Running Theiagen's kSNP4 Workflow in Terra for Phylogenetic Analysis of Bacterial Genomes	
Document TG-kSNP4-01, Version 1	
Date:	Workflow Version:
May/2026	PHB v4.1.0

1. PURPOSE/SCOPE

To standardize the process of running bacterial foodborne outbreak isolates' next generation sequencing (NGS) data using Theiagen's kSNP4 workflow in Terra to generate a phylogenetic tree. Acceptable data types include assembly FASTA file format and metadata imported in TSV file format. Follow all instrument, raw read, and assembly quality control (QC) procedures determined internally and by the sequencing program (e.g. PulseNet, GenomeTrakr, etc), as appropriate, prior to performing phylogenetic analysis.

2. REQUIRED RESOURCES

- Computer
- Internet browser
    - Google Chrome, Firefox, or Edge
- Google account
- Terra account, linked to Google account
- Illumina PE raw read files uploaded to Terra workspace, see TG-TER-04
- Theiagen's kSNP4 Workflow in Terra, see appendix 10.1

IMPORTANT NOTES

- The kSNP4 workflow is a SET-level workflow, meaning it cannot be run on individual samples; a sample set may, however, be created before launching kSNP4 as indicated in section 4.1, bullet 9
- FASTA assemblies generated using TheiaProk are located in the sample-level data table under the assembles_fasta column

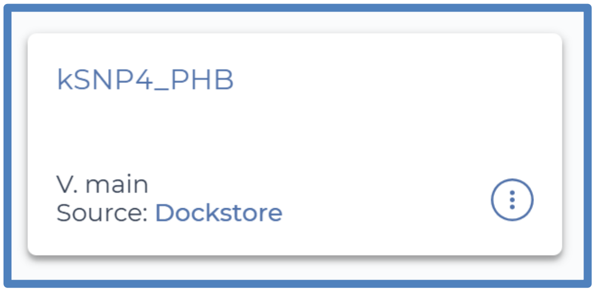
3. RELATED DOCUMENTS

Document Number	Document Name
TG-TER-04	Linking BaseSpace and Importing BaseSpace Reads to Terra Workspace

4. PROCEDURE

4.1 RUNNING THE kSNP4 WORKFLOW

1. Open Terra and navigate to the *workflows* tab within the workspace containing the data of interest
2. Select the *kSNP4_PHB* workflow (Figure 1)

