	Analyzing Bacterial Data in Terra using Theiagen's TheiaProk ONT Workflow	
	Document TG-TP-ONT, Version 2	
	Date:	Workflow Version
	5/1/2025	PHB v3.0.0

1. PURPOSE/SCOPE

To standardize the process of analyzing bacterial next generation sequence (NGS) data using Theiagen's TheiaProk_ONT_PHB workflow in Terra to generate assemblies, quality control (QC) metrics, and determine 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 ONT 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 docs here.

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
- ONT raw sequencing read files uploaded to Terra workspace, see TG-TER-03
- Theaigen's TheiaCoV_ONT_PHB workflow in Terra, see TG-TER-03 appendix 9.2

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

4. PROCEDURE

4.1 CREATE A SAMPLE METADATA FILE (TSV FILE) FOR RAW READS OR SRA FETCH

1. In Excel, *create a list* containing the following sample information:
   - **a. For all analyses:**
     - i. Column 1 header (Fig 1): *BacterialONT_id* where *BacterialONT* is the name of the data table/group of samples to be analyzed
     - ii. List all *sample IDs* in column 1

	A	B
1	BacterialONT_id	reads
2	Sample_01	gs://fc-a0493432-4788-42e6-8b
3	Sample_02	gs://fc-a0493432-4788-42e6-8b
4	Sample_03	gs://fc-a0493432-4788-42e6-8b
5	Sample_04	gs://fc-a0493432-4788-42e6-8b

Figure 1: Raw Read Metadata File.

b. For analysis from raw sequencing reads (Fig 1)

i. Column 2 header: *reads*

ii. List the *full file paths* to read1 files in the cloud

c. For analysis using SRA fetch (Fig 2):

i. Column 2 header: *sra_accession*

BacterialONT_id	sra_accession	run_id
Sample_01	SRX17082331	SEQ137
Sample_07	SRX17082330	SEQ137
Sample_11	SRX17082329	SEQ137
		SEQ137

Figure 2: SRA Accession Metadata File.

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

e. 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

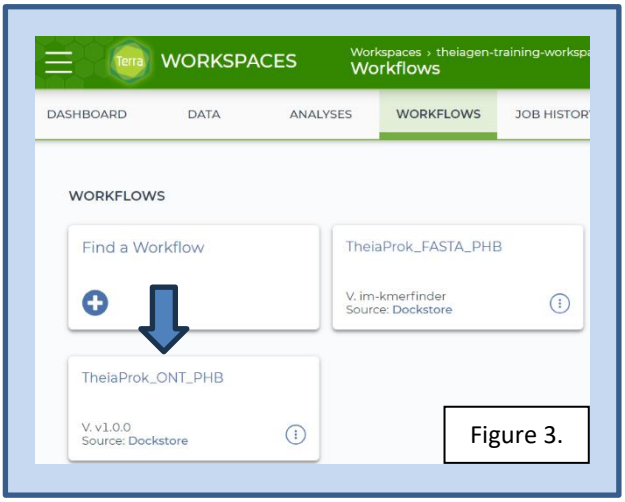

Figure 3.

4.2 RUNNING THE THEIAPROK WORKFLOW

1. Open Terra and navigate to the *workflows* tab of the workspace containing bacterial data
2. Select the *TheiaProk_ONT_PHB* workflow (Fig 3)
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 4, a)
4. Select the second bullet to *run workflow(s) with inputs defined by data table* (Fig 4, b)
5. Select the relevant data table under the *select data table* dropdown (Fig 4, c), e.g. BacterialONT
6. Click *select data* (Fig 4, d)

