Comparing Terra Data Tables using Theiagen's TheiaValidate Workflow

Document TG-VAL-01, Version 3	
Date:	Workflow Version:
5/1/2025	PHB v3.0.0

1. PURPOSE/SCOPE

To automate the process of comparing data between two Terra tables using Theiagen's TheiaValidate_PHB workflow. No files are required for this procedure, however an optional user-defined validation criteria .tsv or .txt file may be input with user definitions of comparison criteria. Read more here: Theiagen's Public Health Bioinformatics Documentation

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
- Sample comparison data in Terra workspace/s
- TheiaValidate_PHB workflow in Terra; see appendix 10.1

WORKFLOW REQUIREMENTS

- Tables to compare must contain identical sample names in column 1 and an equal number of samples
- Columns to compare must have identical names in data tables 1 and 2
- To input user-defined validation criteria a .tsv or .txt file is required; see section 4.2

3. RELATED DOCUMENTS

Document Number	Document Name
None	N/A

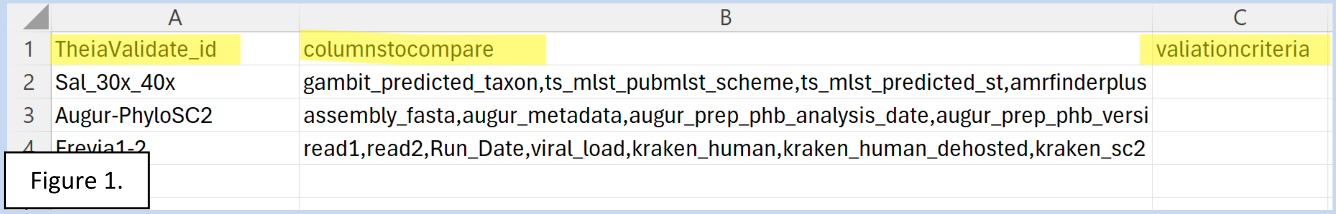
4. PROCEDURE

4.1 CREATING A VALIDATION DATA TABLE AND ADDING SAMPLE SETS

1. When using the TheiaValidate workflow for the first time, *create a new Terra data table* to specify validation parameters and record results; otherwise skip to step 2
   a. Create a *new tsv file* in Excel (Fig 1)
   b. Title cell A1 as *validations_id*, *TheiaValidate_id*, or something similar
   c. Specify the *name of each data table comparison* that will be run under column A without using spaces (e.g. Sal_50x_40x, Sal_40x_30x, etc)
   d. Title cell B1 as *columnstocompare*
   e. Under column B, without spaces, *create a comma separated list* of each column to include in sample comparisons (e.g. taxon,MLST_scheme,ST,AMR_Genes,assembly_length,etc)

f. *Optional: For comparisons where user-defined criteria will be used, title cell C1 as* validationcriteria*; once a validation criteria txt file has been created the file path will be pasted in this column in Terra;* see section 4.2 for details

g. *Optional: Create other columns and add details as desired (e.g. notes, etc)*



	A	B	C
1	TheiaValidate_id	columnstocompare	valiationcriteria
2	Sal_30x_40x	gambit_predicted_taxon,ts_mlst_pubmlst_scheme,ts_mlst_predicted_st,amrfinderplus	
3	Augur-PhyloSC2	assembly_fasta,augur_metadata,augur_prep_phb_analysis_date,augur_prep_phb_versi	
4	Freyja1-2	read1,read2,Run_Date,viral_load,kraken_human,kraken_human_dehosted,kraken_sc2	

Figure 1.

2. If a validations data table or equivalent has already been created in Terra, *add a table row and relevant data* for each new comparison that will be run
   a. Manually adding new rows
      i. In the sample data table, click *edit*, *add row*, *name the data row* (e.g. Sal_50x_40x) and click *add*
      ii. Edit the columnstocompare column by hovering the mouse within the relevant columnstocompare cell and clicking the *pencil icon*
      iii. Without using spaces, *create a comma separated value list* for each column to include in the comparison and click *save changes*
         1. Alternatively: If the comparison will use the same columns as listed for a previous comparison already listed in the data table, click on the clipboard icon of the columnstocompare cell to copy, then click the pencil icon to edit the new columnstocompare cell, paste and save the text
   b. Adding multiple rows by downloading and re-uploading the Terra data table
      i. Download the validation data table from Terra by *opening the relevant table*, selecting the *checkbox for all rows*, clicking *export*, and *download as tsv*
      ii. *Add a new row* for each new data comparison, *naming the comparison* in column A, and without spaces *creating a comma separated list* of columns to compare in column B (Fig 1)
      iii. *Name and save the file* in tsv file format, then upload the file to Terra by clicking *import data*, *upload tsv*, *select the relevant file*, and clicking *start import job*

