Analyzing Bacterial Data in Terra using Theiagen's TheiaProk Illumina PE Workflow	
Document TG-TP-PE, Version 4	
Date:	Workflow Version:
5/1/2025	PHB v3.0.0

1. PURPOSE/SCOPE

To standardize the process of running and analyzing bacterial isolates' next generation sequencing (NGS) data using Theiagen's TheiaProk Illumina PE workflow in Terra to perform genome assembly, QC, and characterization for predicted taxonomy, serotype/serogroup, sequence type (ST), AMR profile, and plasmid content. Additional analyses are optional in TheiaProk but are not addressed herein. Acceptable data types include Illumina paired end (PE) raw read file format. Lab-specific QC metrics and acceptance criteria should be established to ensure the integrity of the end-to-end NGS test system. Read the documentation here.

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-03 or TG-TER-04
- Theiagen's TheiaProk_Illumina_PE_PHB Workflow in Terra, see Appendix 10.1

IMPORTANT NOTES

- Metadata column headers and workflow input text indicated in *gray* in this SOP are customizable; *black* is required text
- Terra data table column headers become available as workflow inputs when running workflows, search for them in workflow input dropdowns using the prefix *this.* to filter
- Filter for workspace data and files in workflow input dropdowns using the prefix *workspace.*

3. RELATED DOCUMENTS

Document Number	Document Name
TG-TER-03	Getting Started in Terra: Importing Reads, Metadata, Workflows, and More
TG-TER-04	Linking BaseSpace and Importing BaseSpace Reads to Terra Workspace

4. PROCEDURE

4.1 RUNNING THE THEIAPROK WORKFLOW

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

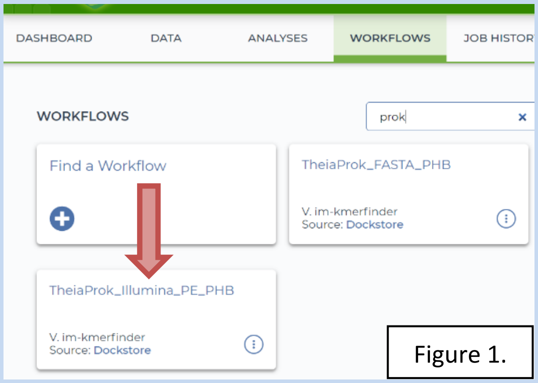
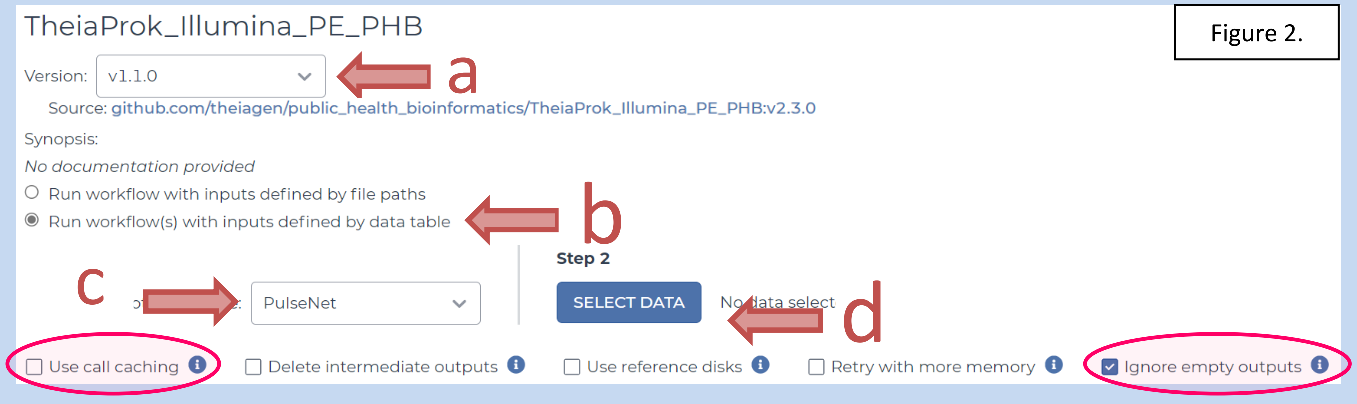


Figure 1.