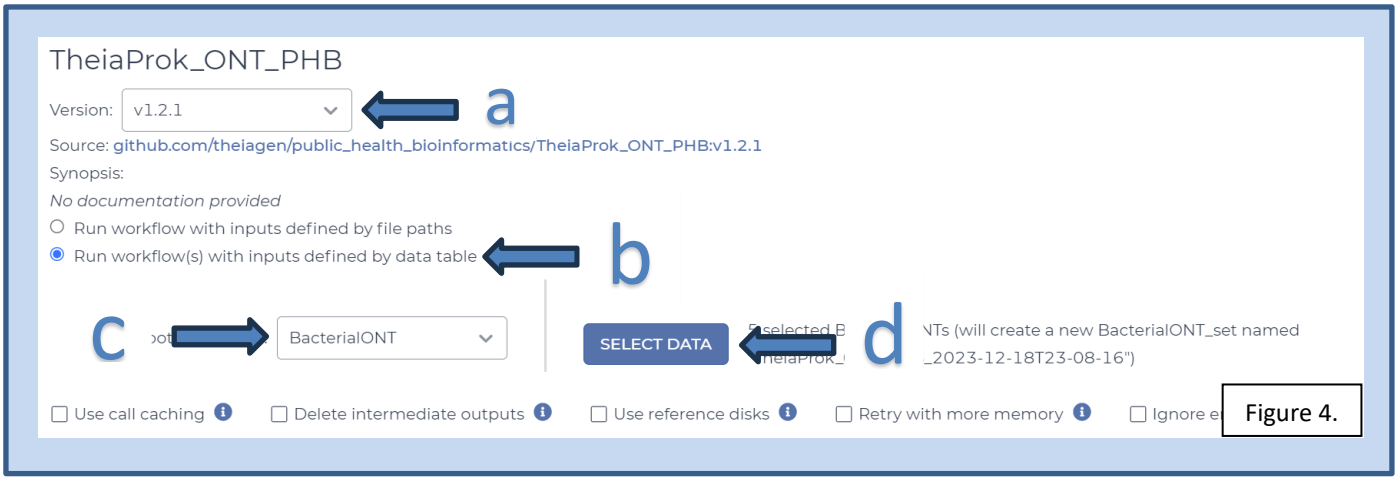

Figure 4.

7. In the pop-up window *select the checkbox* for each sample to be included in the analysis (Fig 5)
   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 The set of samples will be saved in the *BacterialONT_set* Terra data table and can be named for easier identification using the text field in the bottom left of this pop-up window (e.g. the group of samples could be saved as the run ID, HAI, etc)
   e Scroll to the bottom and click *ok*

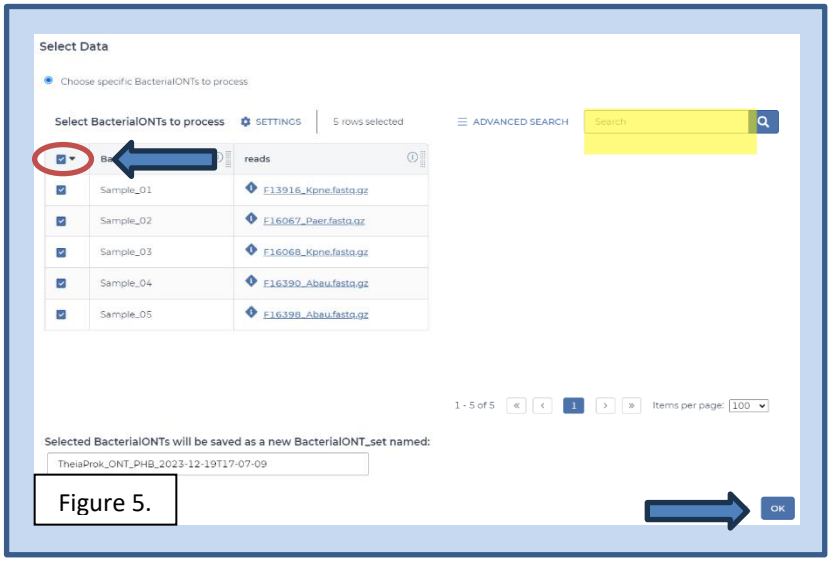

Figure 5.

8. Uncheck *Use call caching* (Fig 4)
9. Set the first two input attributes in the table to *this.reads* and *this.BacterialONT_id*, respectively (Fig 6) where:
   a *this.BacterialONT_id* is the unique name of your data table in Terra

