Comparing Terra Data Tables using Theiagen's TheiaValidate Workflow	
Document TG-VAL-01, Version 2	
Date:	Workflow Version:
8/20/2024	PHB v2

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.

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 *entity:validations_id*, *entity:TheiaValidate_id*, or something similar

	A	B	C
1	entity:TheiaValidate_id	columnstocompare	valiationcriteria
2	Sal_30x_40x	Taxon,MLST_Scheme,ST,AMR_Genes,assembly_length,estimated_coverage	
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,k	

Figure 1.

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)*

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 exactly
   c. Title column B *criteria*
   d. *Define the comparison criteria* to use for each data column
      i. EXACT will fail samples that do not have an exact value match (numerical or text)
      ii. IGNORE will disregard the data values and there no samples will fail
      iii. SET compares a list of items <u>without regard to order</u> 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
         1. Use decimal format (e.g 0.05 for 5% difference)

	A	B
1	column	criteria
2	Taxon	EXACT
3	MLST_Scheme	SET
4	ST	SET
5	AMR_Genes	SET
6	assembly_length	0.05
	ed_coverage	0.05

Figure 2.

   e. *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

1. The project name indicated by Figure 3 is theiagen-training-workspaces
2. The workspace name indicated by Figure 3 is Theiagen_Kropp_Sandbox

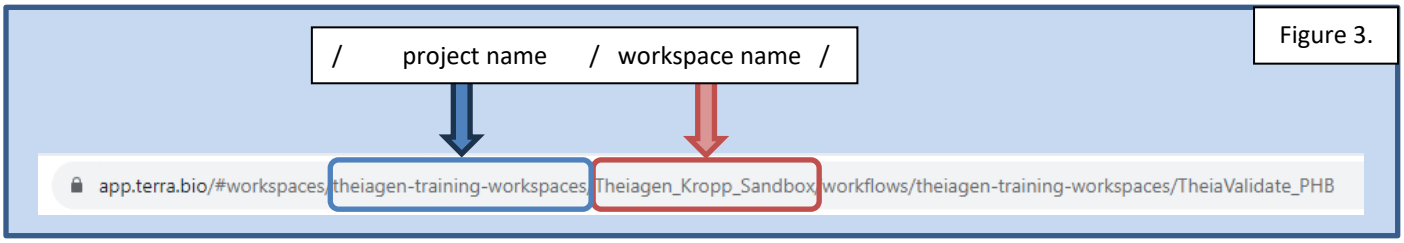

Figure 3.

