Analyzing HAI Pathogens in Terra **using CDC's Phoenix Workflow, Version 2**		
Document TG-PX-V2, Version 1		
Date:	Effective Date:	Workflow Version
7/20/2023	8/2023	v2

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

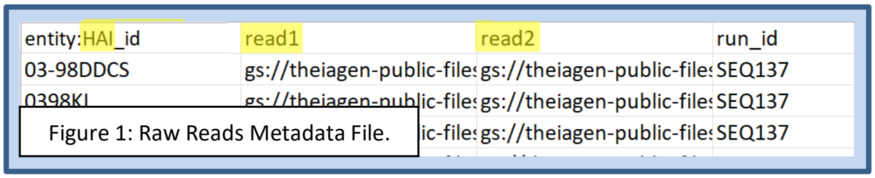
3. RELATED DOCUMENTS

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

4. PROCEDURE

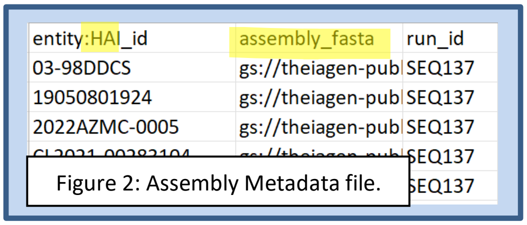
4.1 CREATE A SAMPLE METADATA FILE (TSV FILE) FOR RAW READS, ASSEMBLIES, AND SRA FETCH

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
0398KI	gs://theiagen-public-files	gs://theiagen-public-files	SEQ137
	ic-files	gs://theiagen-public-files	SEQ137

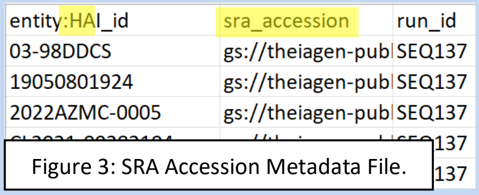
Figure 1: Raw Reads Metadata File.



entity:HAI_id	assembly_fasta	run_id
03-98DDCS	gs://theiagen-publ	SEQ137
19050801924	gs://theiagen-publ	SEQ137
2022AZMC-0005	gs://theiagen-publ	SEQ137
[illegible]	[illegible]	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
		137
		137

Figure 3: SRA Accession Metadata File.

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)

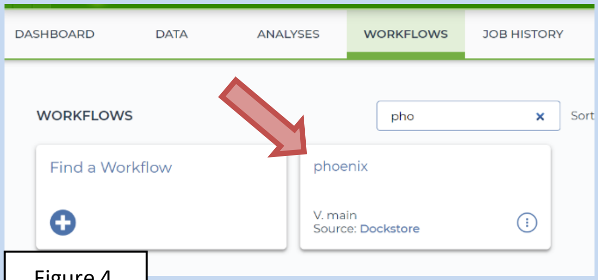
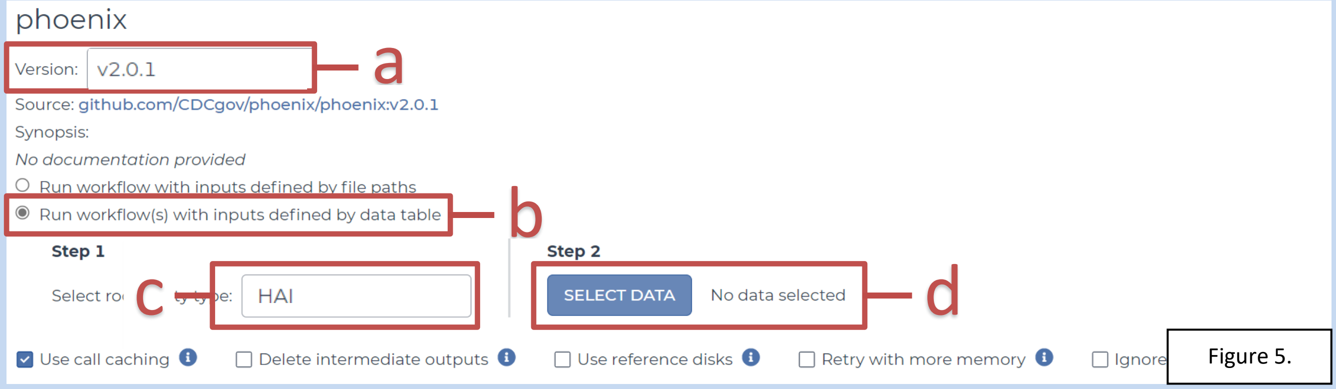


