## 1. PURPOSE/SCOPE

To standardize the process of running and analyzing Healthcare-Associated Infection (HAI) pathogen next generation sequencing (NGS) data using CDC's Phoenix workflow in Terra to generate assemblies, quality control (QC) metrics, and identify and characterize bacterial HAI pathogens for sequence type, antibiotic resistance and hypervirulence genes, and plasmid detection. Acceptable data types include Illumina paired end (PE) raw read files.

## 2. REQUIRED RESOURCES

- 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
- CDC's Phoenix Workflow in Terra, see TG-TER-03 appendix 9.2

## 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 CREATE A SAMPLE METADATA FILE (TSV FILE) FOR RAW READS, ASSEMBLIES, AND SRA FETCH

1. In Excel, *create a list* containing the following sample information:
   a. Column 1 header: *entity:HAI_id*, where *HAI* is the sample group/batch name (Fig 1)
      i. List all *sample IDs* in column 1
   **b. For analysis from raw sequencing reads (Fig 1):**
      i. Column 2 and 3 headers: *read1* and *read2*, respectively
         1. List the *full file paths* to read1 and read2 files in the cloud
   ***c.* For analysis from assembly data (Fig 2):**
      ***i.*** Column 2 header: *assembly_fasta*, or similar

| entity:HAI_id | read1 | read2 | run_id |
|---|---|---|---|
| 03-98DDCS | gs://theiagen-public-files | gs://theiagen-public-files | SEQ137 |
| 0398KL | gs://theiagen-public-files | gs://theiagen-public-files | SEQ137 |
| | ic-files | gs://theiagen-public-files | SEQ137 |

Figure 1: Raw Reads Metadata File.

| entity:HAI_id | assembly_fasta | run_id |
|---|---|---|
| 03-98DDCS | gs://theiagen-publ | SEQ137 |
| 19050801924 | gs://theiagen-publ | SEQ137 |
| 2022AZMC-0005 | gs://theiagen-publ | SEQ137 |
| CL2021-00283104 | gs://theiagen-publ | SEQ137 |
| | | SEQ137 |

Figure 2: Assembly Metadata file.

*d.* **For analysis using SRA fetch to pull read data (Fig 3):**

*i.*Column 2 header: *sra_accession*, or similar

*e.* *Optional: remaining columns may be used to add metadata like additional lab results, sample collection information, demographic data, etc*

f. Do not include spaces in the headers

2. *Save as* a txt or tsv file
3. *Upload* to Terra workspace; see TG-TER-03 for details

| entity:HAI_id | sra_accession | run_id |
|---|---|---|
| 03-98DDCS | gs://theiagen-publ | SEQ137 |
| 19050801924 | gs://theiagen-publ | SEQ137 |
| 2022AZMC-0005 | gs://theiagen-publ | SEQ137 |
| CL2021-00283104 | | 137 |
| | | 137 |

Figure 3: SRA Accession Metadata File.

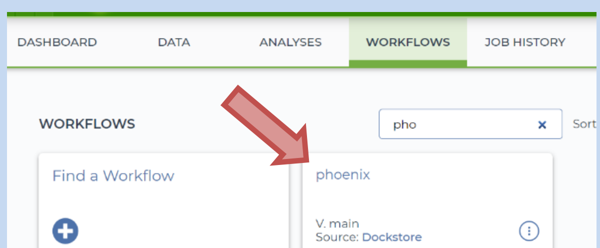


Figure 4.

## 4.2 RUNNING THE PHOENIX WORKFLOW

1. In Terra, open the *workspace* containing the data of interest and click the *workflows* tab
2. Select the *phoenix* workflow (Fig 4)
3. Choose the latest version of *version 2* in the version dropdown field or the internally validated (Fig 5, a)
4. Select the second bullet to *run workflow(s) with inputs defined by data table* (Fig 5, b)
5. Select the relevant data table name under the *select root entity type* dropdown (Fig 5, c)