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

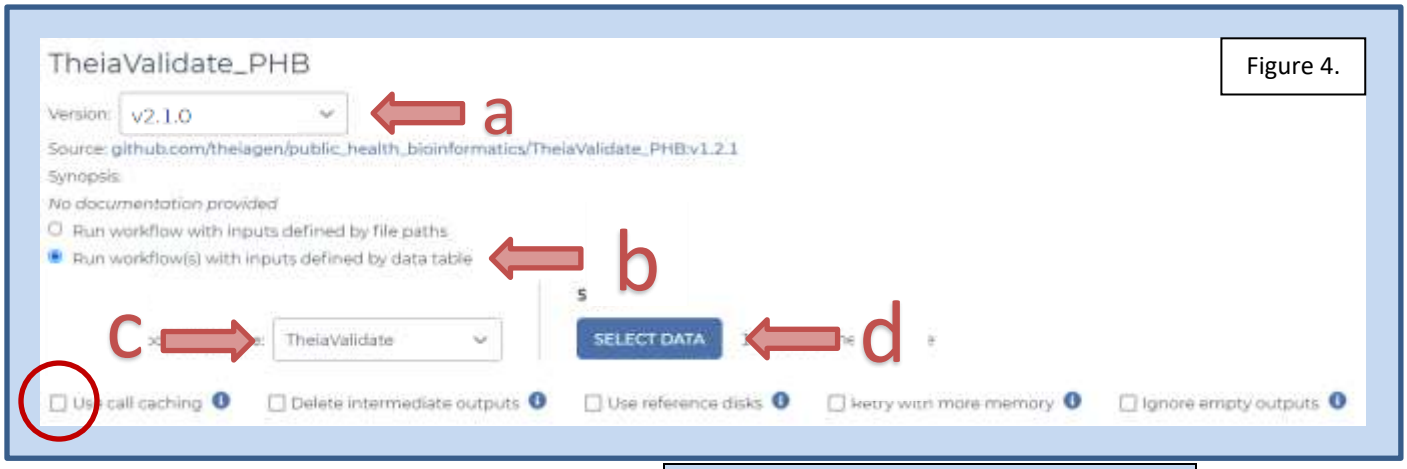

Figure 4.

8. 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
   b. Scroll to the bottom and click *ok*

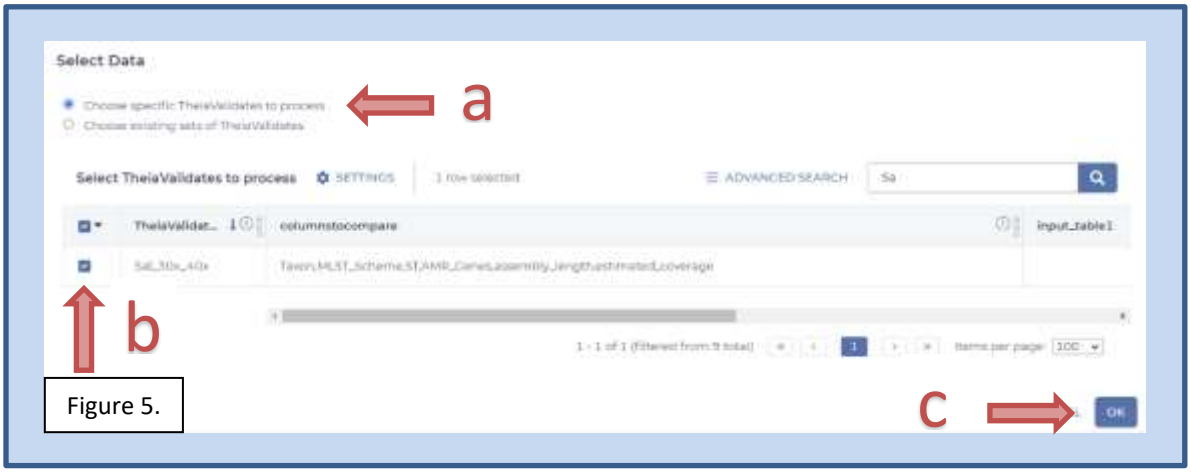

Figure 5.

9. 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*
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*; enter desired comments and click *launch*