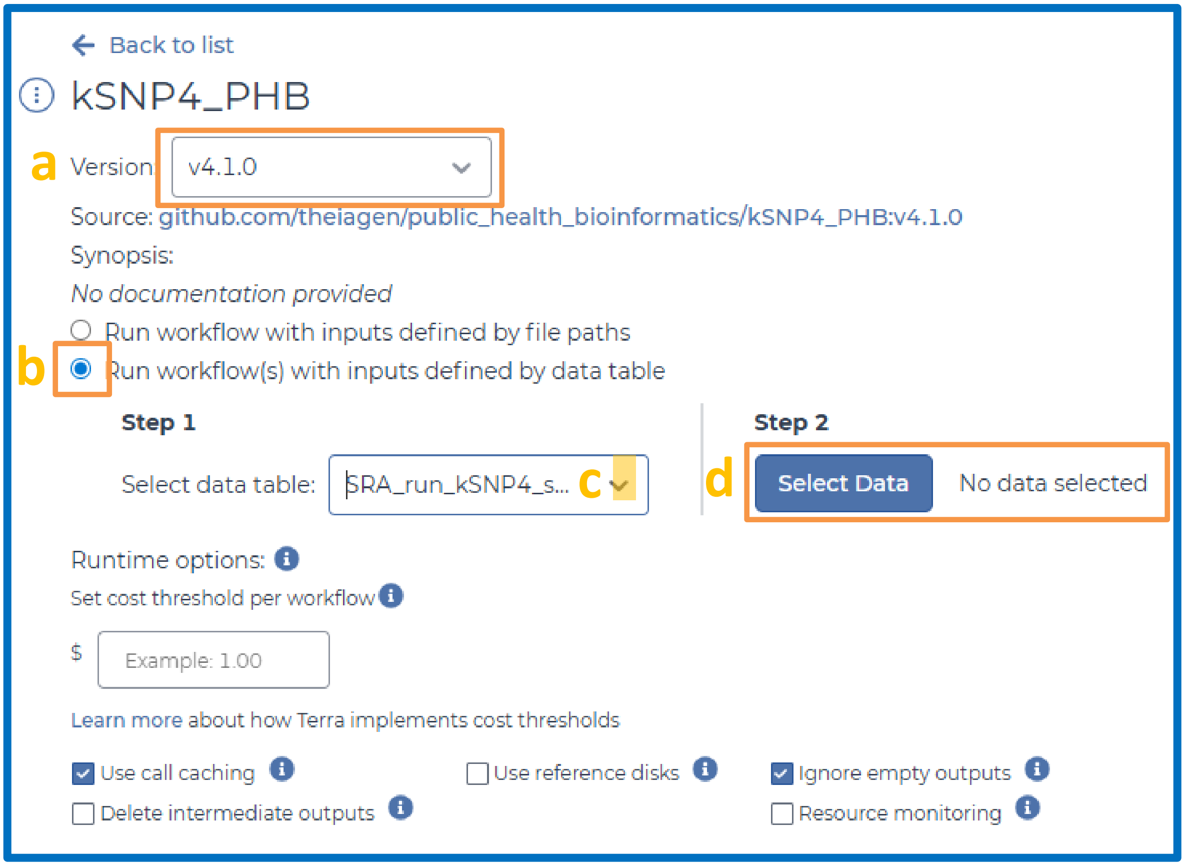


kSNP4_PHB

V. main
Source: Dockstore

Figure 1

3. *Optional:* *Uncheck call caching* (Figure 2)
4. *Optional:* *Check the box to* ignore empty outputs *(Figure 2)*
5. Choose the latest version of the workflow in the *version dropdown field*, or the workflow version that was used during internal assay validation (Figure 2, a)
6. Select the second bullet to *run workflow(s) with inputs defined by data table* (**Error! Reference source not found.**, b)
7. Select the relevant data table name (**the Set data table**) under the select *root entity type* dropdown (Figure 2, c)



kSNP4_PHB

Version: v4.1.0

Source: github.com/theiagen/public_health_bioinformatics/kSNP4_PHB:v4.1.0

Synopsis:

No documentation provided

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

Step 1

Select data table: SRA_run_kSNP4_s...

Step 2

Select Data — No data selected

Runtime options:

Set cost threshold per workflow

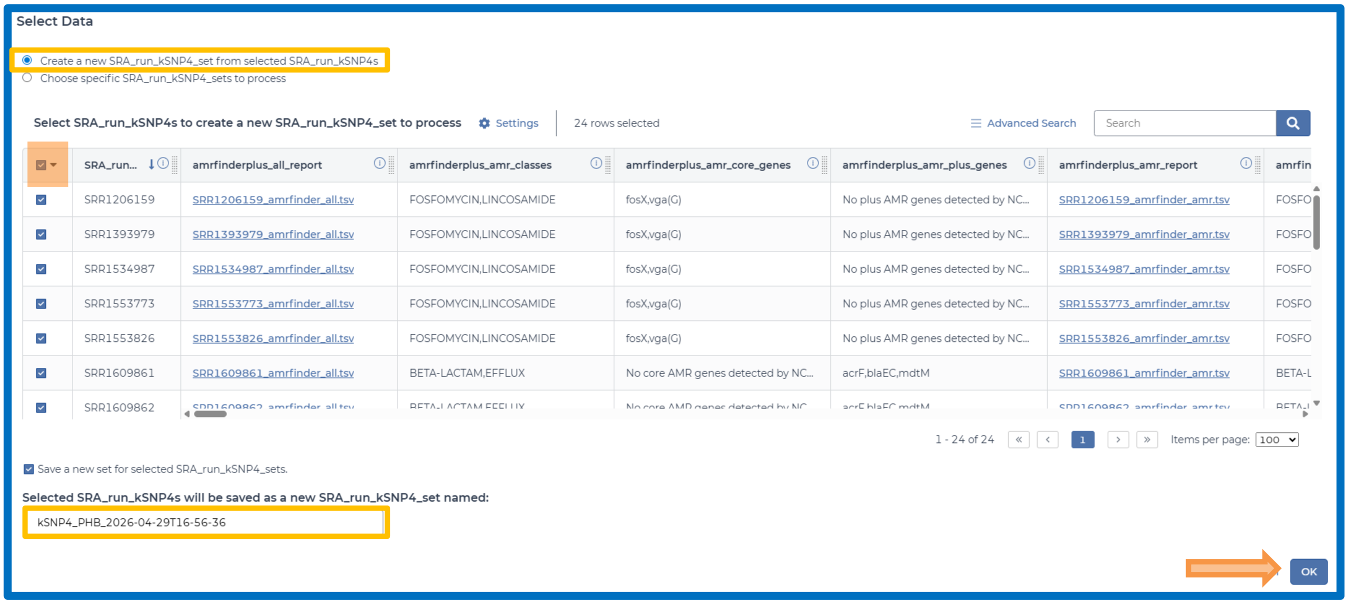
$ Example: 1.00

Learn more about how Terra implements cost thresholds

☑ Use call caching
☐ Use reference disks
☑ Ignore empty outputs
☐ Delete intermediate outputs
☐ Resource monitoring

Figure 2

8. Click *select data* (Figure 2, d)
9. In the pop-up window, choose the first bullet at the top for *Create a new [SRA_run_kSNP4]_set from selected SRA_run_kSNP4s* (Figure 3)
   a. *Select the checkbox* for each sample to be included in the analysis
   b. *The checkbox at the top may be used to select all samples listed*
   c. *Additionally, a subset of samples may be chosen using the search bar to filter before selecting the checkbox at the top to only select samples matching the search criteria*
   Optional: Name the sample set in the box at the bottom left, e.g. *[initials]_SetName_YYMMDD*. *This is a good way of creating groups of samples that will be analyzed together.*
10. Alternatively, a previous sample set may be selected by selecting the second bullet to “Choose specific SRA-run_kSNP4_sets to process” and marking the checkbox for the appropriate samlple set (Figure 3)
11. Scroll to the bottom and click *ok*


Select Data

● Create a new SRA_run_kSNP4_set from selected SRA_run_kSNP4s
○ Choose specific SRA_run_kSNP4_sets to process