Figure 4.



phoenix

Version: v2.0.1

Source: github.com/CDCgov/phoenix/phoenix:v2.0.1

Synopsis:

No documentation provided

○ Run workflow with inputs defined by file paths

● Run workflow(s) with inputs defined by data table

Step 1

Select ro ty type: HAI

Step 2

SELECT DATA No data selected

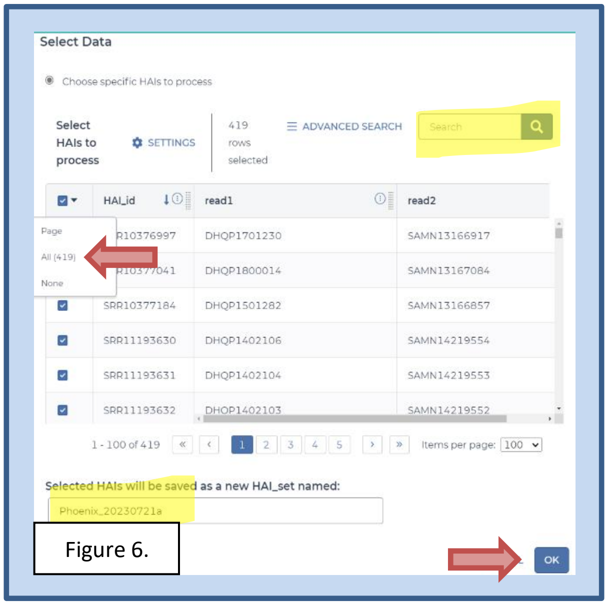
☑ Use call caching ☐ Delete intermediate outputs ☐ Use reference disks ☐ Retry with more memory ☐ Ignore

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*

8. In the *inputs* tab, set the first 3 attributes to the following, respectively (Fig 7)
   a. *"CDC_PHOENIX"* or *"PHOENIX"*
   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)



Select Data

● Choose specific HAIs to process

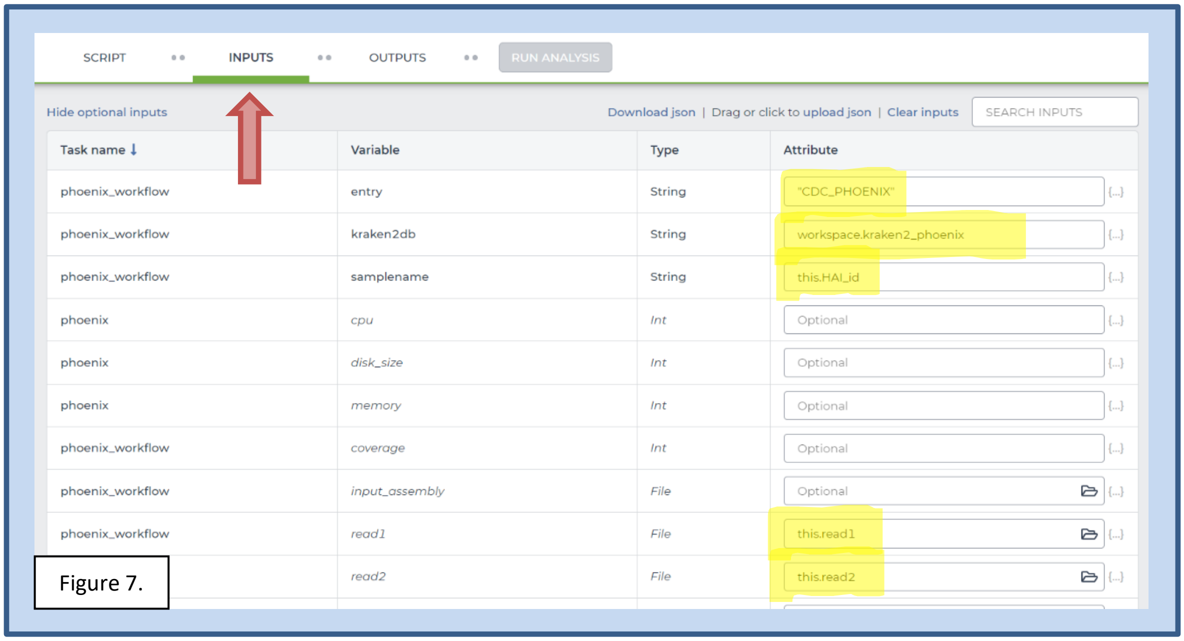
Select HAIs to process   SETTINGS   419 rows selected   ADVANCED SEARCH   Search