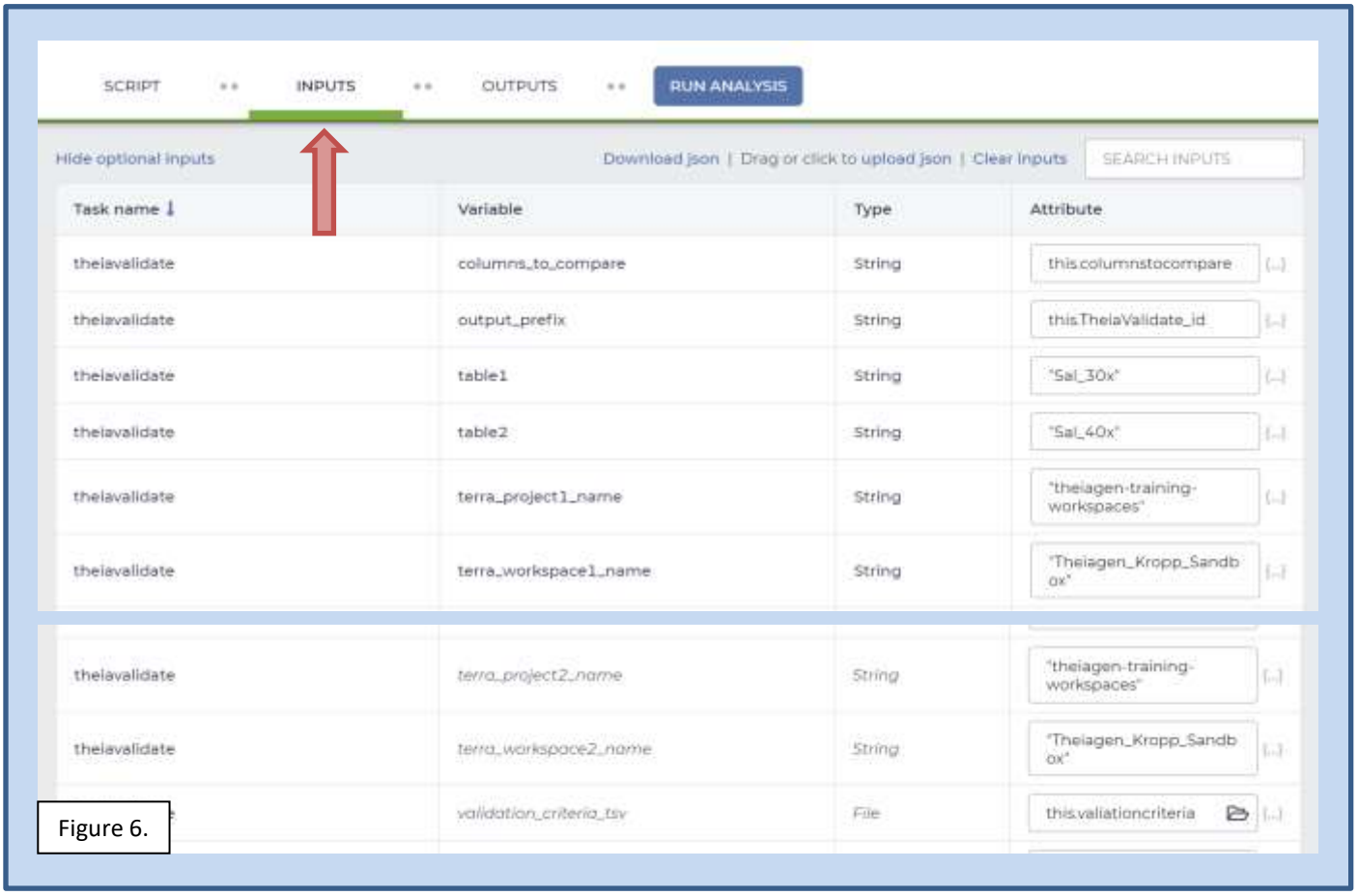

Figure 6.