Select SRA_run_kSNP4s to create a new SRA_run_kSNP4_set to process

24 rows selected

☑	SRA_run...	amrfinderplus_all_report	amrfinderplus_amr_classes	amrfinderplus_amr_core_genes	amrfinderplus_amr_plus_genes	amrfinderplus_amr_report	amrfin
☑	SRR1206159	SRR1206159_amrfinder_all.tsv	FOSFOMYCIN,LINCOSAMIDE	fosX,vga(G)	No plus AMR genes detected by NC...	SRR1206159_amrfinder_amr.tsv	FOSFO
☑	SRR1393979	SRR1393979_amrfinder_all.tsv	FOSFOMYCIN,LINCOSAMIDE	fosX,vga(G)	No plus AMR genes detected by NC...	SRR1393979_amrfinder_amr.tsv	FOSFO
☑	SRR1534987	SRR1534987_amrfinder_all.tsv	FOSFOMYCIN,LINCOSAMIDE	fosX,vga(G)	No plus AMR genes detected by NC...	SRR1534987_amrfinder_amr.tsv	FOSFO
☑	SRR1553773	SRR1553773_amrfinder_all.tsv	FOSFOMYCIN,LINCOSAMIDE	fosX,vga(G)	No plus AMR genes detected by NC...	SRR1553773_amrfinder_amr.tsv	FOSFO
☑	SRR1553826	SRR1553826_amrfinder_all.tsv	FOSFOMYCIN,LINCOSAMIDE	fosX,vga(G)	No plus AMR genes detected by NC...	SRR1553826_amrfinder_amr.tsv	FOSFO
☑	SRR1609861	SRR1609861_amrfinder_all.tsv	BETA-LACTAM,EFFLUX	No core AMR genes detected by NC...	acrF,blaEC,mdtM	SRR1609861_amrfinder_amr.tsv	BETA-L
☑	SRR1609862	SRR1609862_amrfinder_all.tsv	BETA-LACTAM,EFFLUX	No core AMR genes detected by NC	acrF,blaEC,mdtM	SRR1609862_amrfinder_amr.tsv	BETA-L

1 - 24 of 24 Items per page: 100

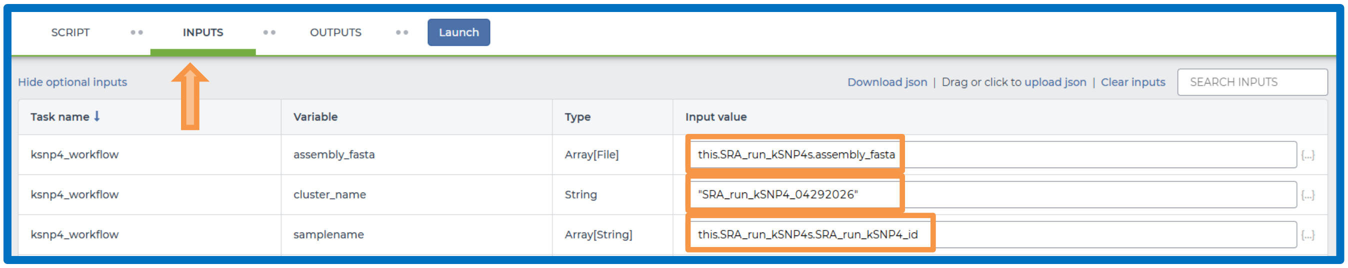
☑ Save a new set for selected SRA_run_kSNP4_sets.

Selected SRA_run_kSNP4s will be saved as a new SRA_run_kSNP4_set named:

kSNP4_PHB_2026-04-29T16-56-36

Figure 3

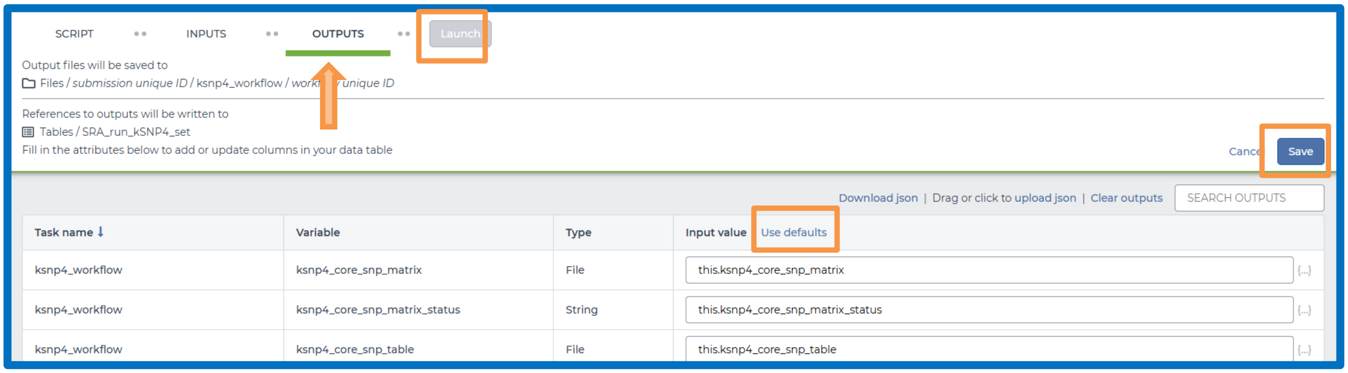
12. Click on *inputs* and set the first three attributes in the table to the following, respectively (Figure 4):
   a. *this.SRA_run_kSNP4s.assembly_fasta*
   b. *“SRA_run_kSNP4_Sal_5012026”*
      I. Where *Cluster_name* is a free text input used to label output files.
   c. *this. SRA_run_kSNP4s_SRA_run_kSNP4_id*
      I. Where *SRA_run_kSNP4* is the unique name of your data table in Terra.