6. Click *select data* (Fig 5, d)

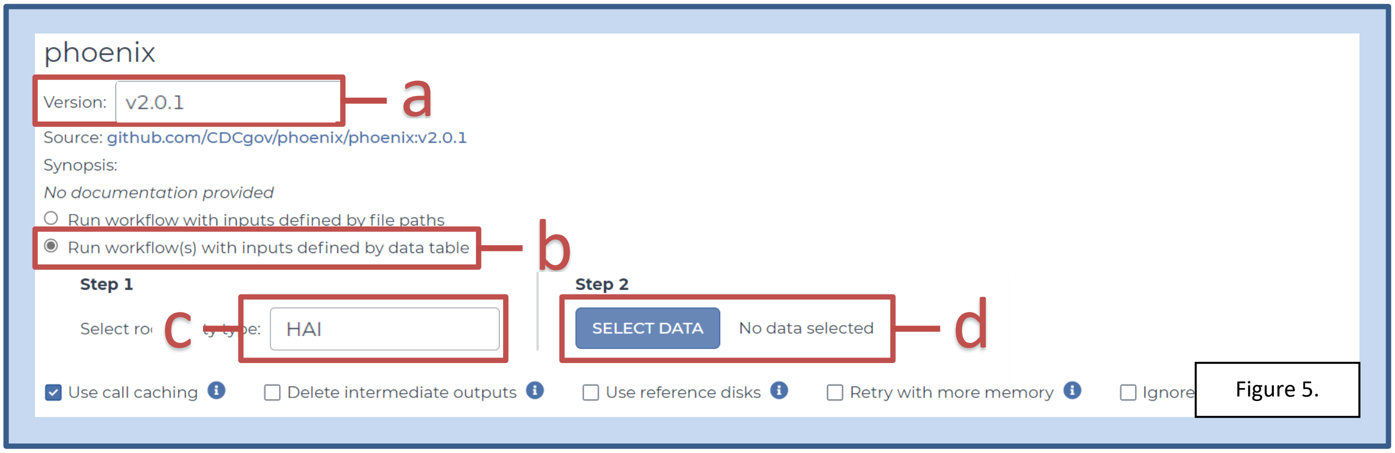


Figure 5.

7. In the pop-up window, *select the checkbox* for each sample to be included in the analysis (Fig 6)
   a. *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*
   b. *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*
   c. *Optional: name the output set name to differentiate this analysis from others, e.g. Phoenix_YYYYMMDDn; this populates a new row to the SET data table*
   d. Click *ok*

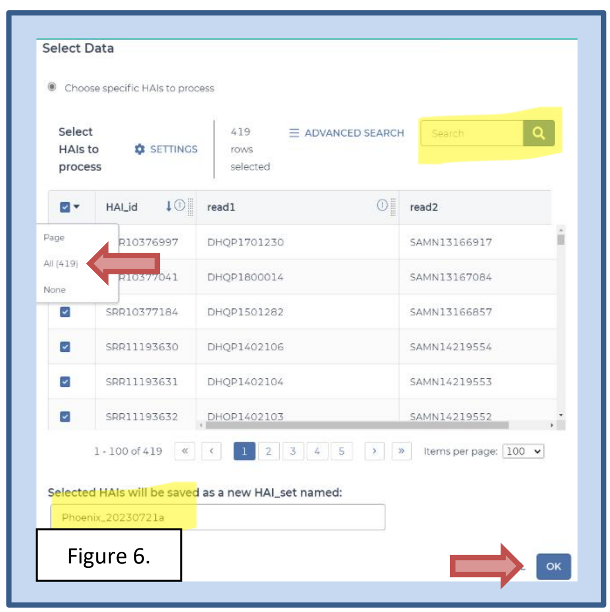


Figure 6.

8. In the *inputs* tab, set the first 3 attributes to the following, respectively (Fig 7)
   a. *"CDC_PHOENIX"* or *"PHOENIX"*
      i. Alternatively, to run Phoenix using assembly fasta files, input *"CDC_SCAFFOLDS"* or *"SCAFFOLDS"*
         1. NOTE: Assembly fields must be gzipped (fa.gz or fasta.gz) for analysis using Phoenix scaffolds
   b. *workspace.kraken2_phoenix*
      i. *kraken2_phoenix must be uploaded as a workspace data element; see appendix 10.1*
   c. *this.HAI_id*
      i. Where *HAI* is the column name in the data table containing sample IDs
9. Additionally specify sequencing data location:

a. **For raw reads and sra_fetch data**, specify in the read1 and read2 attribute fields as:

i. *this.read1**

ii. *this.read2**

1. *Where *read1* and *read2* are the metadata file column names containing the relevant files (section 4.1b)