4.2 (OPTIONAL): **ADDING USER-DEFINED VALIDATION CRITERIA**

1. Create a tsv file using the format in Figure 2
    a. Title column A *column*
    b. Find the output names in the relevant data tables and *create a list of all columns to compare* under column A using one row per output; these names must match column headers exactly
    c. Title column B *criteria*
    d. OPTIONAL: Title column C *delimiter*
        - For results that include non-comma delimiters (e.g. BUSCO)
        - Can take delimiters of multiple types (e.g. BUSCO)
    e. *Define the comparison criteria* to use for each data column
        - *Note: The* **black** *criteria below indicates the string to be used, while the* ***gray*** *criteria indicates the use of the proceeding values*
        - For all criteria, dates and object-type values are ignored and indicate 0 failures.

i. **EXACT** will fail samples that do not have an exact value match (numerical or text)

ii. **IGNORE** will disregard the data values and no samples will fail. Useful for including data of interest in the results (e.g. coverage or taxonomic values)

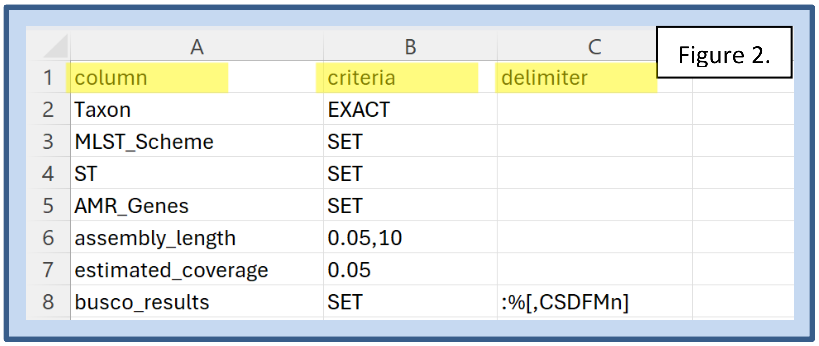


Figure 2.

	A	B	C
1	column	criteria	delimiter
2	Taxon	EXACT	
3	MLST_Scheme	SET	
4	ST	SET	
5	AMR_Genes	SET	
6	assembly_length	0.05,10	
7	estimated_coverage	0.05	
8	busco_results	SET	:%[,CSDFMn]

iii. **SET** compares a list of items without regard to order and samples will fail when any items between lists are not identical (e.g. helpful for AMR result comparison)

iv. **PERCENT_DIFF** will fail samples when two values differ more than the indicated percentage
- Use *decimal* format (e.g **0.05** for 5% difference)

v. **RANGE** will fail samples when two values differ more than the indicated integer
- Use *integer* format (e.g. **10** for calculating a >10 difference (units are ignored)

vi. **CRITERIA1,CRITERIA2,...** will fail samples after checking both criteria (e.g. if criteria1 fails, then check criteria2, etc. before failing)
- Indicated for use of mixed-type criteria (e.g. **0.05,10** for checking 5% first, then 10 units)
- MUST be a comma-delimited list without spaces

   f. *Save the file* with a relevant title (e.g. ValidationCriteria_Sal) and *upload to Terra*
      i. To upload a file to Terra, in the Terra workspace, scroll to the bottom and open *files* in the left sidebar
      ii. Click *upload*, *select the validation criteria tsv*, and click *ok*
2. In the Terra workspace where the validation criteria file was uploaded, navigate to the workspaces Files by scrolling to the bottom of the left sidebar and click *files*
3. Find the validation criteria file in the files table, hover the mouse over the relevant cell and click the *clipboard icon* to copy the file location
4. Navigate to the validation data table and *paste* the validation criteria file location into the *validationcriteria* column for the corresponding sample set (e.g. Sal_30x_40x)