Task name	Variable	Type	Input value
ksnp4_workflow	assembly_fasta	Array[File]	this.SRA_run_kSNP4s.assembly_fasta
ksnp4_workflow	cluster_name	String	"SRA_run_kSNP4_04292026"
ksnp4_workflow	samplename	Array[String]	this.SRA_run_kSNP4s.SRA_run_kSNP4_id

Figure 4

13. Add *"-ML"* to ksnp4_args input to output a maximum likelihood tree in the ksnp4_ml_tree output column that is compatible with analysis using boostrapping values
14. Specify outputs by clicking on the *outputs* tab and *use defaults* (Figure 5)
15. Click *save*
16. Launch the workflow by clicking *run analysis*; enter desired comments and click *launch.*



Output files will be saved to
Files / *submission unique ID* / ksnp4_workflow / *workf* *unique ID*

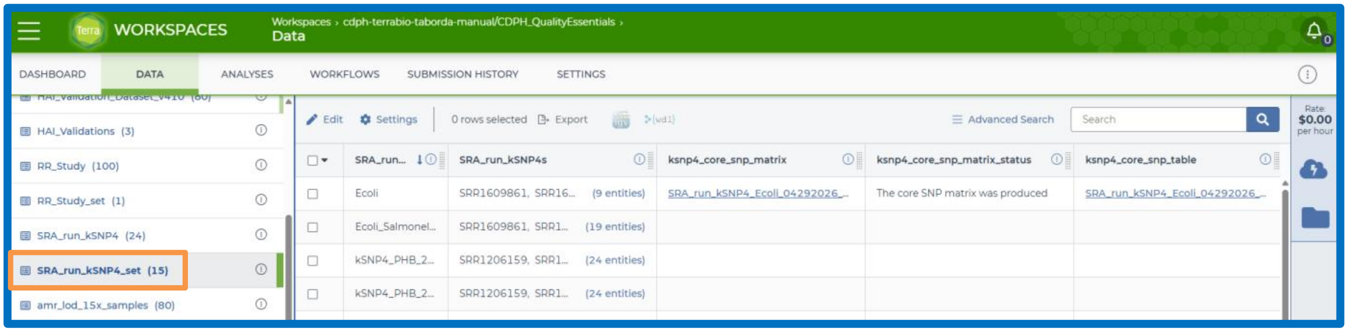
References to outputs will be written to
Tables / SRA_run_kSNP4_set
Fill in the attributes below to add or update columns in your data table

Task name	Variable	Type	Input value
ksnp4_workflow	ksnp4_core_snp_matrix	File	this.ksnp4_core_snp_matrix
ksnp4_workflow	ksnp4_core_snp_matrix_status	String	this.ksnp4_core_snp_matrix_status
ksnp4_workflow	ksnp4_core_snp_table	File	this.ksnp4_core_snp_table

Figure 5

4.2 VISUALIZING THE PHYLOGENETIC TREE USING iTOL

1. In the *data tab*, *select the SET data table* containing phylogenetic data (Figure 6)



HAI_Validation_Dataset_v410 (80)
HAI_Validations (3)
RR_Study (100)
RR_Study_set (1)
SRA_run_kSNP4 (24)
SRA_run_kSNP4_set (15)
amr_lod_15x_samples (80)

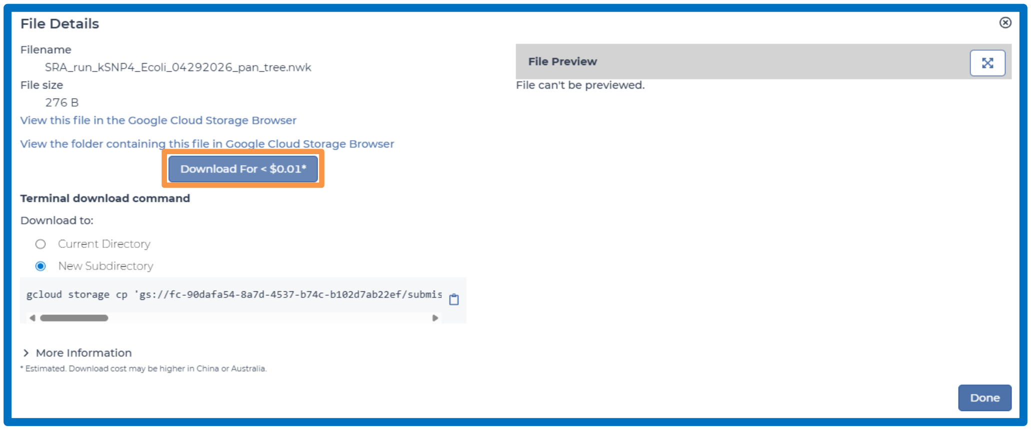
SRA_run...	SRA_run_kSNP4s	ksnp4_core_snp_matrix	ksnp4_core_snp_matrix_status	ksnp4_core_snp_table
Ecoli	SRR1609861, SRR16... (9 entities)	SRA_run_kSNP4_Ecoli_04292026_...	The core SNP matrix was produced	SRA_run_kSNP4_Ecoli_04292026_...
Ecoli_Salmonel...	SRR1609861, SRR1... (19 entities)			
kSNP4_PHB_2...	SRR1206159, SRR1... (24 entities)			
kSNP4_PHB_2...	SRR1206159, SRR1... (24 entities)			

Figure 6

☐	SRA_run...		ksnp4_pan_snp_matrix	ksnp4_pan_tree	ksnp4_snp_dists_version	ksnp4_snps
☑	Ecoli		SRA_run_kSNP4_Ecoli_04292026_...	SRA_run_kSNP4_Ecoli_04292026_...	snp-dists 0.8.2	SRA_run_kSNP4_Ecoli_0429
☑	Ecoli_Salmonel...					