TheiaProk_Illumina_PE_PHB

Version: v1.1.0

Source: github.com/theiagen/public_health_bioinformatics/TheiaProk_Illumina_PE_PHB:v2.3.0

Synopsis:

No documentation provided

○ Run workflow with inputs defined by file paths

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

PulseNet

Step 2

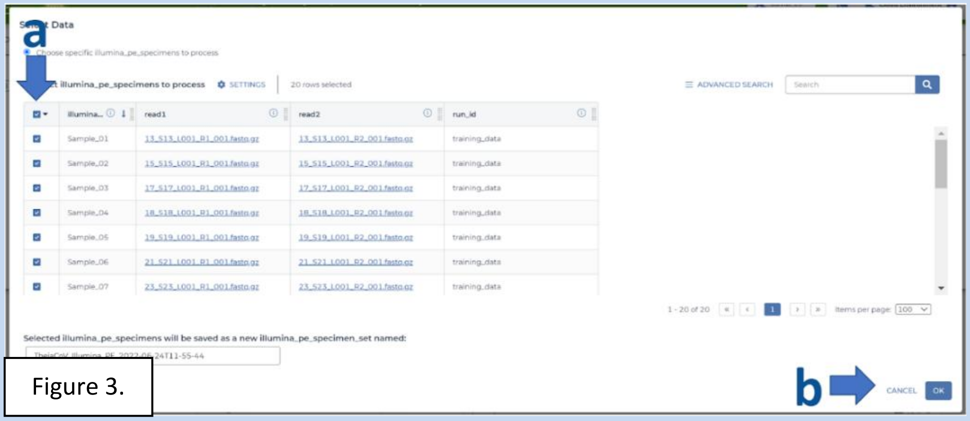
SELECT DATA

No data select

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

Figure 2.

3. Choose the latest version of the workflow in the *version dropdown field*, or the workflow version that was used during internal assay validation (Fig 2, a)
4. Select the second bullet to *run workflow(s) with inputs defined by data table* (Fig 2, b)
5. Select the relevant data table name under the *select data table* dropdown (Fig 2, c)
6. Click *select data* (Fig 2, d)
7. In the pop-up window *select the checkbox* for each sample to be included in the analysis (Fig 3)
   a. *The checkbox at the top may be used to select all samples listed on the page*
   b. *Click the down arrow and select all to process all specimens*
   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*
   d. Scroll to the bottom and click *ok*
8. Uncheck call caching (Fig 2)
9. *Optional: Check the box to* ignore empty outputs (Fig 2)



Select Data

● Choose specific illumina_pe_specimens to process

illumina_pe_specimens to process ⚙ SETTINGS | 20 rows selected

☰ ADVANCED SEARCH Search

☑	illumina_...	read1	read2	run_id
☑	Sample_01	13_S13_L001_R1_001.fastq.gz	13_S13_L001_R2_001.fastq.gz	training_data
☑	Sample_02	15_S15_L001_R1_001.fastq.gz	15_S15_L001_R2_001.fastq.gz	training_data
☑	Sample_03	17_S17_L001_R1_001.fastq.gz	17_S17_L001_R2_001.fastq.gz	training_data
☑	Sample_04	18_S18_L001_R1_001.fastq.gz	18_S18_L001_R2_001.fastq.gz	training_data
☑	Sample_05	19_S19_L001_R1_001.fastq.gz	19_S19_L001_R2_001.fastq.gz	training_data
☑	Sample_06	21_S21_L001_R1_001.fastq.gz	21_S21_L001_R2_001.fastq.gz	training_data
☑	Sample_07	23_S23_L001_R1_001.fastq.gz	23_S23_L001_R2_001.fastq.gz	training_data

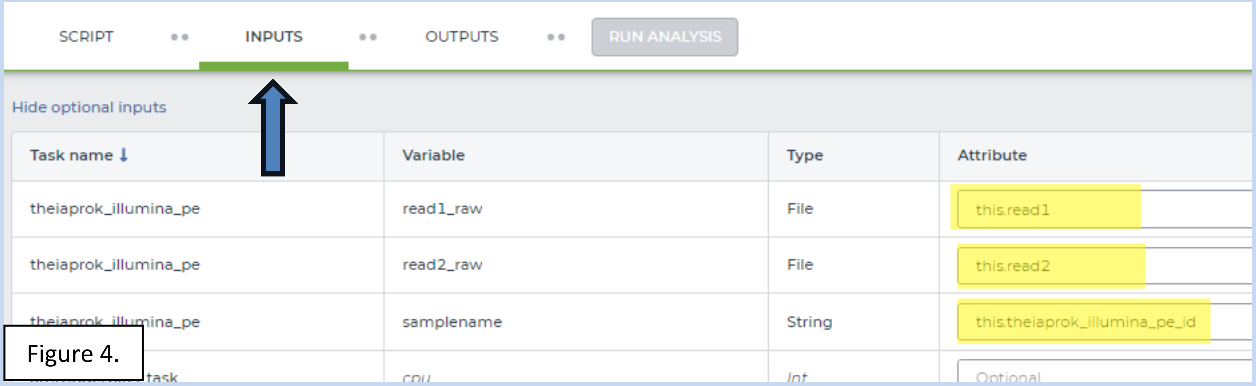
1 - 20 of 20 « ‹ 1 › » Items per page: 100