☑	HAI_id	read1	read2
	R10376997	DHQP1701230	SAMN13166917
	R10377041	DHQP1800014	SAMN13167084
☑	SRR10377184	DHQP1501282	SAMN13166857
☑	SRR11193630	DHQP1402106	SAMN14219554
☑	SRR11193631	DHQP1402104	SAMN14219553
☑	SRR11193632	DHQP1402103	SAMN14219552

1 - 100 of 419   « ‹ 1 2 3 4 5 › »   Items per page: 100

Selected HAIs will be saved as a new HAI_set named: Phoenix_20230721a

Figure 6.

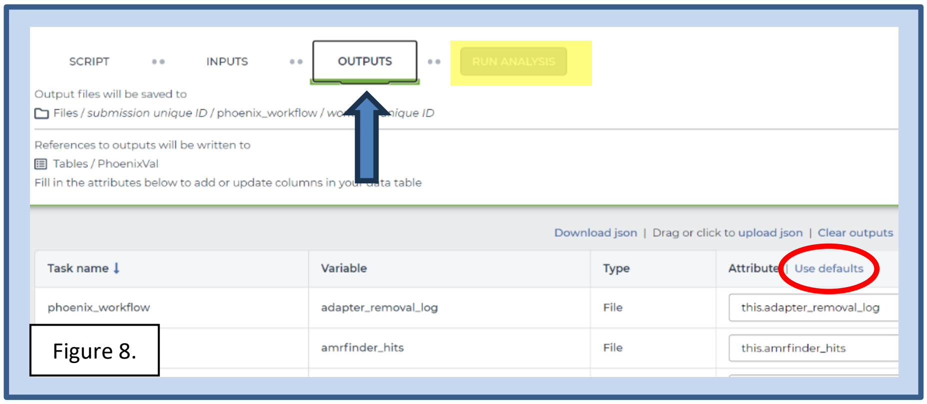


Task name	Variable	Type	Attribute
phoenix_workflow	entry	String	"CDC_PHOENIX"
phoenix_workflow	kraken2db	String	workspace.kraken2_phoenix
phoenix_workflow	samplename	String	this.HAI_id
phoenix	*cpu*	*Int*	Optional
phoenix	*disk_size*	*Int*	Optional
phoenix	*memory*	*Int*	Optional
phoenix_workflow	*coverage*	*Int*	Optional
phoenix_workflow	*input_assembly*	*File*	Optional
phoenix_workflow	*read1*	*File*	this.read1
	read2	*File*	this.read2

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*



Output files will be saved to

Files / *submission unique ID* / phoenix_workflow / wo… *nique ID*

References to outputs will be written to

Tables / PhoenixVal

Fill in the attributes below to add or update columns in you… ta table

Task name	Variable	Type	Attribute | Use defaults
phoenix_workflow	adapter_removal_log	File	this.adapter_removal_log
	amrfinder_hits	File	this.amrfinder_hits

Figure 8.