Figure 7

2. In the column titled **ksnp4_pan_tree**, *click the hyperlinked file* for the relevant set of samples created when running the kSNP4 workflow, e.g. SRA_run_kSNP4_Sal_5012026 (Figure 7)
3. In the file details pop-up window, click *download for <$0.01* (Figure 8)
4. Open a new browser tab and navigate to *https://itol.embl.de/*



Filename
SRA_run_kSNP4_Ecoli_04292026_pan_tree.nwk
File size
276 B
View this file in the Google Cloud Storage Browser
View the folder containing this file in Google Cloud Storage Browser

File can't be previewed.

Terminal download command
Download to:
○ Current Directory
● New Subdirectory

```
gcloud storage cp 'gs://fc-90dafa54-8a7d-4537-b74c-b102d7ab22ef/submis
```

* Estimated. Download cost may be higher in China or Australia.

Figure 8

5. Under Annotate, click *upload a tree* (Figure 9)

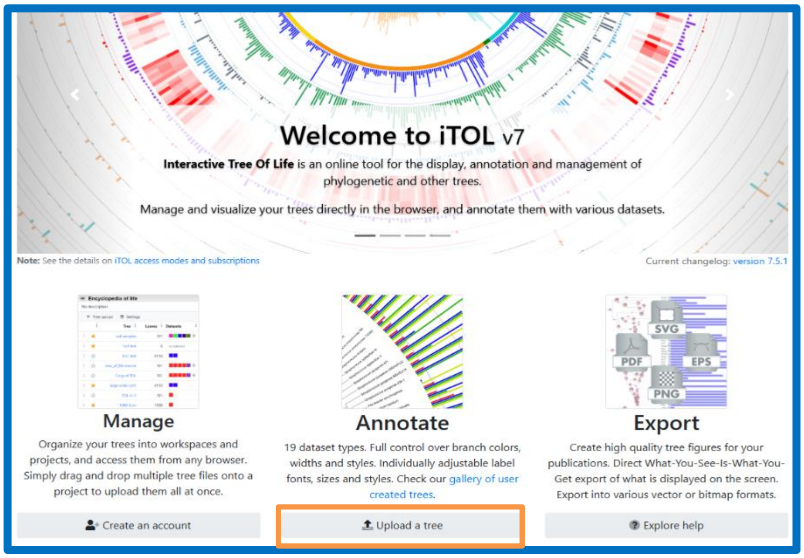
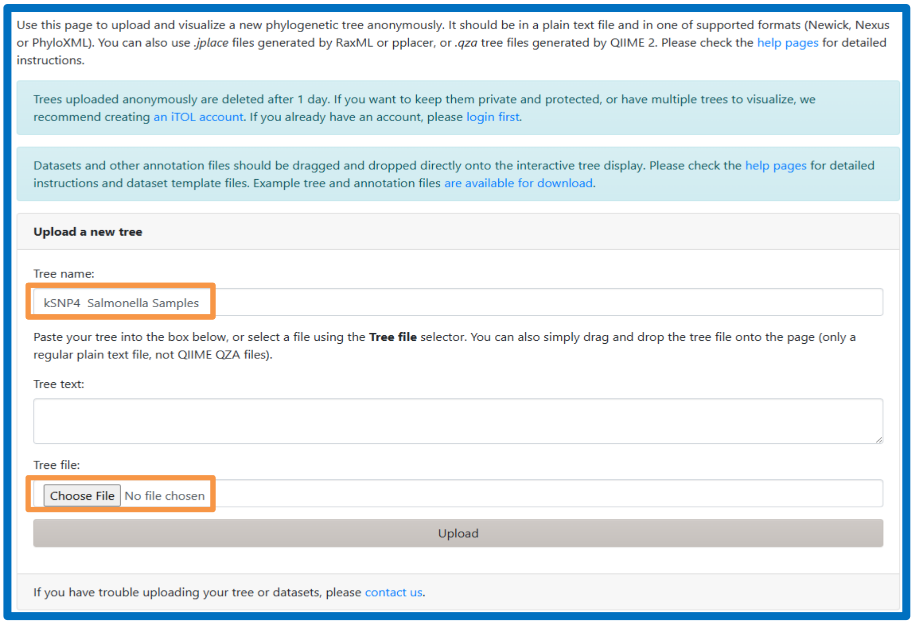


Figure 9

6. *Name the tree*, e.g. kSNP4 Salmonella Samples, *choose the tree file* downloaded from Terra, and click *upload* (Figure 10)



Use this page to upload and visualize a new phylogenetic tree anonymously. It should be in a plain text file and in one of supported formats (Newick, Nexus or PhyloXML). You can also use *.jplace* files generated by RaxML or pplacer, or *.qza* tree files generated by QIIME 2. Please check the help pages for detailed instructions.

Trees uploaded anonymously are deleted after 1 day. If you want to keep them private and protected, or have multiple trees to visualize, we recommend creating an iTOL account. If you already have an account, please login first.