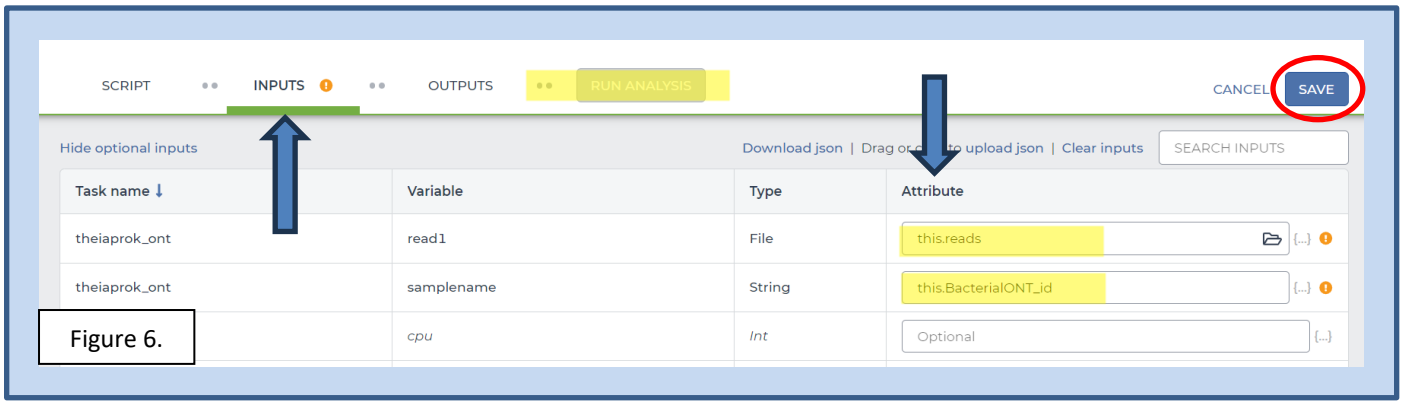

Figure 6.

10. Specify outputs by clicking on the *outputs* tab and *use defaults* (Fig 7)
11. Click *save*
12. Launch the workflow by clicking *run analysis* (Fig 7); enter desired comments and click *launch*

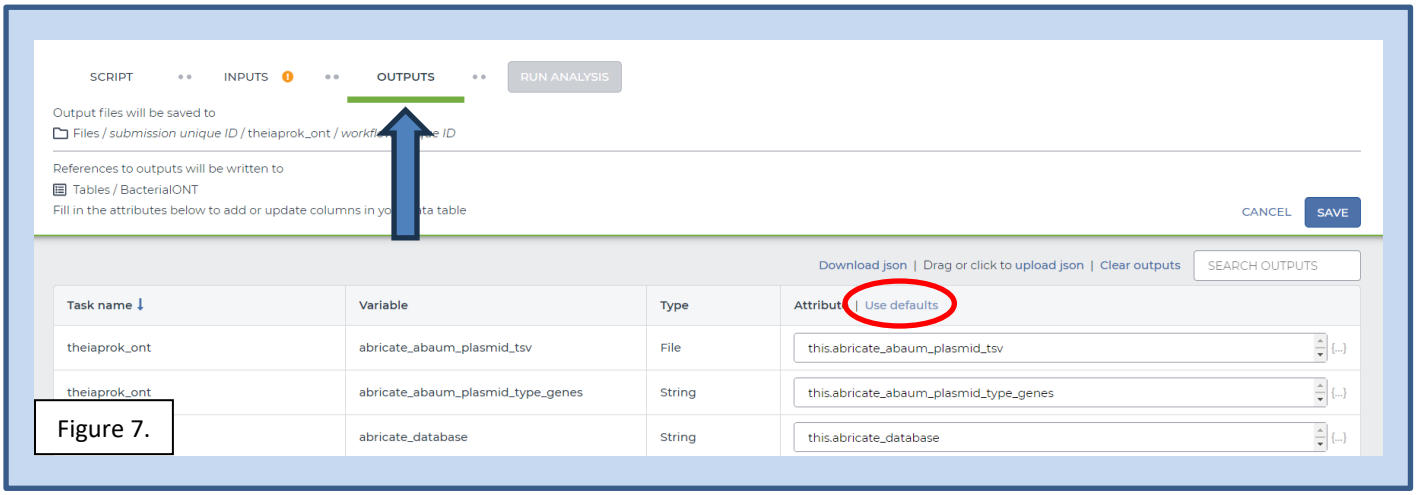

Figure 7.

4.3 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

4.4 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 8)
3. Select the appropriate columns:
   a. *assembly_length*
   b. *est_coverage_clean*
   c. *gambit_predicted_taxon*
   d. *nanoplot_r1_mean_q_clean*
   e. *number_contigs*
   f. *quast_gc_percent*
   g. *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 8)
4. Click *done*
5. Compare all metrics to relevant QC requirements and determine pass/fail calls for all samples