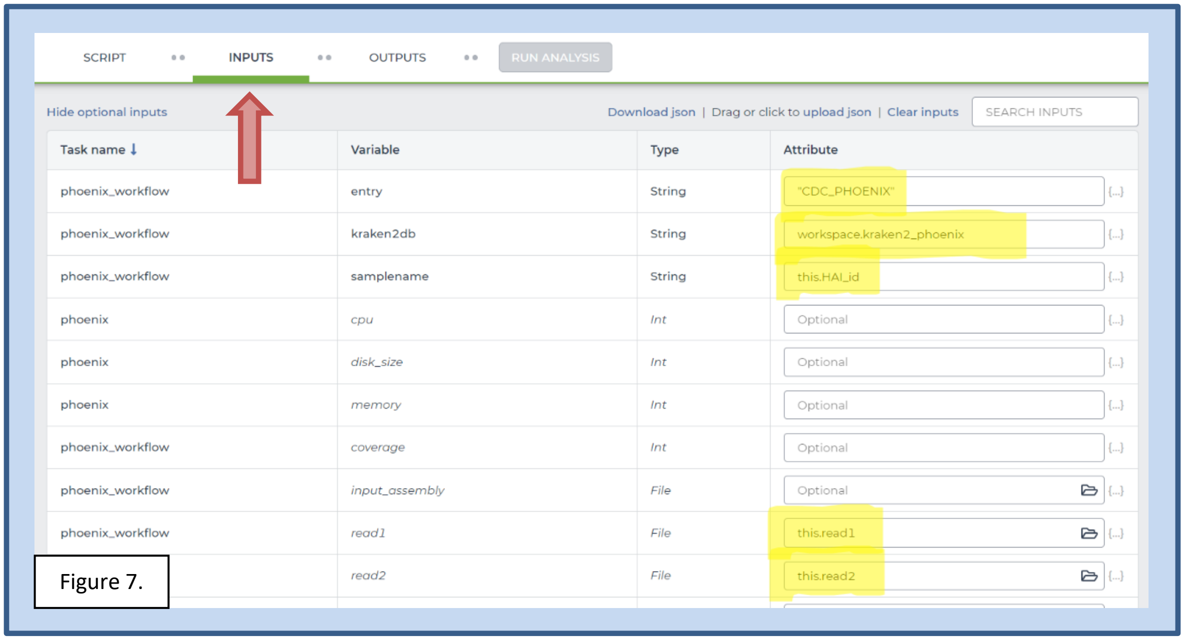


Figure 7.

b. **For assembly input data**, specify in the input_assembly field as:

i. this.assembly_fasta

1. Where assembly_fasta is the metadata file column name containing assemblies (section 4.1c)

10. Specify outputs in the *outputs* tab by clicking *use defaults* (Fig 8)

11. Click *save*

12. Launch the workflow by clicking *run analysis*; enter desired comments and click *launch*

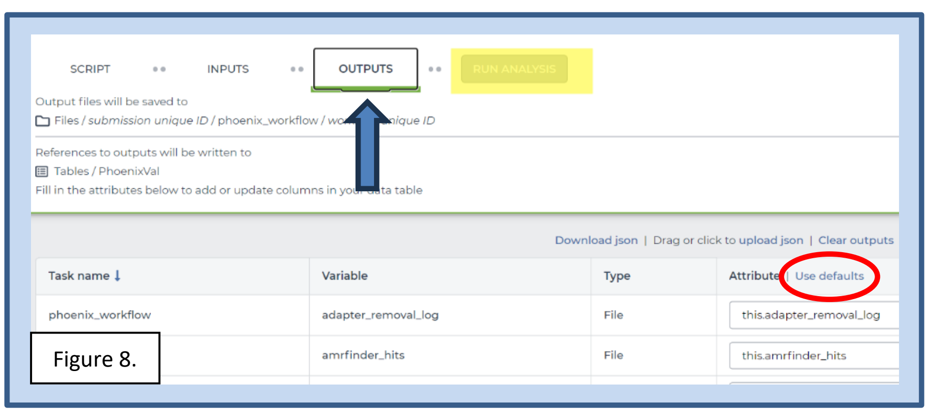


Figure 8.