Datasets and other annotation files should be dragged and dropped directly onto the interactive tree display. Please check the help pages for detailed instructions and dataset template files. Example tree and annotation files are available for download.

Upload a new tree

Tree name: kSNP4 Salmonella Samples

Paste your tree into the box below, or select a file using the **Tree file** selector. You can also simply drag and drop the tree file onto the page (only a regular plain text file, not QIIME QZA files).

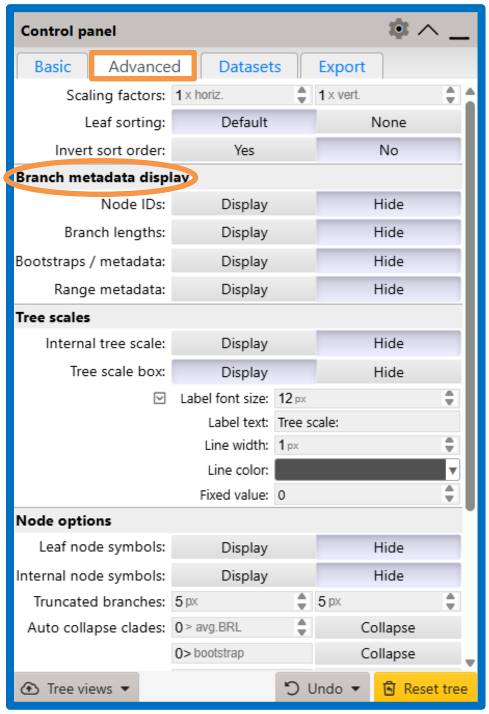
Tree text:

Tree file: No file chosen

If you have trouble uploading your tree or datasets, please contact us.

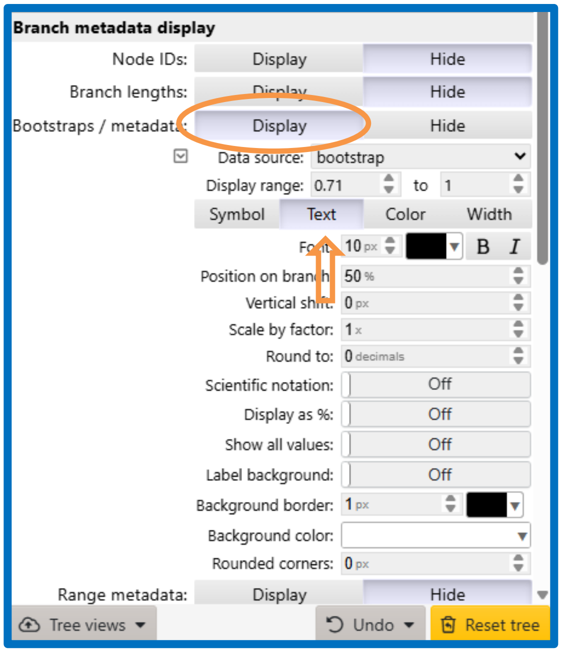
Figure 10

7. In the control panel, navigate to the *advanced tab*, select *display* for bootstraps/metadata, then select text to display the bootstrap values (Figure 11, Figure 12)
    a. *The font size, position, etc may also be adjusted for better visibility*



Scaling factors: 1 x horiz. 1 x vert.
Leaf sorting: Default None
Invert sort order: Yes No
Branch metadata display
Node IDs: Display Hide
Branch lengths: Display Hide
Bootstraps / metadata: Display Hide
Range metadata: Display Hide
Tree scales
Internal tree scale: Display Hide
Tree scale box: Display Hide
Label font size: 12 px
Label text: Tree scale:
Line width: 1 px
Line color:
Fixed value: 0
Node options
Leaf node symbols: Display Hide
Internal node symbols: Display Hide
Truncated branches: 5 px 5 px
Auto collapse clades: 0 > avg.BRL Collapse
0> bootstrap Collapse

Figure 11



Branch metadata display
Node IDs: Display Hide
Branch lengths: Display Hide
Bootstraps / metadata: Display Hide
Data source: bootstrap
Display range: 0.71 to 1
Symbol Text Color Width
Font: 10 px B *I*
Position on branch: 50 %
Vertical shift: 0 px
Scale by factor: 1 x
Round to: 0 decimals
Scientific notation: Off
Display as %: Off
Show all values: Off
Label background: Off
Background border: 1 px
Background color:
Rounded corners: 0 px
Range metadata: Display Hide

Figure 12

8. *View the phylogenetic tree* (Figure 13)

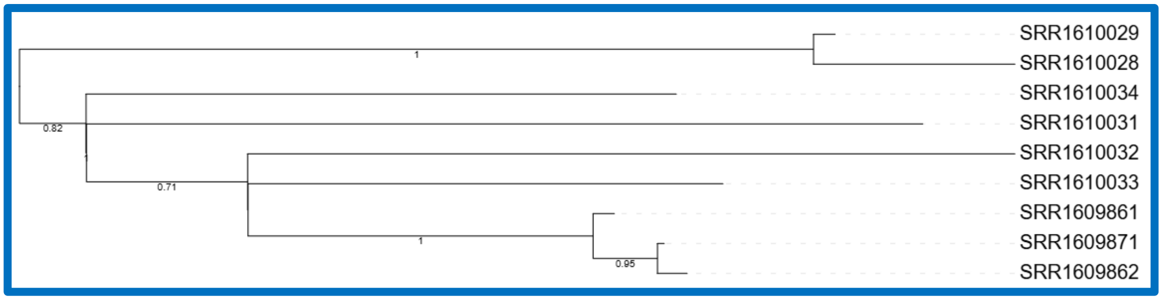


Figure 13

9. *See NGS Result Report Interpretation Guide[1] for more information on trees*
10. Follow lab-specific resulting and reporting procedures