Selected illumina_pe_specimens will be saved as a new illumina_pe_specimen_set named:

TheiaCoV_Illumina_PE_2022-06-24T11-55-44

Figure 3.

10. Click on *inputs* and set the first three attributes in the table to the following, respectively (Fig 4):
    e. *this.read1*
    f. *this.read2*
    g. *this.theiaprok_illumina_pe_id*
        i. Where *theiaprok_illumina_pe* is the unique name of your data table in Terra

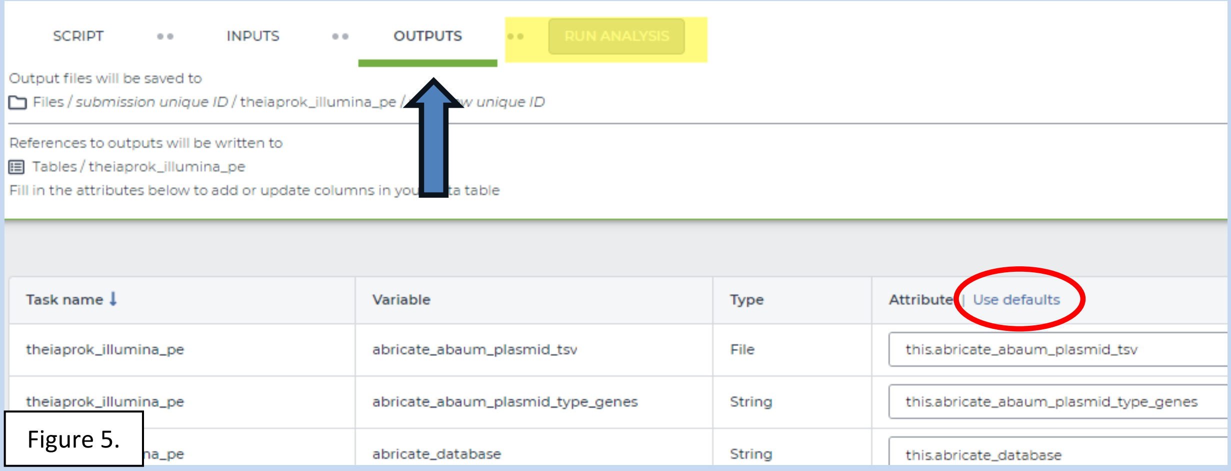


Hide optional inputs

Task name ↓	Variable	Type	Attribute
theiaprok_illumina_pe	read1_raw	File	this.read1
theiaprok_illumina_pe	read2_raw	File	this.read2
theiaprok_illumina_pe	samplename	String	this.theiaprok_illumina_pe_id
task	cpu	Int	Optional

Figure 4.

11. Specify outputs by clicking on the *outputs* tab and *use defaults* (Fig 5)
12. Click *save*
13. Launch the workflow by clicking *run analysis*; enter desired comments and click *launch*



Output files will be saved to
Files / *submission unique ID* / theiaprok_illumina_pe / *w unique ID*

References to outputs will be written to
Tables / theiaprok_illumina_pe
Fill in the attributes below to add or update columns in you ta table

Task name ↓	Variable	Type	Attribute | Use defaults
theiaprok_illumina_pe	abricate_abaum_plasmid_tsv	File	this.abricate_abaum_plasmid_tsv
theiaprok_illumina_pe	abricate_abaum_plasmid_type_genes	String	this.abricate_abaum_plasmid_type_genes
na_pe	abricate_database	String	this.abricate_database

Figure 5.

