



Analyzing Bacterial Data in Terra using Theiagen's TheiaProk Illumina PE Workflow

Document TG-TP-PE, Version 3

Date:

2/11/2025

Workflow Version:

PHB v2.3.0

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.

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	Uploading Local or SRA NGS- Data & Creating a Results Metadata Table in Terra
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)
3. Uncheck call caching (Fig 2)

Figure 1.



Analyzing Bacterial Data in Terra using Theiagen's TheiaProk Illumina PE Workflow

Document TG-TP-PE, Version 3

Date:

2/11/2025

Workflow Version:

PHB v2.3.0

TheiaProk_Illumina_PE_PHB

Version: v1.1.0 ↪ a

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 ↪ b

Step 2

C ↪ PulseNet ↪ Step 2

PulseNet ↪ Step 2

SELECT DATA ↪ Step 2

No data selected ↪ d

Use call caching ↪ e
 Delete intermediate outputs ↪ f
 Use reference disks ↪ g
 Retry with more memory ↪ h
 Ignore empty outputs ↪ i

Figure 2.

4. Optional: Check the box to ignore empty outputs (Fig 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 (Fig 2, a)
6. Select the second bullet to **run workflow(s) with inputs defined by data table** (Fig 2, b)
7. Select the relevant data table name under the **select data table** dropdown (Fig 2, c)
8. Click **select data** (Fig 2, d)
9. 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
 - b. 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
 - c. Scroll to the bottom and click **ok**

a

Select Data

Choose specific illumina_pe_specimens to process

20 rows selected

ADVANCED SEARCH Search

Selected illumina_pe_specimens will be saved as a new illumina_pe_specimen_set named:

illumina_pe_specimen	read1	read2	run_id
Sample_01	13_513_L001_R1_001.fasta.gz	13_513_L001_R2_001.fasta.gz	training_data
Sample_02	15_515_L001_R1_001.fasta.gz	15_515_L001_R2_001.fasta.gz	training_data
Sample_03	17_517_L001_R1_001.fasta.gz	17_517_L001_R2_001.fasta.gz	training_data
Sample_04	18_518_L001_R1_001.fasta.gz	18_518_L001_R2_001.fasta.gz	training_data
Sample_05	19_519_L001_R1_001.fasta.gz	19_519_L001_R2_001.fasta.gz	training_data
Sample_06	21_521_L001_R1_001.fasta.gz	21_521_L001_R2_001.fasta.gz	training_data
Sample_07	23_523_L001_R1_001.fasta.gz	23_523_L001_R2_001.fasta.gz	training_data

1 - 20 of 20 CANCEL OK

b



Analyzing Bacterial Data in Terra using Theiagen's TheiaProk Illumina PE Workflow

Document TG-TP-PE, Version 3

Date:

2/11/2025

Workflow Version:

PHB v2.3.0

10. Click on **inputs** and set the first three attributes in the table to the following, respectively (Fig 4):

- d. `this.read1`
- e. `this.read2`
- f. `this.PulseNet_id`

i. Where `PulseNet` is the unique name of your data table in Terra

Task name ↓	Variable	Type	Attribute
theiaprok_illumina_pe	read1_raw	File	<code>this.read1</code>
theiaprok_illumina_pe	read2_raw	File	<code>this.read2</code>
theiaprok_illumina_pe	samplename	String	<code>this.PulseNet_id</code>
Figure 4. task	CPU	Int	Optional

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**

Task name ↓	Variable	Type	Attribute Use defaults
theiaprok_illumina_pe	abriicate_abbaum_plasmid_tsv	File	<code>this.abriicate_abbaum_plasmid_tsv</code>
theiaprok_illumina_pe	abriicate_abbaum_plasmid_type_genes	String	<code>this.abriicate_abbaum_plasmid_type_genes</code>
Figure 5. na_pe	abriicate_database	String	<code>this.abriicate_database</code>



Analyzing Bacterial Data in Terra using Theiagen's TheiaProk Illumina PE Workflow

Document TG-TP-PE, Version 3

Date:

2/11/2025

Workflow Version:

PHB v2.3.0

4.2 RAW READ AND ASSEMBLY QUALITY ASSESSMENT

- Follow all quality assessment procedures specified by the instrument manufacturer, sequencing program (PulseNet, GenomeTrakr, etc), and those determined during internal validation procedures, as appropriate
- Raw read data quality assessment may include looking at parameters such as average read quality scores; these should be determined during validation activities
- 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

- In the **data tab** of the Terra workspace containing TheiaProk results, **open the relevant data table**
- View **settings** above the data table, select **none** (Fig 6)
- Select columns** to view, as appropriate:
 - assembly_length**
 - combined_mean_q_clean**
 - est_coverage_clean**
 - number_contigs**
- Optional:** save this column group for future use by clicking the **save this column selection** field, naming it (e.g. QC), and clicking **save**
- Click **done**
- Compare QC metrics to relevant acceptance criteria to determine pass/fail calls for each sample

Figure 6.