5. QUALITY RECORDS

- Raw read and assembly files
- Sample read and assembly QC metrics, when applicable
- ksnp4_pan_tree file

6. TROUBLESHOOTING

- Consult with internal staff familiar with this procedure or contact support@theiagen.com

7. LIMITATIONS

- kSNP4 is not suitable for highly divergent genomes
- kSNP4 only takes into account SNP mutations and is not equipped to include other types of mutations like insertions/deletions in the phylogenetic inference
- Using a reference-free based tree building approach (vs. read mapping) does not take into account the context of where in the genome SNPs occur
- It is not possible to mask genome regions that may be prone to false positive mutations, like repetitive regions and phages
- It is not possible to mask recombination, which may skew the degree of relatedness if present

8. REFERENCES

1. "NGS Result Report Interpretation Guide" Centers for Disease Control and Prevention (CDC), URL: https://www.cdc.gov/labquality/qms-tools-and-resources.html?textFilter=result.

9. REVISION HISTORY

Revision	Version	Release Date
Document creation	1	05/2026

10. APPENDICES

10.1 IMPORTING THE kSNP4 WORKFLOW

1. In the *Terra workspace* of interest, open the *workflows* tab and click *find a workflow* (Figure 14)

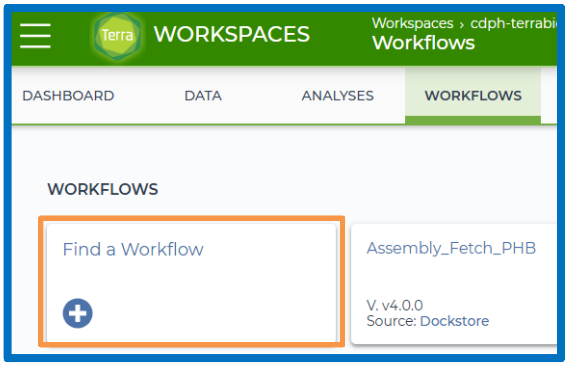


Figure 14

2. At the bottom of the pop-up window, click *dockstore* (Figure 15)


Find a workflow

Dockstore.org

A community repository of best practice workflows that offers integration with GitHub.

Terra Workflow Repository

A repository of WDL workflows that offers private workflows hosted in the platform.

Curated collections from our community:

GATK Best Practices
WDL Analysis Research Pipelines
Long Read Pipelines
Viral Genomics

Visit our documentation to learn how to import and configure your workflow, as well as how to save time and money.

Figure 15

3. Click *Organizations* tab on the top right corner of the (Figure 16)

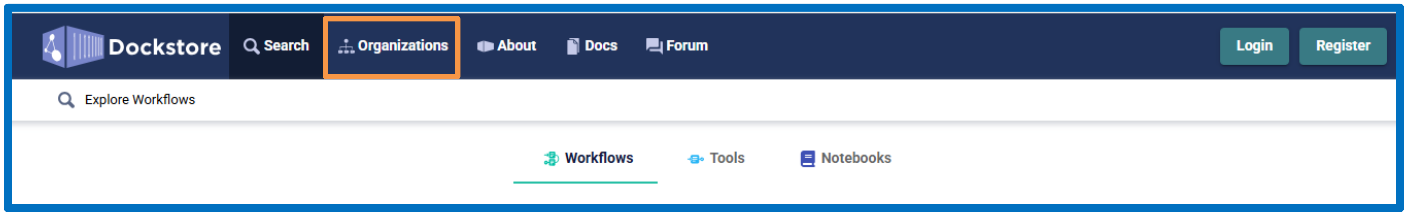


Figure 16

4. Locate *Theiagen Genomics* and click *View*(Figure 17)

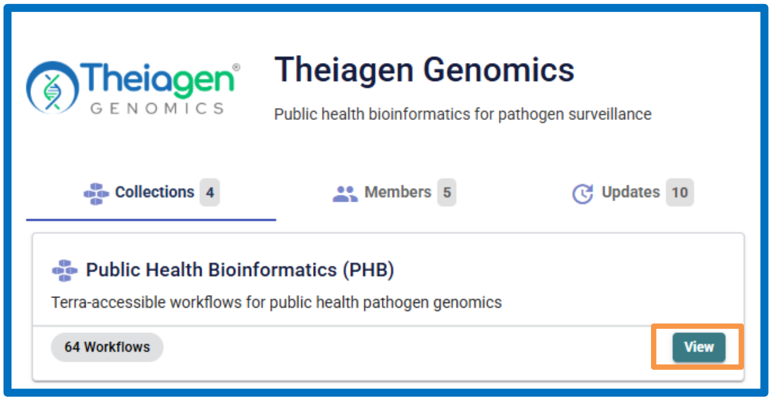
Theiagen Genomics

Public health bioinformatics for pathogen surveillance

https://www.theiagen.com
Denver, CO USA
joel.sevinsky@theiagen.com

Figure 17

5. Select *Public Health Bioinformatics (PHB)* collections by clicking *View* (Figure 18)
6. Locate the workflow "*github.com/theiagen/public_health_bioinformatics/kSNP4_PHB*" and click *View* (Figure 19)