4.2 RAW READ AND ASSEMBLY QUALITY ASSESSMENT

1. Follow all quality assessment procedures specified by the instrument manufacturer, sequencing program (PulseNet, GenomeTrakr, etc), and those determined during internal validation procedures, as appropriate
2. Raw read data quality assessment may include looking at parameters such as average read quality scores; these should be determined during validation activities
3. Assembly-level quality assessment may include evaluating outputs such as average coverage, assembly length, contig number, etc; these should be determined during validation activities

4.3 VIEWING EXAMPLE QUALITY METRICS IN TERRA

1. In the *data tab* of the Terra workspace containing TheiaProk results, *open the relevant data table*
2. Click *Select Columns* above the data table, select *none* (Fig 6)
3. *Select columns* to view, as appropriate:
   a. *assembly_length*
   b. *combined_mean_q_clean*
   c. *est_coverage_clean*
   d. *number_contigs*
   e. *Optional: save this column group for future use by clicking the* save this column selection *field,* naming it *(e.g. QC), and clicking* save (Fig 6)
4. Click *done*

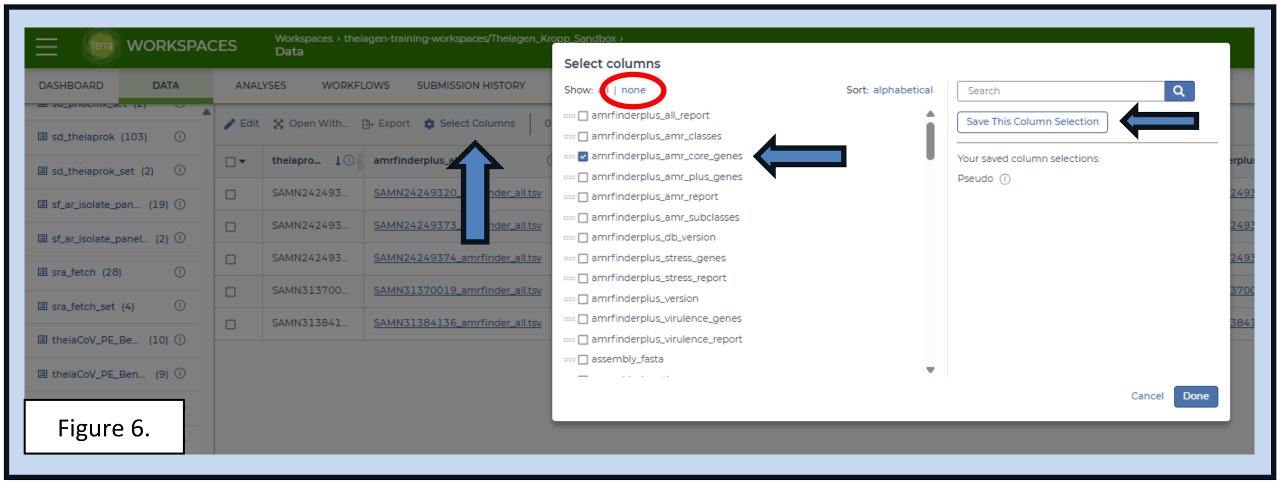


Figure 6.

sd_theiaprok (103)
sd_theiaprok_set (2)
sf_ar_isolate_pan... (19)
sf_ar_isolate_panel... (2)
sra_fetch (28)
sra_fetch_set (4)
theiaCoV_PE_Be... (10)
theiaCoV_PE_Ben... (9)

theiapro...	amrfinderplus_al...
SAMN242493...	SAMN24249320_...nder_all.tsv
SAMN242493...	SAMN24249373_...nder_all.tsv
SAMN242493...	SAMN24249374_amrfinder_all.tsv
SAMN313700...	SAMN31370019_amrfinder_all.tsv
SAMN313841...	SAMN31384136_amrfinder_all.tsv

Select columns

Show: all | none   Sort: alphabetical

□ amrfinderplus_all_report
□ amrfinderplus_amr_classes
☑ amrfinderplus_amr_core_genes
□ amrfinderplus_amr_plus_genes
□ amrfinderplus_amr_report
□ amrfinderplus_amr_subclasses
□ amrfinderplus_db_version
□ amrfinderplus_stress_genes
□ amrfinderplus_stress_report
□ amrfinderplus_version
□ amrfinderplus_virulence_genes
□ amrfinderplus_virulence_report
□ assembly_fasta

Your saved column selections:
Pseudo

5. Compare QC metrics to relevant acceptance criteria to determine pass/fail calls for each sample
6. For samples not passing QC metrics, re-sequence
   a. Failed QC samples may proceed to analysis at the discretion of the laboratory
7. For samples passing QC metrics, continue to analysis section 4.4