4.3 RUNNING THE THEIAVALIDATE WORKFLOW

1. In Terra, navigate to the data tab and view the tables to compare
   a. Take note of the *exact table names*
   b. Verify they contain the *same sample IDs* and *number of samples*, or the workflow will fail
   c. If data tables are in different workspaces, also note the *exact workspace* and *project names*
      i. To identify workspace and project names, *navigate to the data table* and *compare the URL* to Figure 3
2. The project name indicated by Figure 3 is theiagen-training-workspaces
3. The workspace name indicated by Figure 3 is Theiagen_Kropp_Sandbox

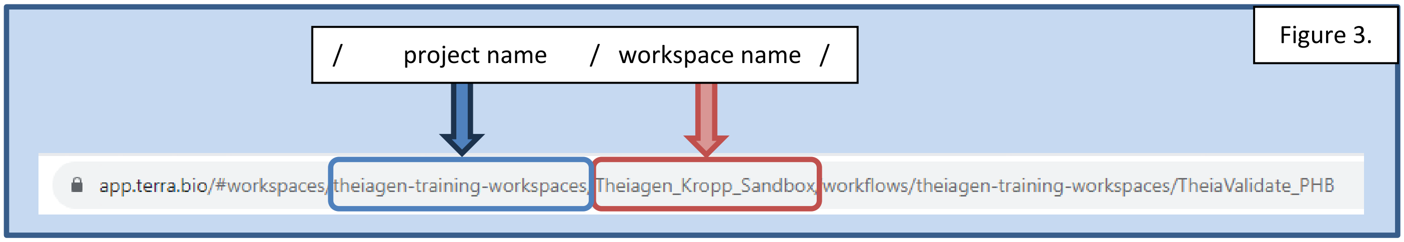


Figure 3.

4. In the workflows tab, open the *TheiaValidate_PHB* workflow
5. *Uncheck call caching* (Fig 4)
6. *Choose the latest version of the workflow* or the version used during internal validation (Fig 4, a)
7. Select the second bullet to *run workflow(s) with inputs defined by data table* (Fig 4, b)
8. Select the relevant data table under the *select root entity type* dropdown (Fig 4, c)
   a. This is the validation data table
9. Click *select data* (Fig 4, d)

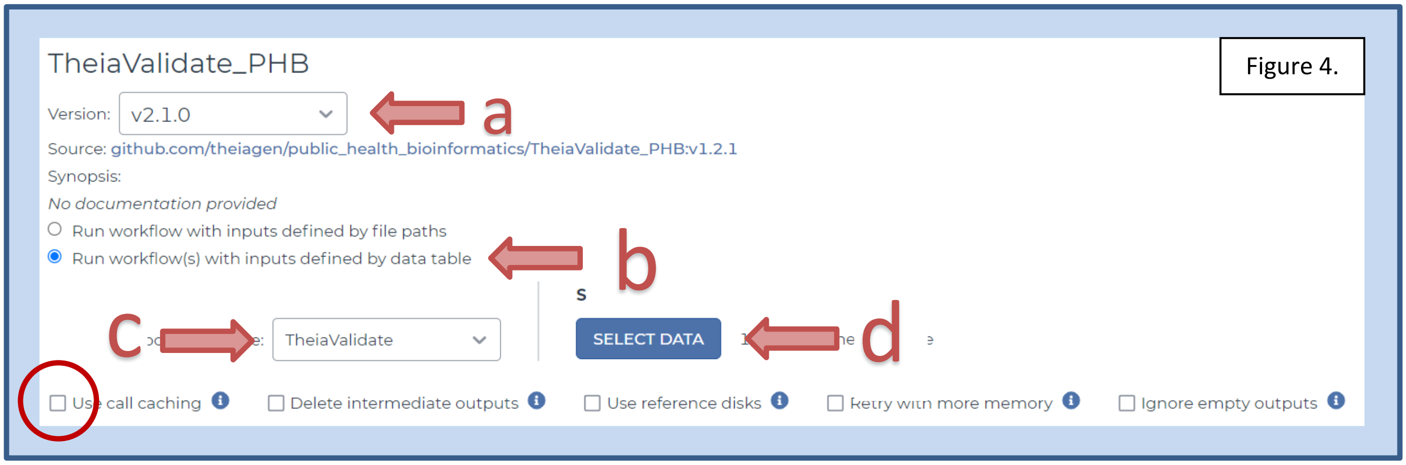


Figure 4.