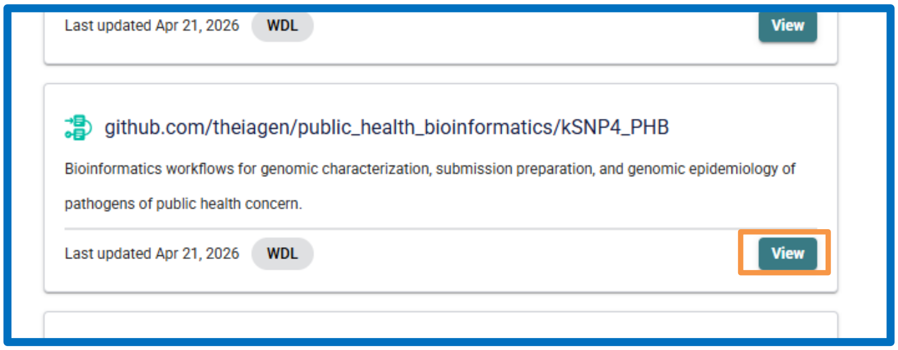
Theiagen Genomics

Public health bioinformatics for pathogen surveillance

Public Health Bioinformatics (PHB)

Terra-accessible workflows for public health pathogen genomics

Figure 18



github.com/theiagen/public_health_bioinformatics/kSNP4_PHB

Bioinformatics workflows for genomic characterization, submission preparation, and genomic epidemiology of pathogens of public health concern.

Figure 19

7. Click *Terra* on the right-hand side to launch the workflow in Terra (Figure 20)

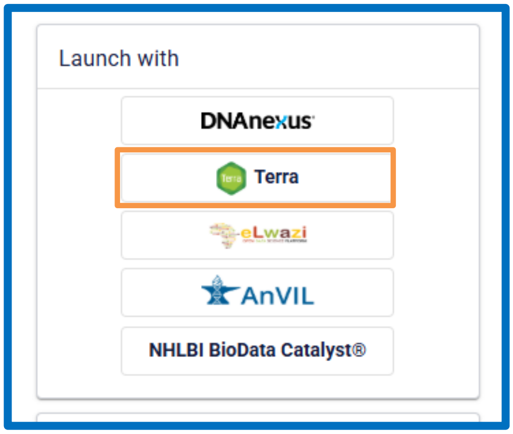
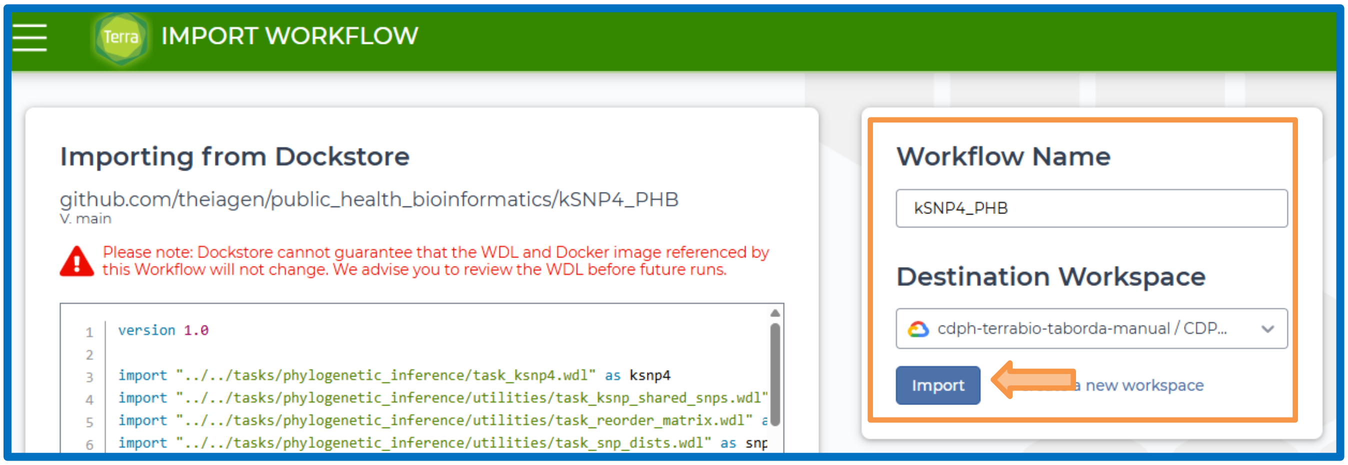


Figure 20

8. *Choose the workspace* destination in the dropdown and click *import* (Figure 21)



github.com/theiagen/public_health_bioinformatics/kSNP4_PHB
V. main

Please note: Dockstore cannot guarantee that the WDL and Docker image referenced by this Workflow will not change. We advise you to review the WDL before future runs.

```
version 1.0

import "../../tasks/phylogenetic_inference/task_ksnp4.wdl" as ksnp4
import "../../tasks/phylogenetic_inference/utilities/task_ksnp_shared_snps.wdl"
import "../../tasks/phylogenetic_inference/utilities/task_reorder_matrix.wdl" a
import "../../tasks/phylogenetic_inference/utilities/task_snp_dists.wdl" as snp
```

Workflow Name: kSNP4_PHB

Destination Workspace: cdph-terrabio-taborda-manual / CDP...

Figure 21