4.4 **GENOMIC CHARACTERIZATION**

1. Navigate to the *data tab* of the Terra workspace containing bacterial data of interest
2. *Open the data table* by clicking on the name of the data table in the left sidebar
3. Click *Select Columns* above the data table, select *none* (Fig 6)
4. *Select columns*, as applicable:

a. *amrfinderplus_amr_core_genes*

b. *gambit_predicted_taxon*

c. *plasmidfinder_results*

d. For serotype and serogroup results:

i. *ectyper_predicted_serotype*: serotype predicted by ECTyper for *Escherichia coli*

ii. *hicap_serotype*: serotype predicted by Hicap for *Haemophilus influenzae*

iii. *kaptive_k_type*: Kaptive predicted K type for *Acinetobacter baumannii*

iv. *kleborate_ktype*: Kleborate predicted K type (capsule) for *Klebsiella spp*. serotyping

v. *kleborate_otype*: Kleborate predicted O type (LPS) for *Klebsiella spp.* serotyping

vi. *lissero_serotype*: serotype predicted by LisSero for *Listeria monocytogenes*

vii. *meningotype_serogroup*: serogroup predicted by meningotype for *Neisseria meningitidis*

viii. *pasty_serogroup*: serogroup predicted by Pasty for *Pseudomonas aeruginosa*

ix. *seqsero2_predicted_serotype*: serotype predicted by SeqSero2 for *Salmonella spp.*

x. *seroba_serotype*: serotype predicted by SeroBA for *Streptococcus pneumoniae*

xi. *seroba_ariba_serotype*: serotype predicted by ARIBA using SeroBA for *Streptococcus pneumoniae*

xii. *serotypefinder_serotype*: serotype predicted by SerotypeFinder for *E. coli* and *Shigella spp.*

xiii. *shigatyper_predicted_serotype*: serotype predicted by ShigaTyper for *Shigella sonnei*

xiv. *shigeifinder_serotype*: serotype predicted by ShigEiFinder for *Shigella sonnei*

xv. *sistr_predicted_serotype*: serotype predicted by SISTR for *Salmonella spp.*

xvi. *sistr_serogroup*: serogroup predicted by SISTR for *Salmonella spp.*

xvii. *srst2_vibrio_serogroup*: O1 and O139 serotype prediction by SRST2 for *Vibrio spp.*

e. For sequence type (ST) results:

i. *kleborate_mlst_sequence_type*: Kleborate predicted ST for *Klebsiella spp.*

ii. *legsta_predicted_sbt*: Legsta predicted sequence-based typing for *Legionella pneumophila*

iii. *ngmaster_ngmast_sequence_type*: Ngmast predicted ST for *Neisseria gonorrohoeae*

iv. *ngmaster_ngstar_sequence_type*: Ngstar predicted ST for *Neisseria gonorrohoeae*

v. *sistr_serotype_cgmlst*: SISTR predicted ST for *Salmonella spp.*

vi. *ts_mlst_predicted_st*: Torsten Seemann predicted ST for bacterial 7-gene MLST

f. *Optional: save this column group for future use by clicking the* save this column selection *field,* naming it *(e.g. PulseNet_Results), and clicking* save (Fig 6)

5. Click *done*
6. Determine the predicted results for each sample by viewing the respective columns, as applicable
7. Compare QC metrics to relevant acceptance criteria to determine pass/fail calls for each result, as applicable
8. Follow lab-specific resulting and reporting procedures, as applicable

5. QUALITY RECORDS

- Raw read files and assemblies
- Sample read, assembly, and result-specific QC metrics, when applicable
- Result-specific determinations

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

1. This workflow only runs on bacterial, Illumina PE NGS data
2. Poor base quality, short read length, and nonuniform sequencing depth can impact the ability to adequately perform de novo assembly and affect downstream tool predictions
3. Sequencing of mixed/contaminated cultures will affect the accuracy of result predictions
4. GAMBIT may misclassify certain *E. coli* samples as *Shigella* species; therefore, interpret results in combination with additional tests or tools to ensure confident *E. coli* identification.

8. REFERENCES

1. Timme, Ruth E et al. "Optimizing open data to support one health: best practices to ensure interoperability of genomic data from bacterial pathogens." One health outlook vol. 2,1 (2020): 20. doi:10.1186/s42522-020-00026-3
2. 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.
3. Theiagen Genomics Public Health Bioinformatics Workflow Documentation. https://theiagen.github.io/public_health_bioinformatics

9. REVISION HISTORY

Revision	Version	Release Date
Document creation	1	12/2023
Revision	2	3/2024
Minor edits to align workflow inputs throughout, minor formatting changes, added assembly as quality record, added appendix 10.1	3	2/2025
Minor edits to align with Terra interface, add new outputs for species typing, improve formatting	4	5/2025

10. APPENDICES

10.1 Find and Import the TheiaProk_Illumina_PE_PHB Workflow

1. Navigate to the *workflows tab* of the workspace (Fig 7).
2. Workspaces that already have the workflow can *select TheiaProk_Illumina_PE_PHB* (Fig 7) and proceed to Running the TheiaProk Workflow section of this SOP.
3. To import the workflow, click *find a workflow* (Fig 7).