10. In the pop-up window, the second bullet to *choose specific TheiaValidates to process* should be selected where *TheiaValidate* is the name of the data table created to record TheiaValidate results (Fig 5)
    a. *Select one sample comparison* row to analyze
        i. Only one sample comparison can be performed at once since workflow inputs require table 1 and table 2 specific information

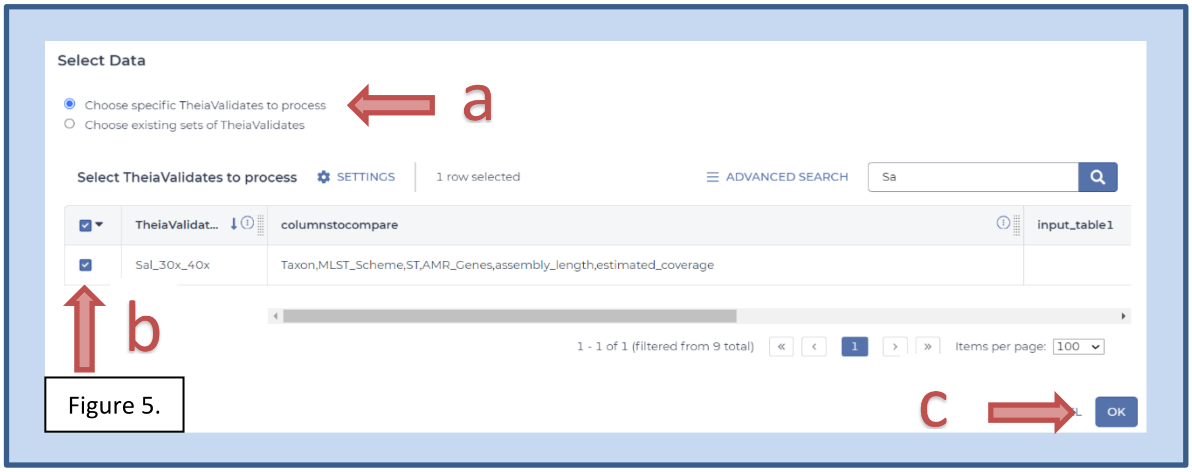


Figure 5.

    b. Scroll to the bottom and click *ok*
11. In the inputs tab, specify the following input fields, respectively (Fig 6):
    a. columns_to_compare: *this.columnstocompare* – the column name of the validation data table specifying which columns to analyze
    b. output_prefix: *this.TheiaValidate_id* – the Terra data table to output TheiaValidate results
    c. table1: *"Sal_30x"* – the name of the first Terra data table to compare