Analyzing Bacterial Data in Terra using Theiagen's TheiaProk Illumina PE Workflow

Document TG-TP-PE, Version 3

Date:

2/11/2025

Workflow Version:

PHB v2.3.0

6. For samples not passing QC metrics, resequence
 - 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. View [settings](#) 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. [shigefinder_serotype](#): serotype predicted by ShigEiFinder for *Shigella sonnei*
 - xv. [sistri_predicted_serotype](#): serotype predicted by SISTR for *Salmonella spp.*
 - xvi. [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 gonorrhoeae*



Analyzing Bacterial Data in Terra using Theiagen's TheiaProk Illumina PE Workflow

Document TG-TP-PE, Version 3

Date:

2/11/2025

Workflow Version:

PHB v2.3.0

- iv. `ngmaster_ngstar_sequence_type`: Ngstar predicted ST for *Neisseria gonorrhoeae*
 - v. `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`
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

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

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



Analyzing Bacterial Data in Terra using Theiagen's TheiaProk Illumina PE Workflow

Document TG-TP-PE, Version 3

Date:

2/11/2025

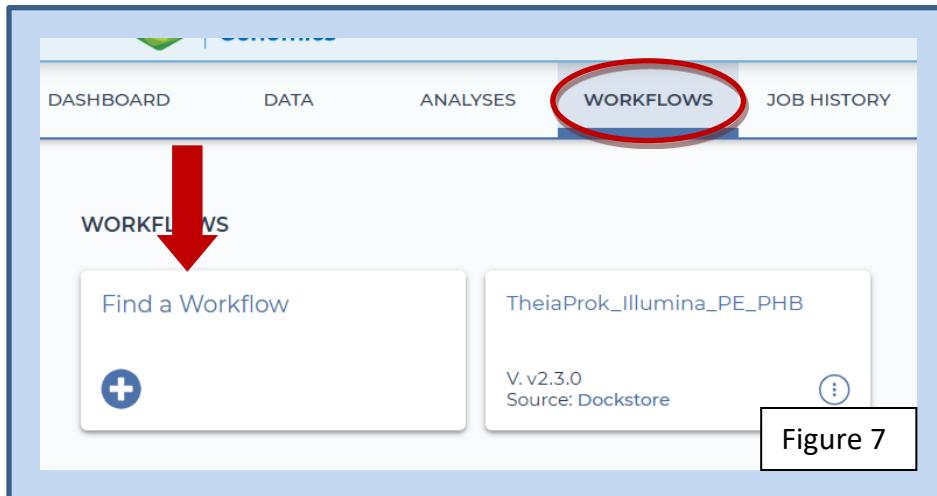
Workflow Version:

PHB v2.3.0

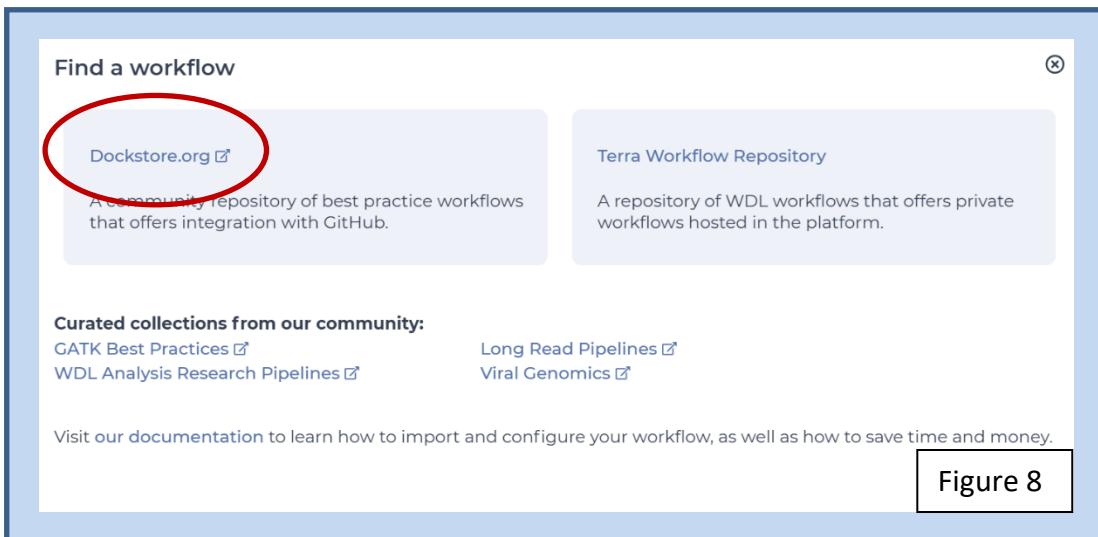
10. APPENDICES

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



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



Analyzing Bacterial Data in Terra using Theiagen's TheiaProk Illumina PE Workflow

Document TG-TP-PE, Version 3

Date:

2/11/2025

Workflow Version:

PHB v2.3.0

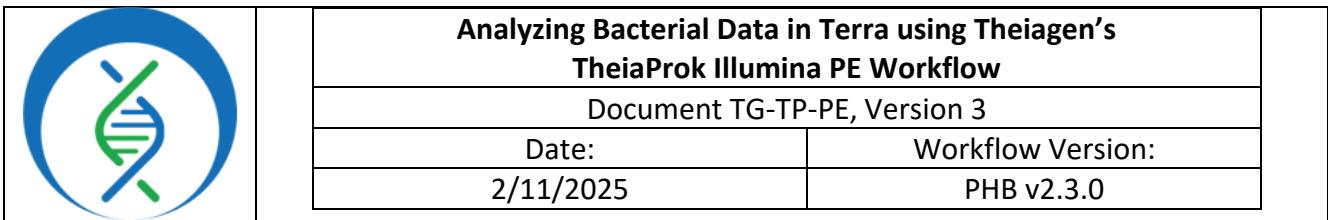
The screenshot shows the Terra platform's organization search results. The 'Organizations' tab is highlighted with a red oval. A search bar at the top right contains the text 'Theiagen'. Below the search bar, there is a dropdown menu showing '66 organizations' and a 'Sort by...' option. A red arrow points downwards from the search bar towards the organization card for 'Theiagen Genomics'. The organization card includes the 'Theiagen GENOMICS' logo, the name 'Theiagen Genomics', a tagline 'Public health bioinformatics for pathogen surveillance', and a star rating of '24'.

Figure 9

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

The screenshot shows the 'Theiagen Genomics' organization page. At the top, the organization's logo and name are displayed, along with the tagline 'Public health bioinformatics for pathogen surveillance'. Below this, there are three navigation links: 'Collections 4', 'Members 2', and 'Updates 10'. A prominent feature is a collection card for 'Public Health Bioinformatics (PHB)', which is highlighted with a red box. The card includes the collection's name, a description 'Terra-accessible workflows for public health pathogen genomics', and a link to '52 Workflows'. A red arrow points downwards from the collection card towards the bottom of the page.

Figure 10



This screenshot shows a GitHub repository page. The repository name is 'github.com/theiagen/public_health_bioinformatics/TheiaProk_Illumina_PE_PHB:v1.0.0'. The description below the name reads: 'Bioinformatics workflows for genomic characterization, submission preparation, and genomic epidemiology of pathogens of public health concern.' A note at the bottom left says 'Last updated Feb 11, 2025'. To the right of the update date is a 'WDL' button. In the top right corner of the main content area is a 'View' button. A callout box labeled 'Figure 11' is positioned in the bottom right corner of the screenshot area.

7. Click on the **Terra icon** (Fig 11) to import the workflow into a Terra workspace.

This screenshot shows the same GitHub repository page as Figure 11. A red box highlights the 'Terra' icon in the 'Launch with' dropdown menu on the right side of the page. A callout box labeled 'Figure 12' is positioned in the bottom left corner of the screenshot area.

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

This screenshot shows a 'Import Workflow' dialog box. It has two main sections: 'Workflow Name' and 'Destination Workspace'. The 'Workflow Name' section contains a text input field with the value 'TheiaProk_Illumina_PE_PHB'. The 'Destination Workspace' section contains a dropdown menu with 'Training_demo' selected. Below the dropdown are two buttons: 'Import' and 'Or create a new workspace'. A callout box labeled 'Figure 13' is positioned in the bottom right corner of the dialog box.