## 4.3 DETERMINING: TAXONOMY, AMR PROFILE, HYPERVIRULENCE, AND PLASMID MARKERS

1. In the Terra *workspace* containing Phoenix data, navigate to the *data* tab
2. *Open the data table* by clicking on the name of the data table in the left sidebar
3. View *settings* above the data table, select *none* (Fig 9)
   a. Select lab-specific QC metric columns needed to make a sample pass/fail determination
   b. Additionally, select the following result columns: (Fig 9)
      i. *amrfinder_point_mutations*
      ii. *beta_lactam_resistance_genes*
      iii. *hypervirulence_genes*
      iv. *mlst1*
      v. *mlst2*
      vi. *mlst_scheme_1*
      vii. *mlst_scheme_2*
      viii. *other_ar_genes*
      ix. *plasmid_incompatability_replicons*
      x. *species*
         1. *<u>Optional</u>: save this column group for future use by clicking the* save this column selection *field,* naming it *(e.g. PhoenixResults), and clicking* save
   c. Click *done*
4. Determine the predicted taxonomy, sequence type, and AMR, hypervirulence, and plasmid characterization for each sample by viewing the corresponding columns
5. Follow lab-specific QC assessment, resulting, and reporting procedures, as applicable

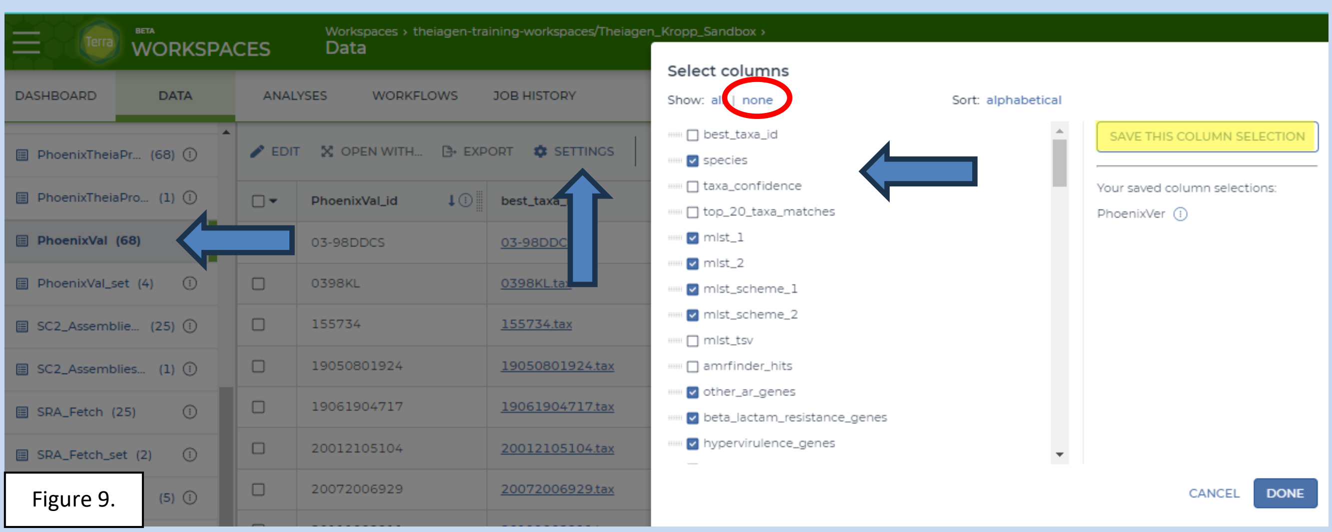


Figure 9.

## 5. QUALITY RECORDS

1. Raw reads
2. Metadata (tsv)
3. All Phoenix workflow outputs relevant to results

## 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. INTERFERENCES

N/A

## 8. REFERENCES

None

## 9. REVISION HISTORY

| Revision | Version | Release Date |
|---|---|---|
| Document creation | 1 | 7/2023 |
| Added "SCAFFOLDS" to entry field to run assemblies & gzip requirement | 2 | 5/2025 |
| | | |