d. table2: *"Sal_40x"* – the name of the second Terra data table to compare

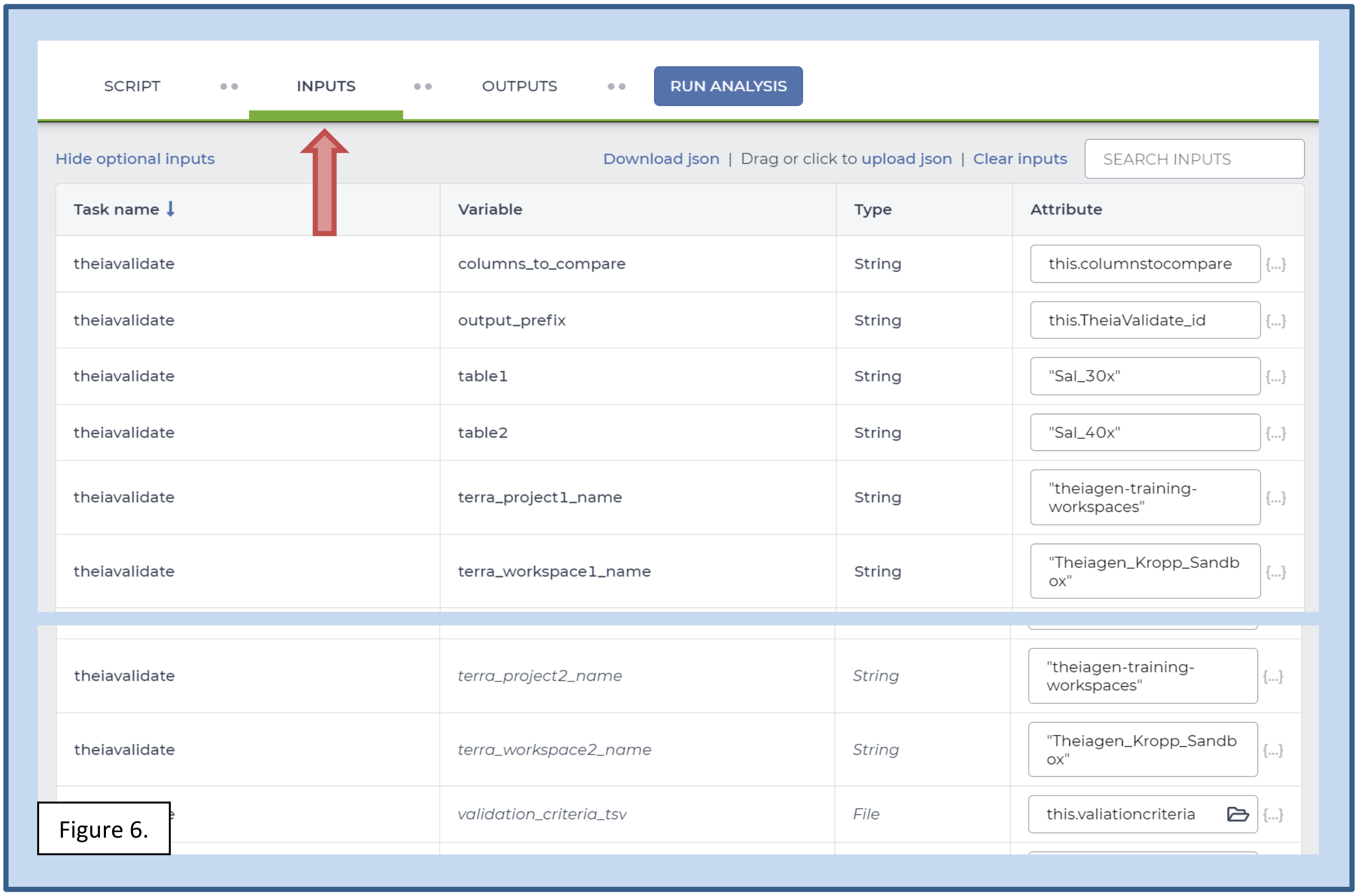
e. terra_project1_name: *"theiagen-training-workspaces"* - the Terra project name where data table 1 is located; see section 4.3, step 1c for details regarding finding this information

f. terra_workspace1_name: *"Theiagen_Kropp_Sandbox"* - the workspace name where data table 1 is located; see section 4.3, step 1c for details regarding finding this information

g. terra_project2_name: *"theiagen-training-workspaces"* – the Terra project name for table 2

h. terra_workspace2_name: *"Theiagen_Kropp_Sandbox"* – the workspace name for table 2

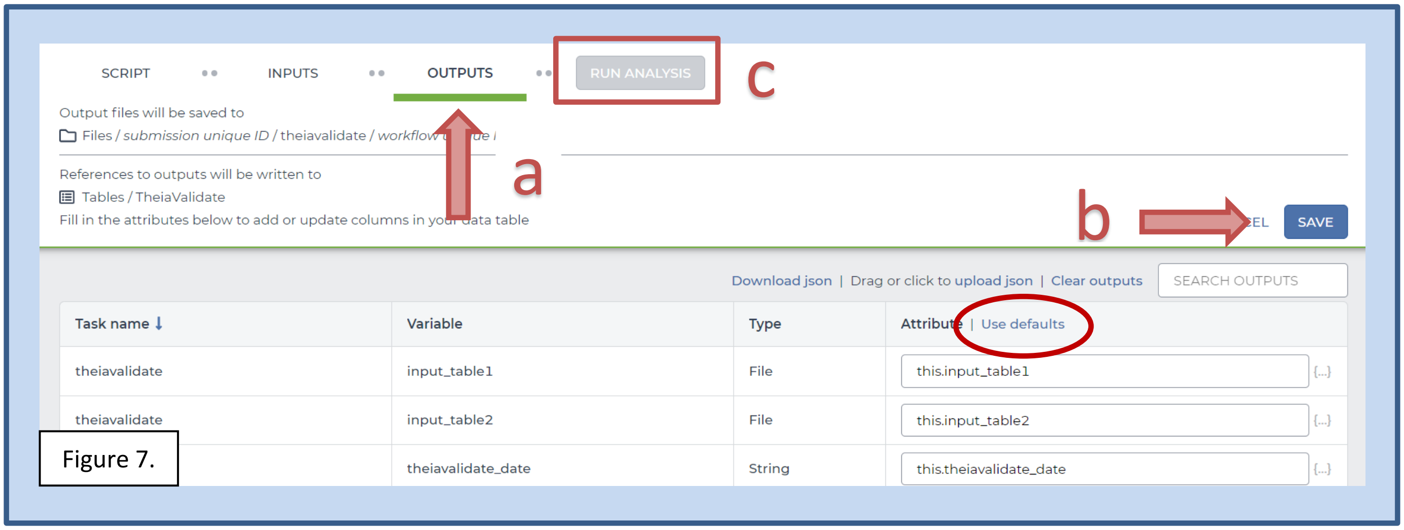
i. *Optional: validation_criteria_tsv:* this.validationcriteria – *see section 4.2 for instructions*