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 4.5

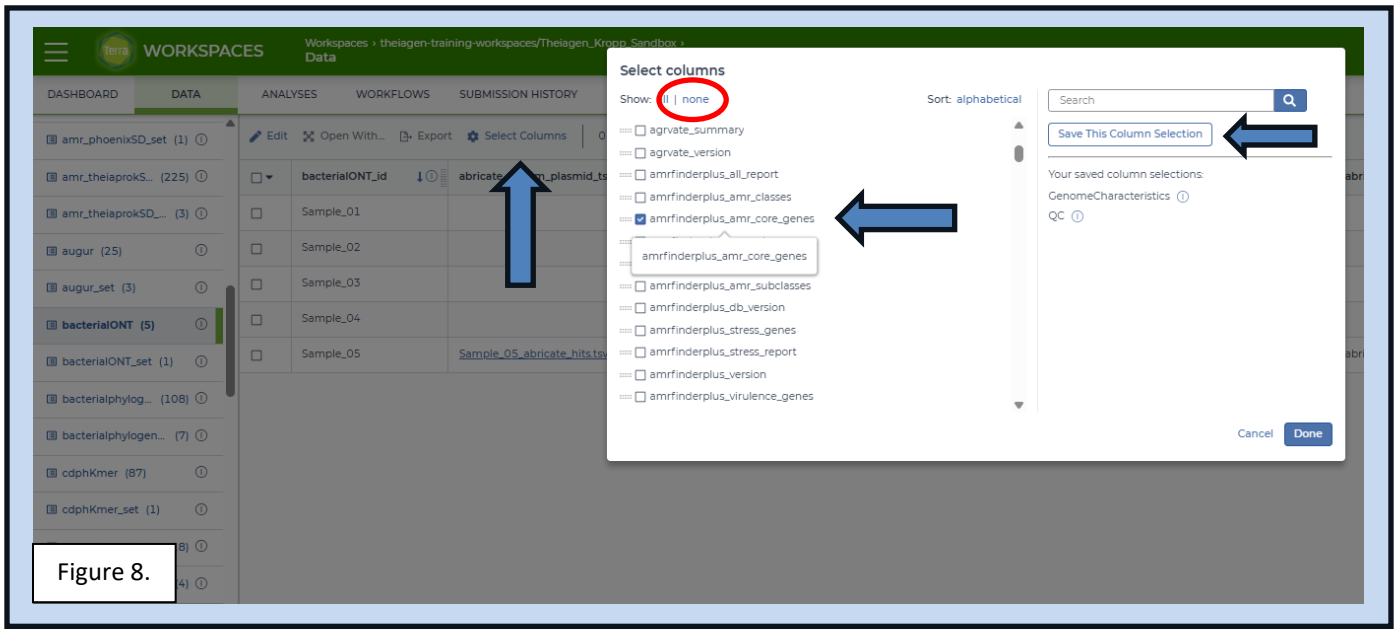

Figure 8.

4.5 DETERMINING GENOME CHARACTERISTICS

1. Navigate to the *data* tab of the Terra workspace containing bacterial data
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 8)
4. Select the following columns:
   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. *kaptive_k_type*: Kaptive predicted K type for *Acinetobacter baumannii*
      iii. *kleborate_ktype*: Kleborate predicted K type (capsule) for *Klebsiella spp*. serotyping
      iv. *kleborate_otype*: Kleborate predicted O type (LPS) for *Klebsiella spp.* serotyping
      v. *lissero_serotype*: serotype predicted by LisSero for *Listeria monocytogenes*
      vi. *meningotype_serogroup*: serogroup predicted by meningotype for *Neisseria meningitidis*
      vii. *pasty_serogroup*: serogroup predicted by Pasty for *Pseudomonas aeruginosa*

viii. *seqsero2_predicted_serotype*: serotype predicted by SeqSero2 for *Salmonella spp.*
ix. *serotypefinder_serotype*: serotype predicted by SerotypeFinder for *E. coli* and *Shigella spp.*
x. *shigatyper_predicted_serotype*: serotype predicted by ShigaTyper for *Shigella sonnei*
xi. *shigeifinder_serotype*: serotype predicted by ShigEiFinder for *Shigella sonnei*
xii. *sistr_predicted_serotype*: serotype predicted by SISTR for *Salmonella spp.*
xiii. *sistr_serogroup*: serogroup predicted by SISTR for *Salmonella spp.*
xiv. *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. Pseudo), and clicking* save (Fig 8)

5. Click *done*
6. Determine the predicted results for each sample by viewing the respective columns, as desired
7. Follow lab-specific resulting and reporting procedures, as applicable

5. QUALITY RECORDS

- Raw read files
- Sample read, assembly, and result-specific QC metrics
- All workflow outputs relevant to results including tool and database versions

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. The TheiaProk_ONT_PHB workflow only runs on bacterial, single-end ONT 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. 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
2. 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
Minor edits to align with Terra interface, add new outputs for species typing, improve formatting	2	5/2025

10. APPENDICES

None