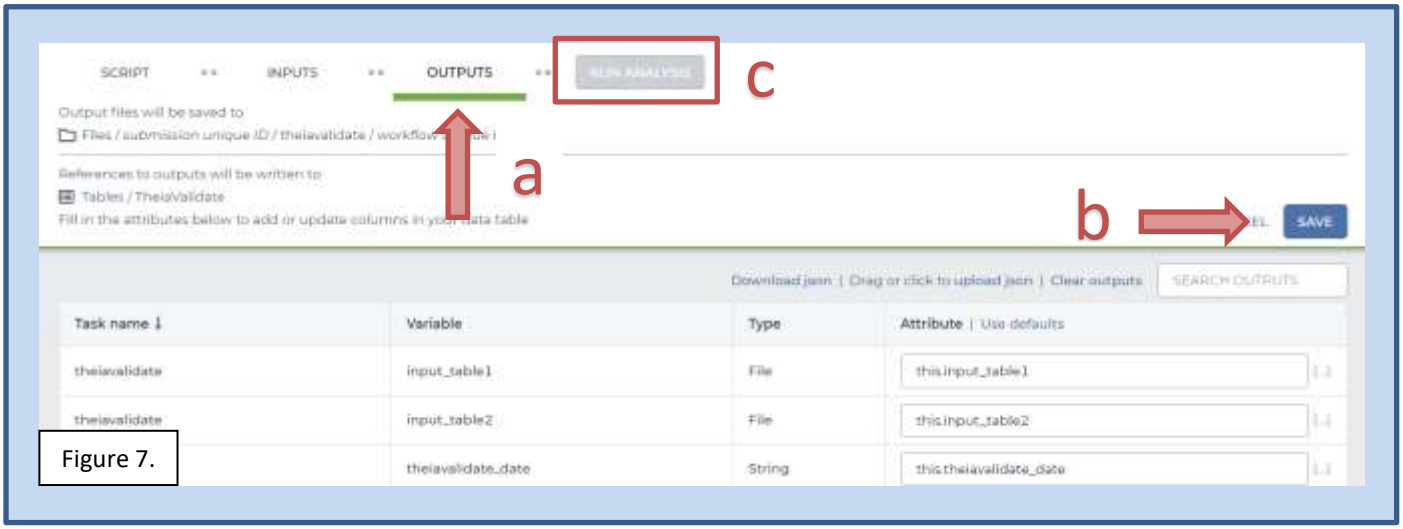

Figure 7.

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 to make the comparison after removing 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 fail 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 and 3 indicate the number of samples from each data table that had data (samples without data for the respective field/s are not included in these counts)*
   d. *Column 4 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 5 notes the user-defined validation criteria used for reporting sample failures in column 6*
      i. *See the validation criteria listed at the bottom of the report for definitions*
   f. *Column 6 reports the number of samples that fail per user-defined criteria*
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 2023-07-29.

Column 1	Column 2	Column 3	4	5	6
	Number of samples populated in SD_PhoenixAnalysis	Number of samples populated in SD_TheiaProkAnalysis	Number of differences (exact match)	Validation Criteria	Number of samples failing the validation criteria
AMR_Genes	89	100	14	SET	14
assembly_length	92	98	98	PERCENT_DIFF: 5.00%	1
estimated_coverage	92	98	98	PERCENT_DIFF: 5.00%	80
MLST_Scheme	92	98	36	SET	36
ST	92	98	28	SET	28
Taxon	92	98	13	EXACT	13

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)

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

None

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

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*

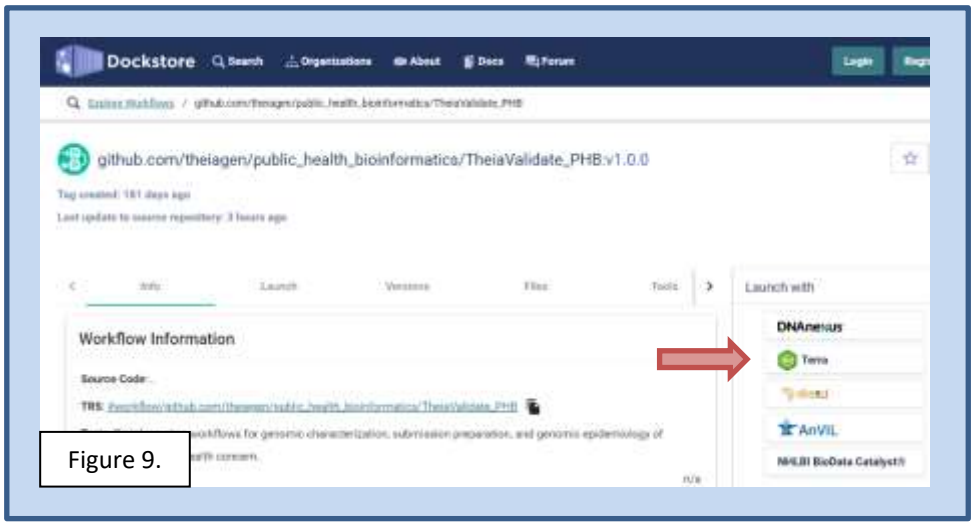

Figure 9.

Figure 10.