample package for submission.
    a. Metadata requirements for the Terra_2_NCBI_PHB workflow are dictated by **the NCBI BioSample package** that will be used for submission. BioSample packages represent types of samples and specify the attributes that must be described by the submitted metadata. A list of NCBI BioSample packages and their metadata requirements are available here. Currently, Terra_2_NCBI_PHB only supports Pathogen, Virus, Microbe, and SARS-CoV-2 Wastewater Surveillance BioSample packages. Definitions of packages supported by Terra_2_NCBI_PHB are listed below with more requirements provided via the links:
        i. Pathogen.cl - any clinical or host-associated pathogen
        ii. Pathogen.env - environmental, food or other pathogen (no metadata formatter available at this time)
        iii. Microbe - bacteria or other unicellular microbes that do not fit under the MIxS, Pathogen, or Virus packages.
        iv. Virus - viruses **not** directly associated with disease
        v. SARS-CoV-2.wwsurv - SARS-CoV-2 wastewater surveillance samples

2. Collate sample metadata in metadata formatter.

a. For each package, Theiagen has created a metadata template spreadsheet to ensure proper inclusion and formatting of metadata.
   i. Terra_2_NCBI-PATHOGEN-metadata-2024-04-30.xlsx
   ii. Terra_2_NCBI-MICROBE-metadata-2022-07-11.xlsx
   iii. Terra_2_NCBI-VIRUS-metadata-2022-09-09.xlsx
   iv. Terra_2_NCBI-SC2WW-metadata-2025-01-10.xlsx

3. Upload the *Metadata* tab of the relevant metadata formatter to the Terra workspace according to TG-TER-03.

10.3 Establishing a Google bucket for transfer of sequencing reads to NCBI

1. Theiagen Genomics has established a public Google bucket which may be used by laboratories who do not have another public Google bucket available: *"gs://theiagen_sra_transfer"*. In order to utilize this resource, every member of your laboratory who will launch Terra_2_NCBI_PHB submissions **must** email support@theiagen.com to be granted write access to this bucket. **Terra_2_NCBI_PHB workflow submissions will fail for any user who does not have access to this bucket.**