| | | |
|---|---|---|
| | **Analyzing Bacterial Data in Terra using Theiagen's TheiaProk Illumina PE Workflow** | |
| | Document TG-PA-PE, Version 2 | |
| | Date: | Workflow Version: |
| | 3/4/2024 | PHB 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
  - 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 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.

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 *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*

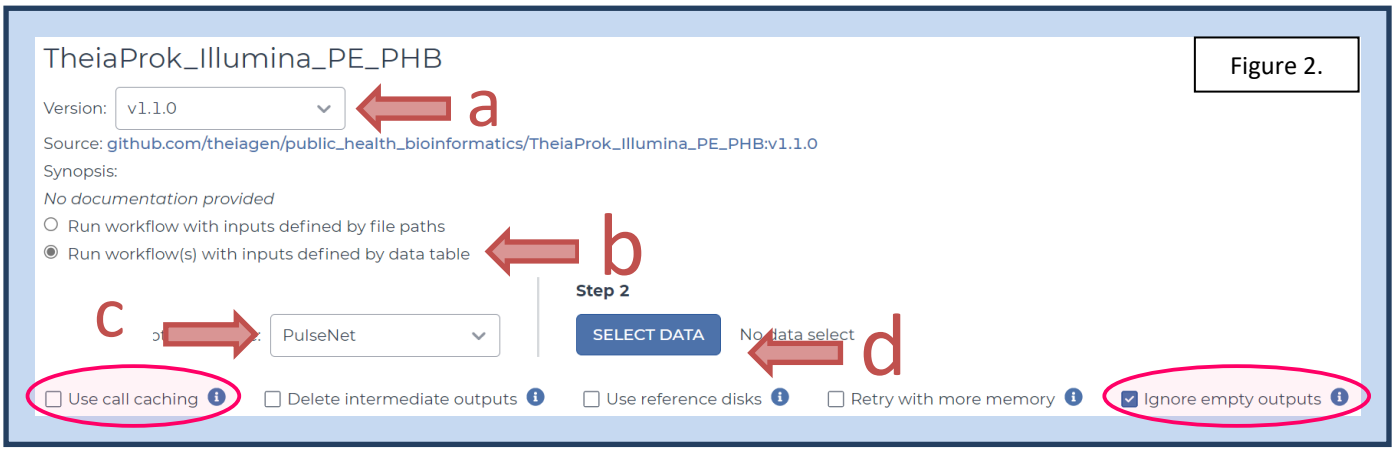


Figure 2.