4.3 DETERMINING: TAXONOMY, AMR CHARACTERIZATION, 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. *Optional: 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

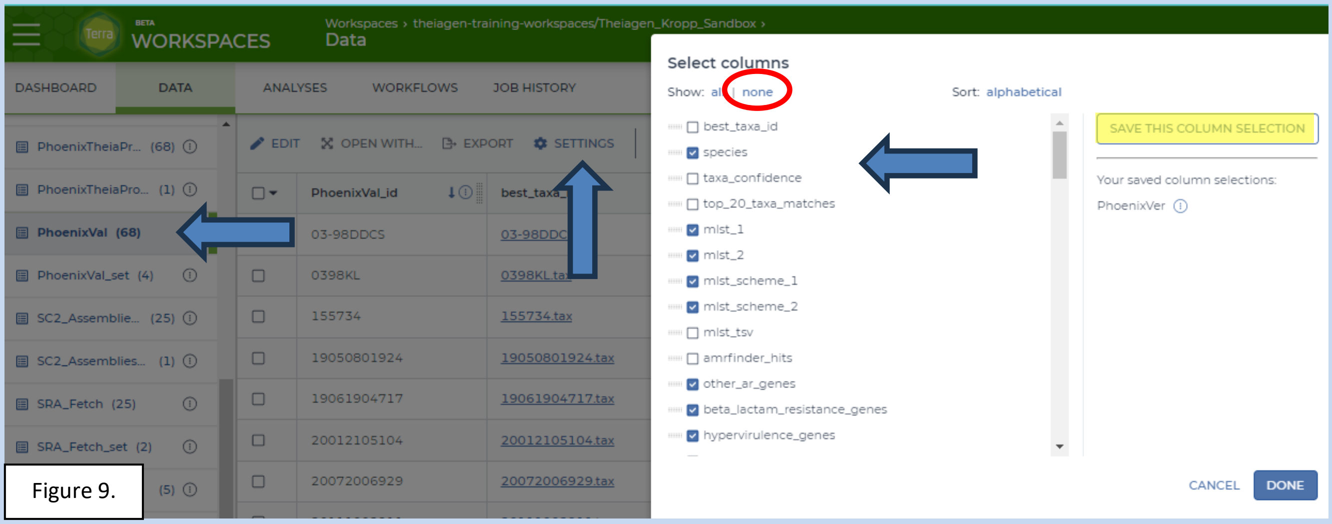


Figure 9.

☐	PhoenixVal_id	best_taxa_
	03-98DDCS	03-98DDC
☐	0398KL	0398KL.ta
☐	155734	155734.tax
☐	19050801924	19050801924.tax
☐	19061904717	19061904717.tax
☐	20012105104	20012105104.tax
☐	20072006929	20072006929.tax

- ☐ best_taxa_id
- ☑ species
- ☐ taxa_confidence
- ☐ top_20_taxa_matches
- ☑ mlst_1
- ☑ mlst_2
- ☑ mlst_scheme_1
- ☑ mlst_scheme_2
- ☐ mlst_tsv
- ☐ amrfinder_hits
- ☑ other_ar_genes
- ☑ beta_lactam_resistance_genes
- ☑ hypervirulence_genes