Task name	Variable	Type	Attribute
theiavalidate	columns_to_compare	String	this.columnstocompare
theiavalidate	output_prefix	String	this.TheiaValidate_id
theiavalidate	table1	String	"Sal_30x"
theiavalidate	table2	String	"Sal_40x"
theiavalidate	terra_project1_name	String	"theiagen-training-workspaces"
theiavalidate	terra_workspace1_name	String	"Theiagen_Kropp_Sandbox"
theiavalidate	terra_project2_name	String	"theiagen-training-workspaces"
theiavalidate	terra_workspace2_name	String	"Theiagen_Kropp_Sandbox"
	validation_criteria_tsv	File	this.valiationcriteria

Figure 6.

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



Output files will be saved to
Files / *submission unique ID* / theiavalidate / *workflow unique ID*

References to outputs will be written to
Tables / TheiaValidate
Fill in the attributes below to add or update columns in your data table

Task name	Variable	Type	Attribute
theiavalidate	input_table1	File	this.input_table1
theiavalidate	input_table2	File	this.input_table2
	theiavalidate_date	String	this.theiavalidate_date

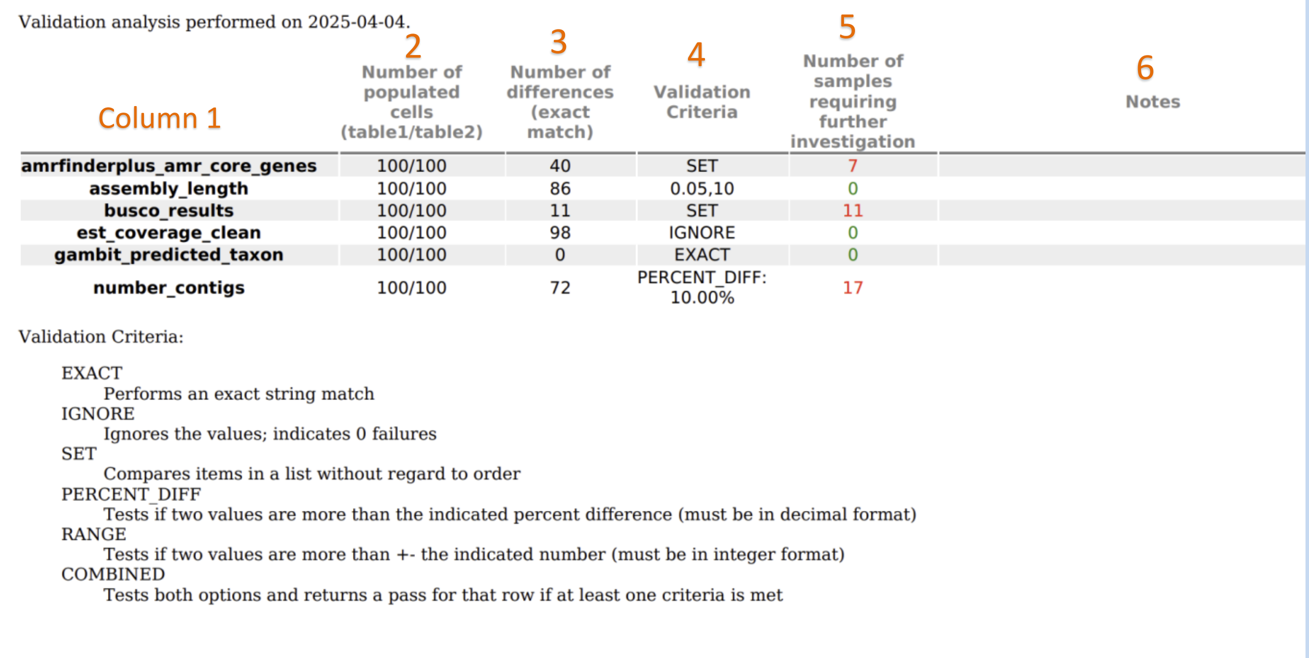
Figure 7.