# 10. APPENDICES

## 10.1 ADD A WORKSPACE DATA ELEMENT

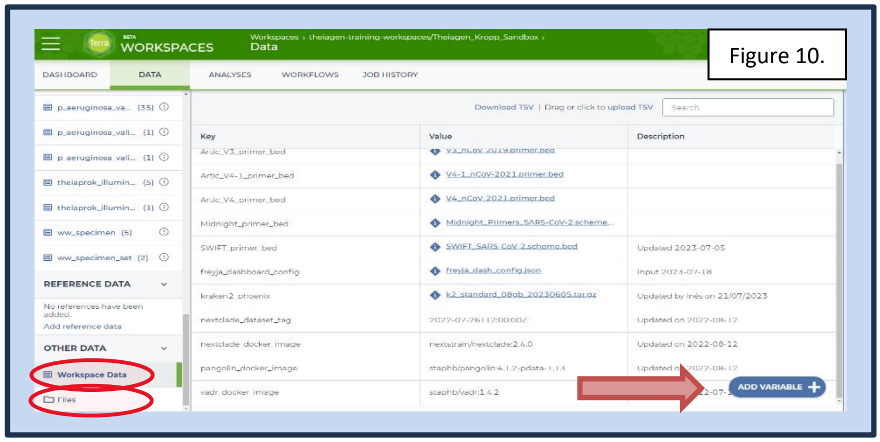

Figure 10.

1. Navigate to the *Terra workspace* where Phoenix will be run
2. To upload local files, open the *Files* tab in the bottom left of the workspace (Fig 10)
   a. Click *upload*
   b. Once the upload is complete, *right click* on the file name and click *copy link*
3. Open the *workspace data* tab (Fig 10) and click the *blue plus symbol* in the bottom right (Fig 10)
4. Click in the *key field* and *name the element* being added (Fig 11)
   a. E.g. to add the Kraken2 database, the key *kraken2_phoenix* may be used to specify its use with the Phoenix workflow
5. In the value field, choose *string* as the value type
   a. *Paste the file path*
      i. E.g. for the kraken2 database, paste *gs://theiagen-public-files-rp/terra/theiaprok-files/k2_standard_08gb_20230605.tar.gz*
   b. For other string elements like docker images and dataset tags, *paste the ID value*
      i. E.g. for the nextclade docker image, add *nextstrain/nextclade:2.13.0*
      ii. Always ensure the docker images and dataset tags are aligned with versions used for internal validation procedures

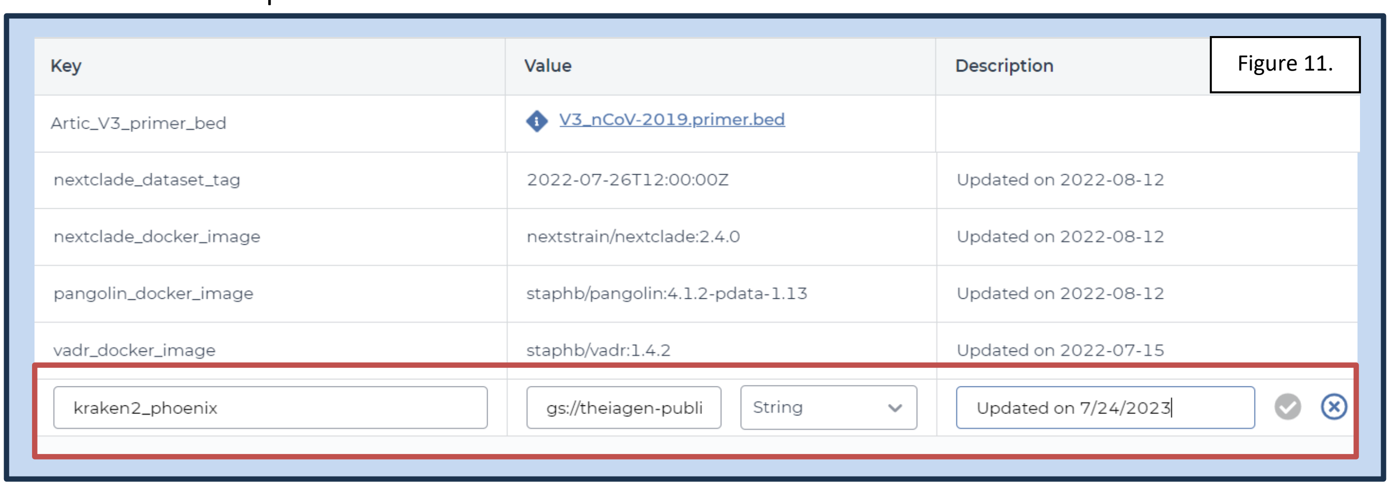

Figure 11.