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

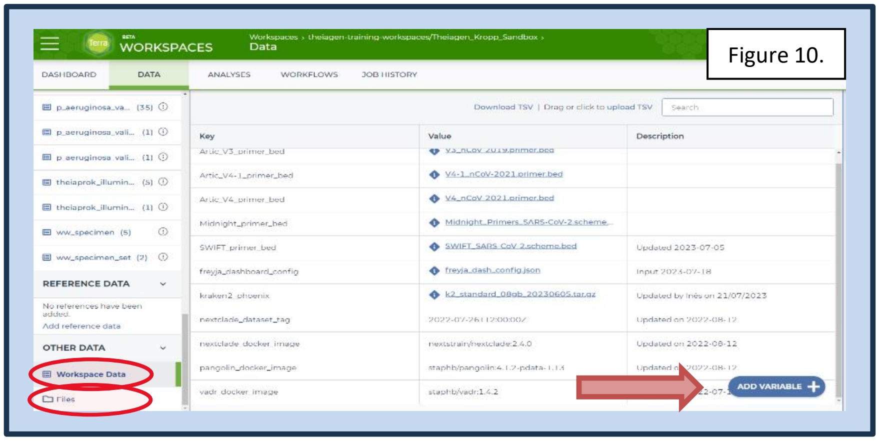
9. REVISION HISTORY

Revision	Version	Release Date
Document creation	1	7/2023

10. APPENDICES

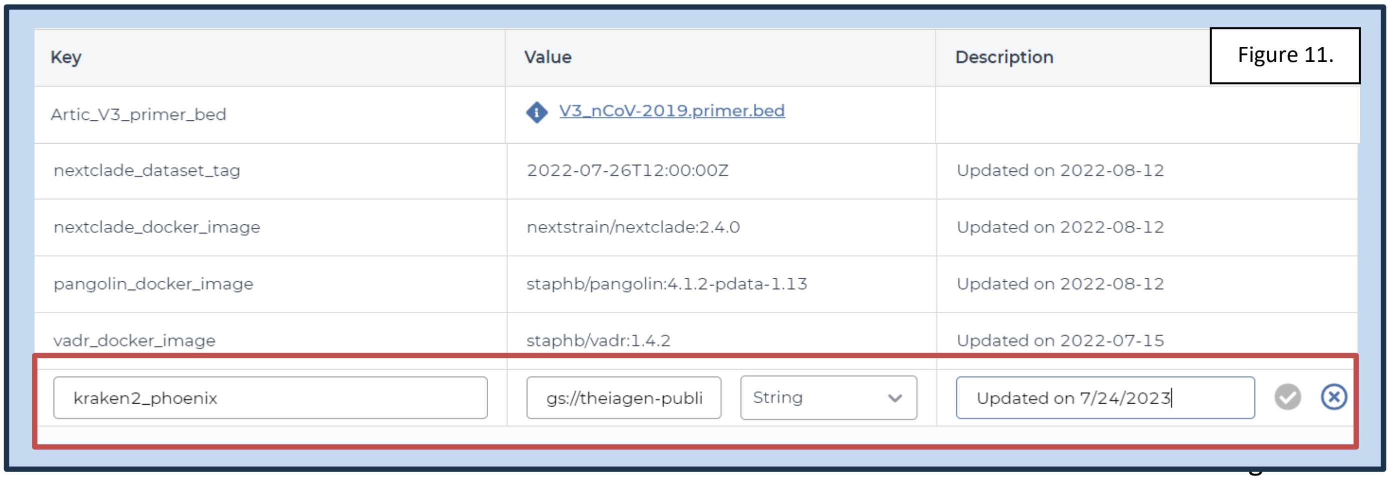
10.1 ADD A WORKSPACE DATA ELEMENT

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



Key	Value	Description
Artic_V3_primer_bed	V3_nCoV-2019.primer.bed	
Artic_V4-1_primer_bed	V4-1_nCoV-2021.primer.bed	
Artic_V4_primer_bed	V4_nCoV-2021.primer.bed	
Midnight_primer_bed	Midnight_Primers_SARS-CoV-2.scheme...	
SWIFT_primer_bed	SWIFT_SARS_CoV_2.scheme.bed	Updated 2023-07-05
freyja_dashboard_config	freyja_dash_config.json	Input 2023-07-18
kraken2_phoenix	k2_standard_08gb_20230605.tar.gz	Updated by Inés on 21/07/2023
nextclade_dataset_tag	2022-07-26T12:00:00Z	Updated on 2022-08-12
nextclade_docker_image	nextstrain/nextclade:2.4.0	Updated on 2022-08-12
pangolin_docker_image	staphb/pangolin:4.1.2-pdata-1.13	Updated on 2022-08-12
vadr_docker_image	staphb/vadr:1.4.2	22-07-1

Figure 10.



Key	Value		Description
Artic_V3_primer_bed	V3_nCoV-2019.primer.bed		
nextclade_dataset_tag	2022-07-26T12:00:00Z		Updated on 2022-08-12
nextclade_docker_image	nextstrain/nextclade:2.4.0		Updated on 2022-08-12
pangolin_docker_image	staphb/pangolin:4.1.2-pdata-1.13		Updated on 2022-08-12
vadr_docker_image	staphb/vadr:1.4.2		Updated on 2022-07-15
kraken2_phoenix	gs://theiagen-publi	String	Updated on 7/24/2023

Figure 11.