4.4 EVALUATING THE DATA COMPARISON

1. Once the TheiaValidate job has successfully completed, navigate to the relevant *validation data table* in the respective Terra workspace
2. *For the data comparison of interest, there should be an additional ten data columns from workflow* ***outputs****, six of which are output files*
    a. *Output Files*
        i. *Theiavalidate_filtered_input_table1 and theiavalidate_filtered_input_table2 outputs are files of the two tables that were used for validation, excluding unexamined columns*
        ii. *theiavalidate_exact_differences is an output file of all non-exact matches between samples without accounting for any user-defined validation criteria*
        iii. *theiavalidate_criteria_differences is an output file listing only the difference that fails to meet user-defined validation criteria*
        iv. *validation_report is a pdf displaying the results of the data comparison for both exact matches and user-defined criteria; see below for details*
    b. *Run information*
        i. *theiavalidate_date and theiavalidate_status indicate the date the workflow was run and if the attempted validation was successful (validation attempted) or failed (validation failure)*
        ii. *theiavalidate_version and theiavalidate_wf_version are the versions of the TheiaValidate Python Docker and the PHB repository, respectively*
3. Find the associated validation report in the Terra validation data table (TheiaValidate) under the column titled theiavalidation_report for the appropriate comparison (e.g. row titled Sal_30x_40x)
    a. *The top of the validation report indicates the date the workflow was run*
    b. *Column 1 of the data table shows all data columns included in the comparison*

   c. *Column 2 indicates the number of populated cells per field per each data table.*
      - Samples without data for the respective field are not counted, and discordance between the tables will be noted.
   d. *Column 3 reports the number of samples between tables 1 and 2 that do not match exactly, regardless of the user-defined criteria for that field*
   e. *Column 4 notes the user-defined validation criteria used for reporting sample failures in column 5*
      - See the validation criteria listed at the bottom of the report for definitions
   f. *Column 5 reports the number of samples that failed the designated criteria and may require further investigation*
      i. *Values will be in red if more than one sample requires investigation and in green if not*
   g. *Column 6 provides space for the user to add information regarding the investigation*

4. Use the validation_differences_table output to see specific sample data values regarding exact match differences between samples
5. Refer to the theiavalidate_criteria_differences file to assess differences per user-defined validation criteria



Validation analysis performed on 2025-04-04.

Column 1	2 Number of populated cells (table1/table2)	3 Number of differences (exact match)	4 Validation Criteria	5 Number of samples requiring further investigation	6 Notes
amrfinderplus_amr_core_genes	100/100	40	SET	7	
assembly_length	100/100	86	0.05,10	0	
busco_results	100/100	11	SET	11	
est_coverage_clean	100/100	98	IGNORE	0	
gambit_predicted_taxon	100/100	0	EXACT	0	
number_contigs	100/100	72	PERCENT_DIFF: 10.00%	17	

Validation Criteria:

EXACT
  Performs an exact string match
IGNORE
  Ignores the values; indicates 0 failures
SET
  Compares items in a list without regard to order
PERCENT_DIFF
  Tests if two values are more than the indicated percent difference (must be in decimal format)
RANGE
  Tests if two values are more than +- the indicated number (must be in integer format)
COMBINED
  Tests both options and returns a pass for that row if at least one criteria is met

Figure 8.

5. QUALITY RECORDS

- Terra input_table1 and input_table2
- theiavalidate_exact_differences file
- theiavalidate_criteria_differences file
- theiavalidate_report
- Workflow versioning and input parameters (validationcriteria file, as applicable)

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. Tables to compare must contain identical sample names and an equal number of samples
2. Columns to compare must have identical names for data tables 1 and 2

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
3. Theiagen's PHB GitHub Repository - TheiaValidate README

