

Document TG-PA-PE, Version 2Date:Workflow Version:3/4/2024PHB v1.3.0

1. PURPOSE/SCOPE

To standardize the process of running and analyzing bacterial isolates' next generation sequencing (NGS) data using Theiagen's TheiaProk Illumine 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. Metrics should always be assessed on the run level using instrument generated sequencing quality data prior to upload and analysis using Terra.

2. REQUIRED RESOURCES

- Computer
- Internet browser
 - o 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 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.

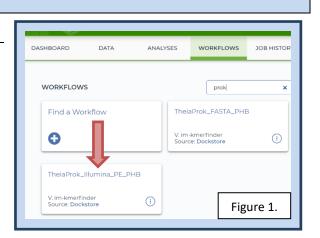
3. RELATED DOCUMENTS

Document Number	Document Name
TG-TER-03	Uploading Local or SRA NGS- Data & Creating a
IG-IEK-05	Results Metadata Table in Terra
	Linking BaseSpace and Importing BaseSpace
TG-TER-04	Reads to Terra Workspace

4. PROCEDURE

4.1 RUNNING THE THEIAPROK WORKFLOW

- Open Terra and navigate to the workflows tab within the workspace containing bacterial data of interest
- Select the <u>TheiaProk_Illumina_PE_PHB</u> workflow (Fig 1)
- 3. Uncheck call caching (Fig 2)

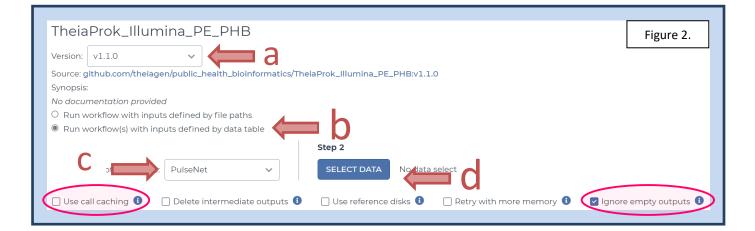




Document TG-PA-PE, Version 2

Date:	Workflow Version:
3/4/2024	PHB v1.3.0

- 4. <u>Optional:</u> 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 *root entity type* 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



5	illumina_pe_speci	imens to process 🎄 SETTINGS	20 rows selected		E ADVANCED SEARCH	Search
	Illumina 🛈 🛔	read1 ()	read2 0	run_id		
2	Sample_01	13_513_1001_P1_001.fasto.gz	13_513_L001_R2_001_fasto-gz	training_data		
	Sample, 02	15_515_L001_R1_001.fasto.gz	15_515_L001_R2_001/aste az	training_data		
	Sample_D3	17_517_L001_R1_001 festo.gz	17_517_L001_R2_001.fasto.oz	training_data		
	Sample_04	18_518_1001_R1_001_faster.or	18_518_L001_R2_001 fests or	training_data		
	Sample,05	19.519.1001.R1.001.fasto.oz	19.519.1001.82.001 fasto.oz	training_data		
	Sample_06	21.521.1001.91.001.fasto.gz	21 521 L001 P2_001 fasto.pz	training_data		
	Sample.07	23_523_1001_R1_001.fasto.gz	23.523.1001_R2_001.fasto.oz	training.data		
					1-20 of 20 (4) (4) 1	🕖 👔 Items per page: 🔟 🗸



Document TG-PA-PE, Version 2		
Date:	Workflow Version:	
3/4/2024	PHB v1.3.0	

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.theiaprok_illumina_pe_id

i.Where *theiaprok_illumina_pe* is the unique name of your data table in Terra

SCRIPT •• INPUTS	•• OUTPUTS •• RUN AN	IALYSIS	
lide optional inputs			
Task name ↓	Variable	Туре	Attribute
theiaprok_illumina_pe	read1_raw	File	this.read1
theiaprok_illumina_pe	read2_raw	File	this.read2
theiaprok illumina_pe	samplename	String	thistheiaprok_illumina_pe_id
Figure 4.	CDU	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