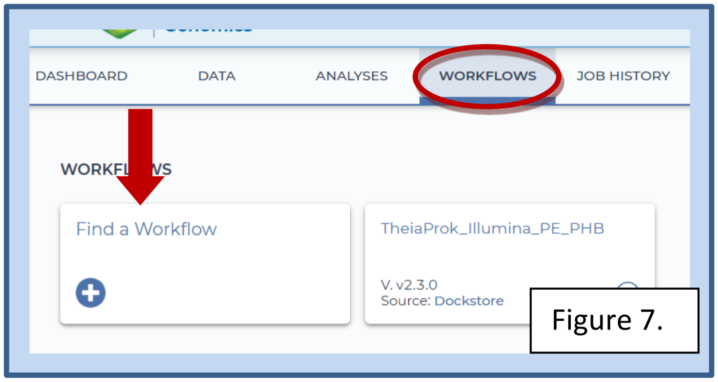
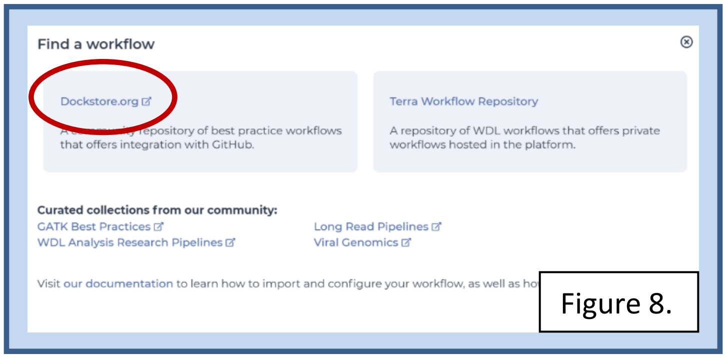


Figure 7.

4. In the pop-up window, click *Dockstore.org* (Fig 8).


Find a workflow
Dockstore.org
Terra Workflow Repository
Curated collections from our community:
GATK Best Practices
Long Read Pipelines
WDL Analysis Research Pipelines
Viral Genomics

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

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

Visit our documentation to learn how to import and configure your workflow, as well as ho

Figure 8.