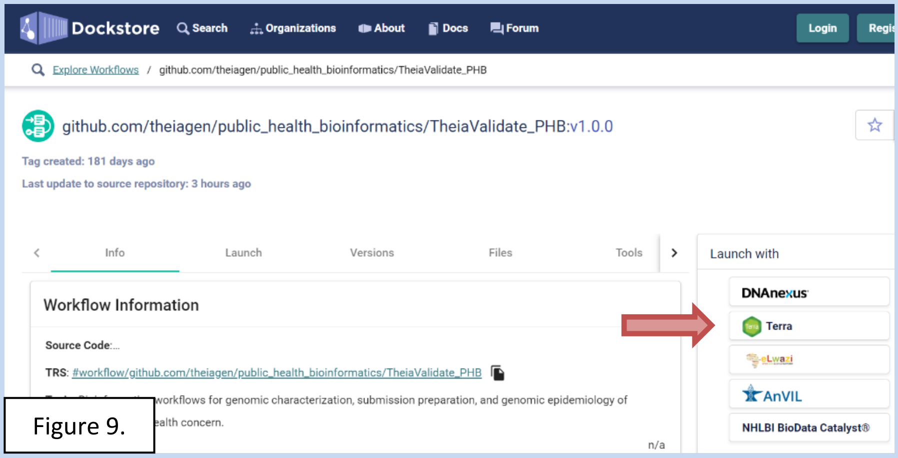
9. REVISION HISTORY

Revision	Version	Release Date
Document creation	1	12/2023
Updated to incorporate theiavalidate_criteria_differences output file and usage, quality records and limitations section updates, minor formatting changes	2	8/2024
Major updates to criteria options and outputs, minor updates to align with Terra interface and improve formatting	3	5/2025

10. APPENDICES

10.1 IMPORTING THE THEIAVALIDATE WORKFLOW FROM DOCKSTORE

1. Navigate to the Dockstore repository for the TheiaValidate workflow at https://dockstore.org/workflows/github.com/theiagen/public_health_bioinformatics/TheiaValidate_PHB:v1.0.0
2. Click on the *Terra icon* to export the workflow to Terra (Fig 9)
3. *Sign in* to Terra if necessary and choose the *destination workspace* to copy the workflow to (Fig 10); click *import*



Explore Workflows / github.com/theiagen/public_health_bioinformatics/TheiaValidate_PHB

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

Tag created: 181 days ago

Last update to source repository: 3 hours ago

Workflow Information

Source Code:...

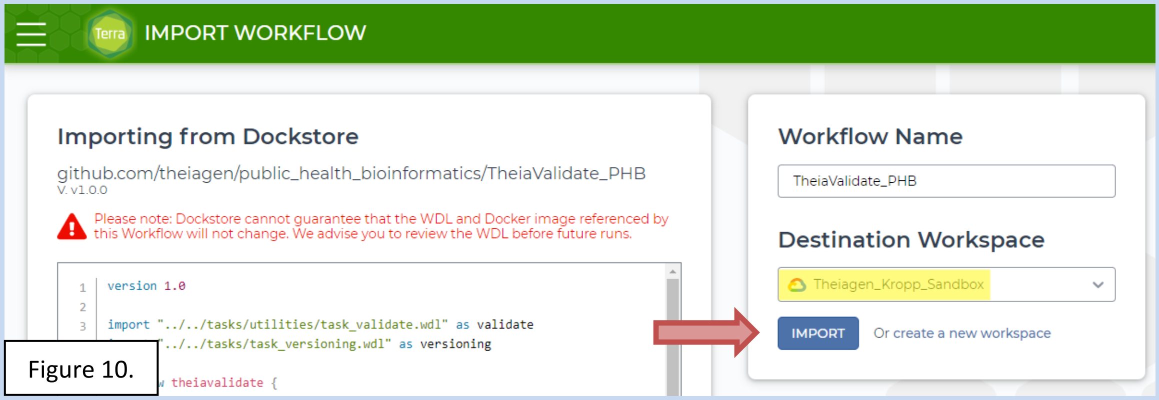
TRS: #workflow/github.com/theiagen/public_health_bioinformatics/TheiaValidate_PHB

workflows for genomic characterization, submission preparation, and genomic epidemiology of

ealth concern.

n/a

Figure 9.



Importing from Dockstore

github.com/theiagen/public_health_bioinformatics/TheiaValidate_PHB
V. v1.0.0

Please note: Dockstore cannot guarantee that the WDL and Docker image referenced by this Workflow will not change. We advise you to review the WDL before future runs.

```
version 1.0

import "../../tasks/utilities/task_validate.wdl" as validate
      "../../tasks/task_versioning.wdl" as versioning

     w theiavalidate {
```

Workflow Name: TheiaValidate_PHB

Destination Workspace: Theiagen_Kropp_Sandbox

Figure 10.