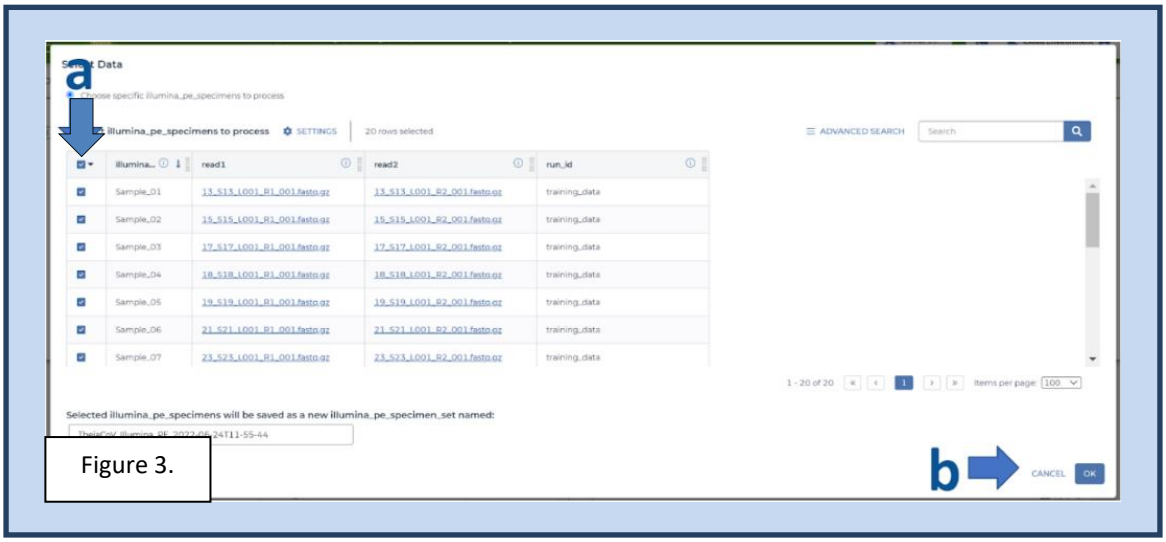


Figure 3.

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

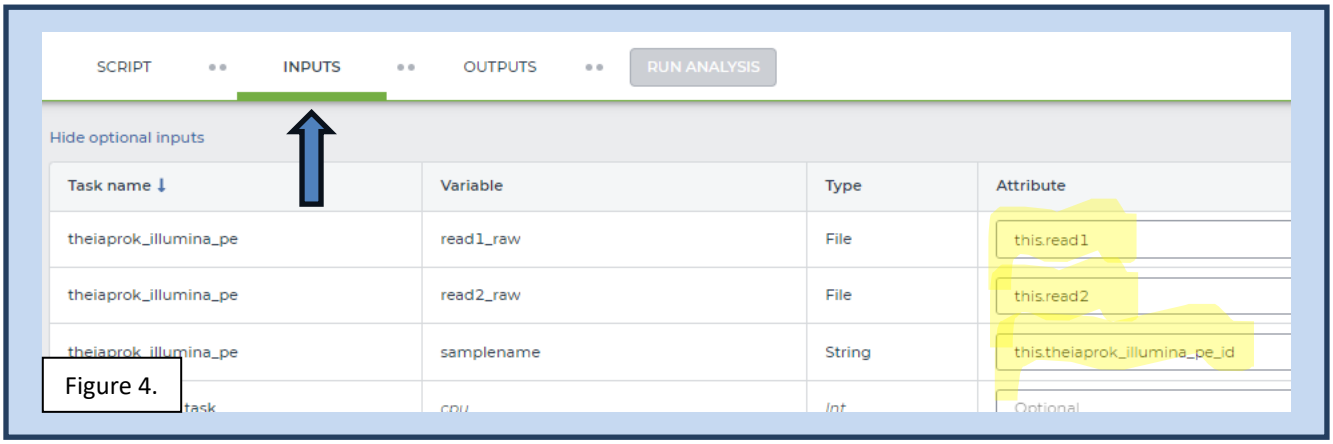

Figure 4.