5. Click *Organizations* in the banner at the top and *search for Theiagen* using the search box (Fig 9).

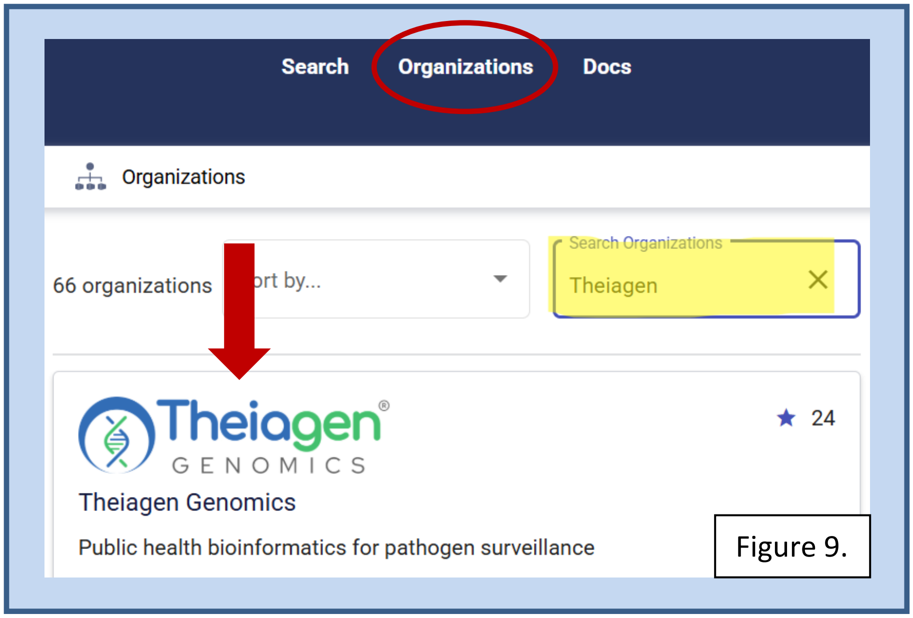
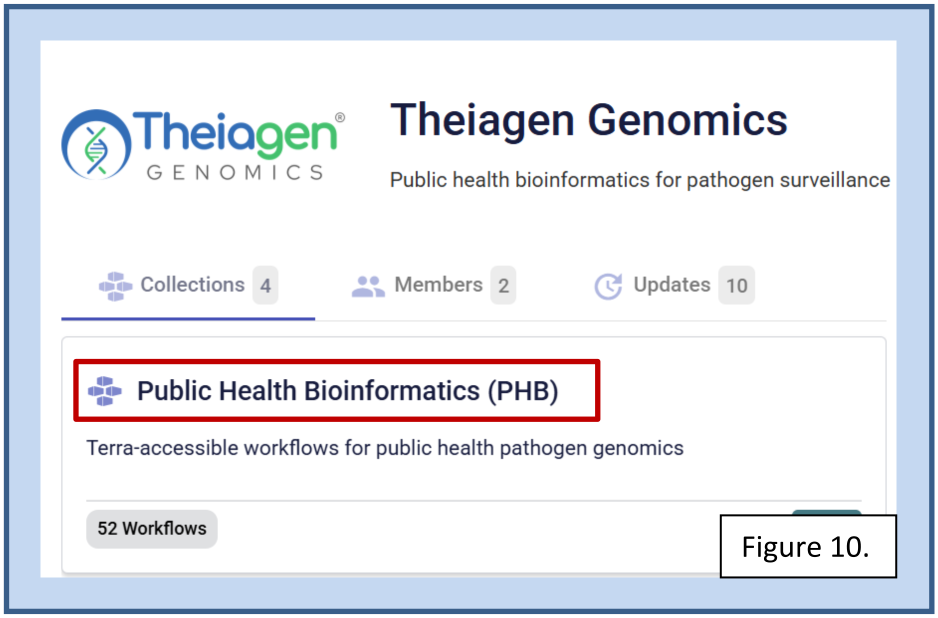


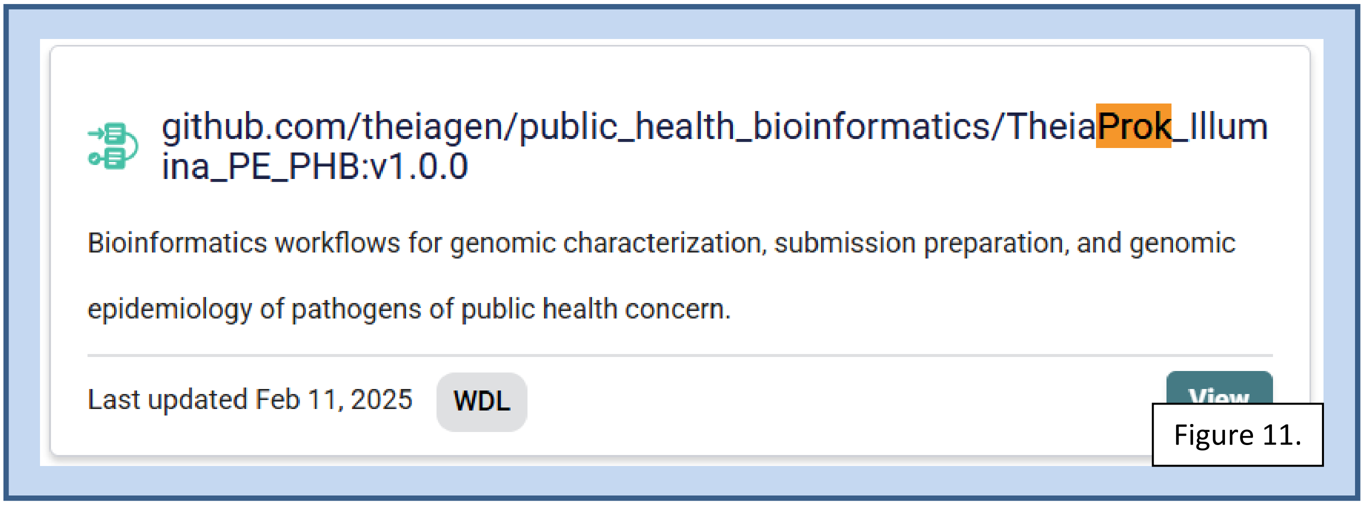
Figure 9.