SCRIPT •• INPUTS	•• OUTPUTS •• RUN ANALYSIS		
Output files will be saved to			
Files / submission unique ID / theiap	prok_illumina_pe / v unique ID		
References to outputs will be written to			
Tables / theiaprok_illumina_pe			
Fill in the attributes below to add or upd	tate columns in your a table		
Task name 🖡	Variable	Туре	Attribute(Use defaults
Task name ↓ theiaprok_illumina_pe	Variable abricate_abaum_plasmid_tsv	Type File	Attribute Use defaults this.abricate_abaum_plasmid_tsv



Document TG-PA-PE, Version 2

Date:	Workflow Version:
3/4/2024	PHB v1.3.0

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. View *settings* above the data table, select *none* (Fig 6)
- 3. Select the appropriate columns:
 - a. assembly_length
 - b. combined_mean_q_clean
 - c. est_coverage_clean
 - d. number_contigs
 - e. <u>Optional</u>: save this column group for future use by clicking the save this column selection field, naming it (e.g. QC), and clicking save
- 4. Click done
- 5. Compare all metrics to relevant QC requirements and determine pass/fail calls for all samples

DASHBOARD DATA	ANA	LYSES WORKFLOWS	JOB HISTORY	Show: all none	Sort: alphabetical	
KK_ClearLabs_Te (2) 🕕		T 🔀 OPEN WITH 🕒 EXPO		amrfinderplus_all_report	A	SAVE THIS COLUMN SELECTION
KK_ClearLabs_Te (1) 🛈		theiaprok_illumina_pe_id		amrfinderplus_amr_core_genes		Your saved column selections:
ONT_Test (3)		chelapiok_indifina_pe_id		amrfinderplus_amr_plus_genes	•	Pseudo 🕕
ONT_Test_set (2)		SAMN24249320	par 7L,gyrA_	amrfinderplus_amr_report amrfinderplus_amr_subclasses		
		SAMN24249373	par uga 7L,catB5	amrfinderplus_db_version		
SC2_Assemblie (25) 🕕		SAMN24249374	crpP,blaPDC-35,	amrfinderplus_stress_genes		
SC2_Assemblies (1) 🔅		SAMN31370019	fosA,parC_S87L,c	amrfinderplus_stress_report amrfinderplus_version		
TheiaCoV_IIIumi (5) 🕕		SAMN31384136	catB7.aph(6)-ld.a	amrfinderplus_virulence_genes		
TheiaCoV_IIIumi (1)		SAMIN31304136	cate 7,aph(6)-10,a	amrfinderplus_virulence_report		
				- assembly_fasta	-	
🔲 illumina_pe_sp (20) 🕕						CANCEL DONE



Document TG-PA-PE, Version 2

Date:	Workflow Version:
3/4/2024	PHB v1.3.0

- 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 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. View *settings* above the data table, select *none* (Fig 6)
- 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. *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. <u>seroba_ariba_serotype</u>: 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 predictd 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. *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



Document TG-PA-PE, Version 2

Date:	Workflow Version:		
3/4/2024	PHB v1.3.0		

- iv. *ngmaster_ngstar_sequence_type*: Ngstar predicted ST for *Neisseria gonorrohoeae*
- v. *ts_mlst_predicted_st*: Torsten Seemann predicted ST for bacterial 7-gene MLST
- f. <u>Optional</u>: save this column group for future use by clicking the save this column selection field, naming it (e.g. Pseudo), and clicking save
- 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, when applicable
- Result-specific determinations

6. TROUBLESHOOTING

- Consult with internal staff familiar with this procedure or contact <u>support@theiagen.com</u> for troubleshooting inquiries
- For document edit requests, contact <u>support@theiagen.com</u>

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

 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

10. APPENDICES

None