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*

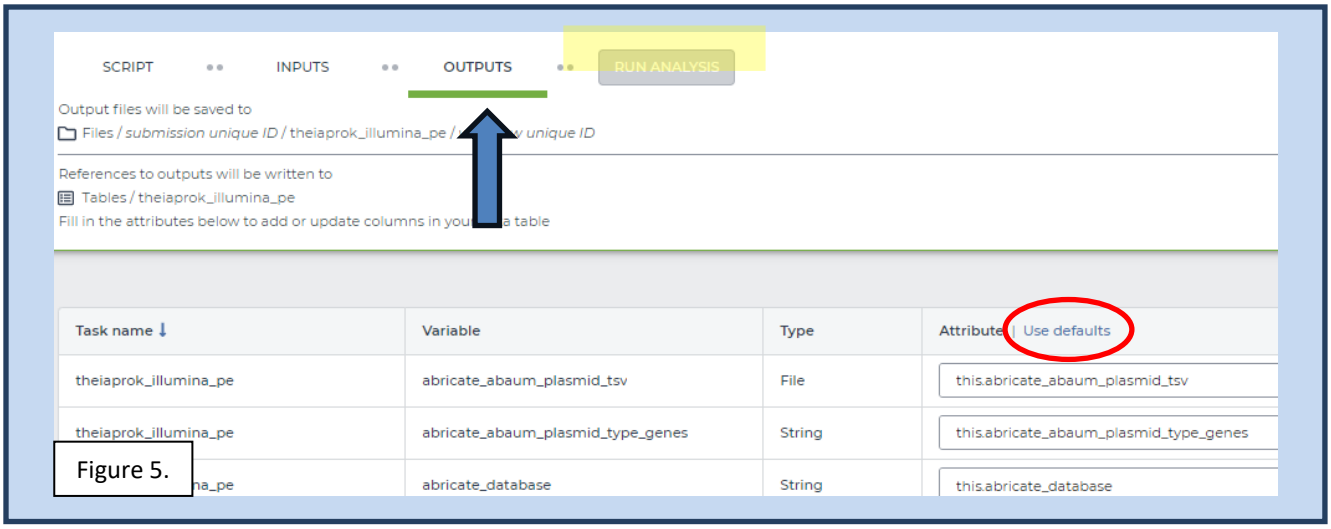

Figure 5.

## 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. 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. Optional: 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

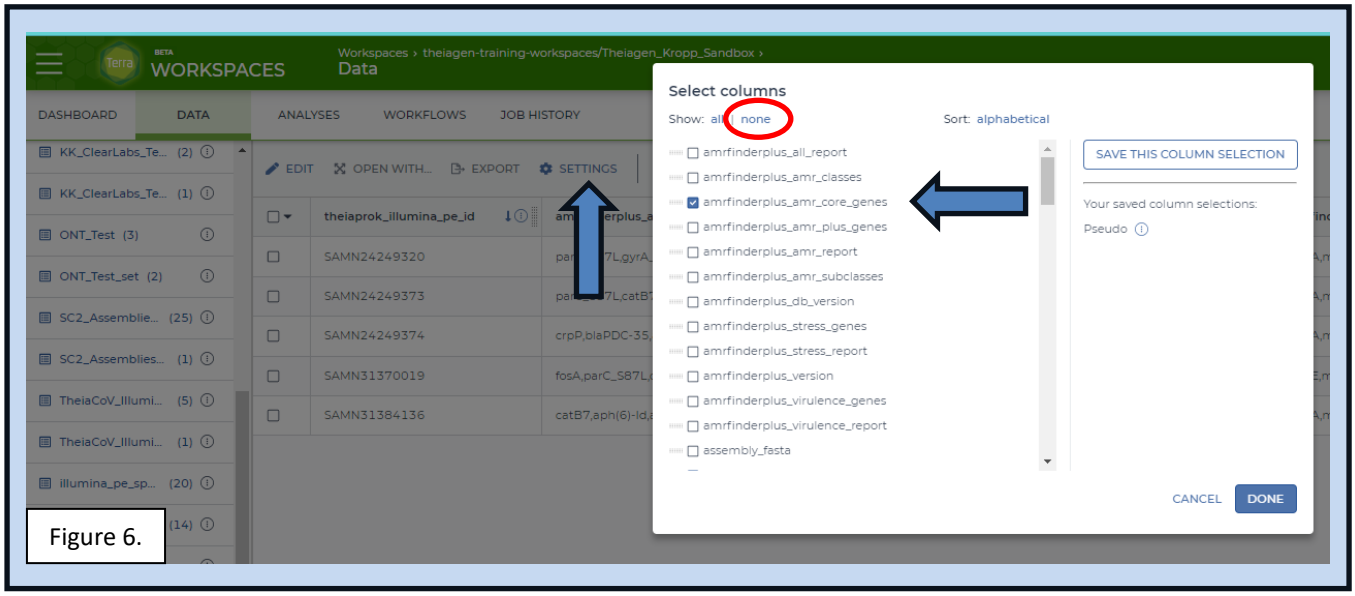

Figure 6.

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. *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 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*

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. 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 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 |
| | | |

## 10. APPENDICES

None