6. Click the *Public Health Bioinformatics (PHB)* collection (Fig 10) and using *ctrl* + *f* on Windows search for "*Prok*."(Fig 11)



Theiagen GENOMICS
Theiagen Genomics
Public health bioinformatics for pathogen surveillance
Collections 4
Members 2
Updates 10
Public Health Bioinformatics (PHB)
Terra-accessible workflows for public health pathogen genomics
52 Workflows

Figure 10.



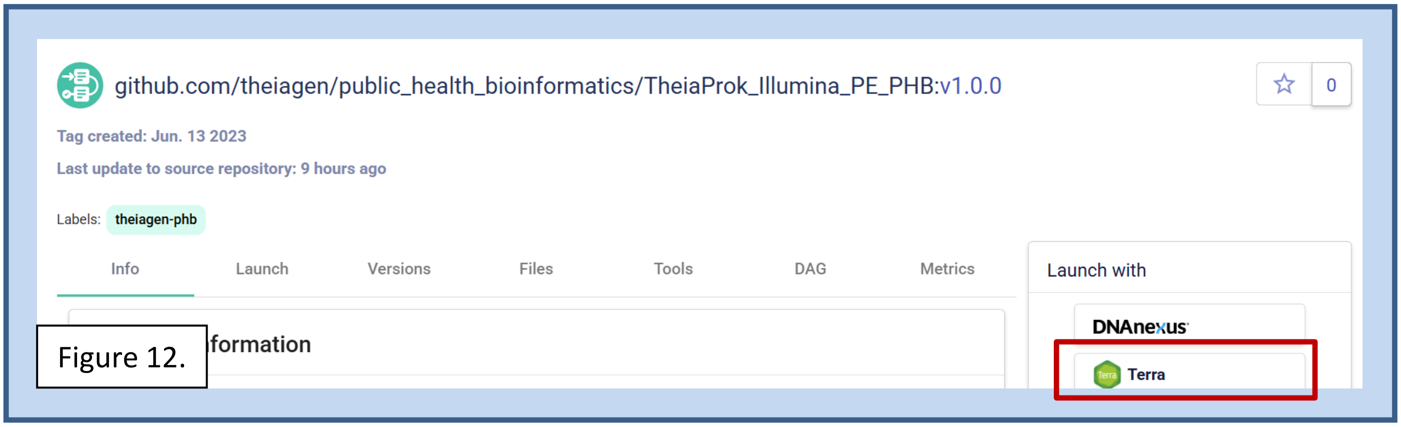
github.com/theiagen/public_health_bioinformatics/TheiaProk_Illumina_PE_PHB:v1.0.0

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

Last updated Feb 11, 2025

Figure 11.

7. Click on the *Terra icon* (Fig 12) to import the workflow into a Terra workspace.



github.com/theiagen/public_health_bioinformatics/TheiaProk_Illumina_PE_PHB:v1.0.0

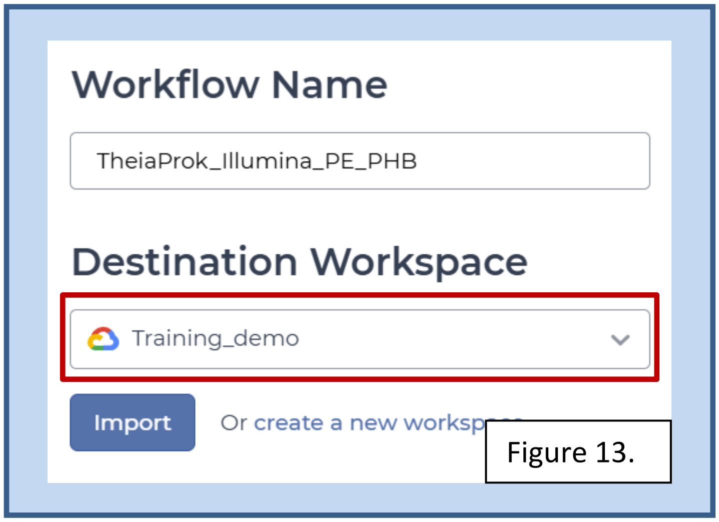
Tag created: Jun. 13 2023

Last update to source repository: 9 hours ago

Labels: theiagen-phb

Figure 12.

8. *Select the workspace* in the destination workspace dropdown field and click *Import* (Fig 13).



Workflow Name: TheiaProk_Illumina_PE_PHB

Destination Workspace: Training_demo

Figure 13.