Running Theiagen's Rasusa Workflow in Terra to Randomly Downsample Reads

Document TG-RASUSA-01, Version 2

Date: 04/20/2024 | Workflow Versions: PHB v1.3.0 and PHB v2

1. PURPOSE/SCOPE

To standardize the procedure of downsampling read files using Theiagen's Rasusa workflow in Terra. Acceptable data types include both short and long read input files.

NOTE: This workflow serves as the initial step required for Limit of Detection (LOD) validation method.

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.
- Raw read files uploaded to Terra workspace.

REQUIRED WORKFLOW INPUT FILES

- Raw read files
- Terra metadata (tsv) file

3. RELATED DOCUMENTS

Document Number	Document Name
TG-TER-03	Uploading Local or SRA NGS Data & Creating a Results Metadata Table in Terra

4. PROCEDURE

4.1 CREATE A NEW TERRA DATA TABLE

1. For first time using Theiagen's Rasusa PHB Workflow in Terra, see Appendix *10.1 IMPORTING THE RASUSA WORKFLOW FROM DOCKSTORE*
2. In the Terra workspace of interest, *open the data table* to view samples that will be downsampled.
3. Click the *checkbox* next to each sample that will be downsampled.
    a. *Click the down arrow in the top left of the sample table and select "all" to process all samples.*

4. Click *settings* (Figure 1) and *none*, select only *read1* and *read2* for paired read data, click *done* (Figure 2).
    a. For single end read data, select only the *reads* column.

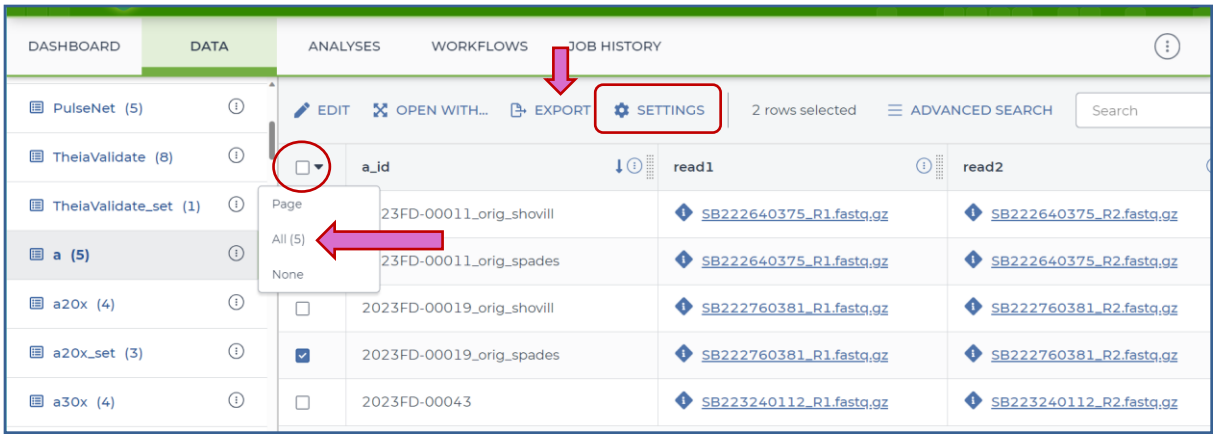

Figure 1

Figure 2

5. Click *export*, *download as tsv* (Figure 1), and *open* the file in excel (Figure 3).
6. Add a data table name suffix in cell A1 by indicating the final coverage, read fraction basepairs, etc that reads will be downsampled e.g. *entity:a_id* is changed to *entidy:a30x_id* to indicate subsampled reads (read#_subsampled) in table a30x will be approximatelt 30X coverage.(Figure 3 and Figure 4)
    a. *Save the file* with a new file name, e.g. a30x.

b. Return to the Terra window and click *import data*, *upload tsv* (Figure 5).
c. In the pop-up window, *click to select* or *drag and drop* the relevant file, click *start import job* (Figure 6)

7. Return to the excel file *and repeat step 6 for every target downsample level* to create separate Terra data tables (e.g. 20X, 30X, 40X, 50X, ect).

	A	B	C
1	entity:a_id	read1	read2
2	2023FD-00011_orig_spades	gs://fc-21(	gs://fc-21(
3	2023FD-00019_orig_spades	gs://fc-21(	gs://fc-21(

Figure 3

	A	B	C
1	entity:a30x_id	read1	read2
2	2023FD-00011_orig_spades	gs://fc-21(	gs://fc-21(
3	2023FD-00019_orig_spades	gs://fc-21(	gs://fc-21(

Figure 4

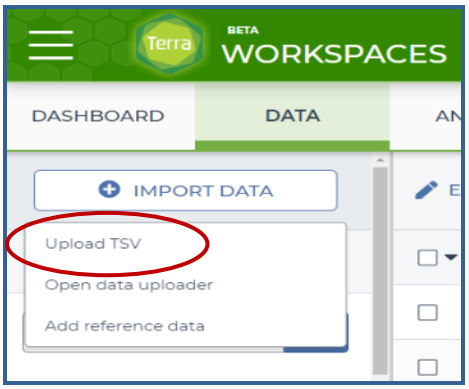

Figure 5

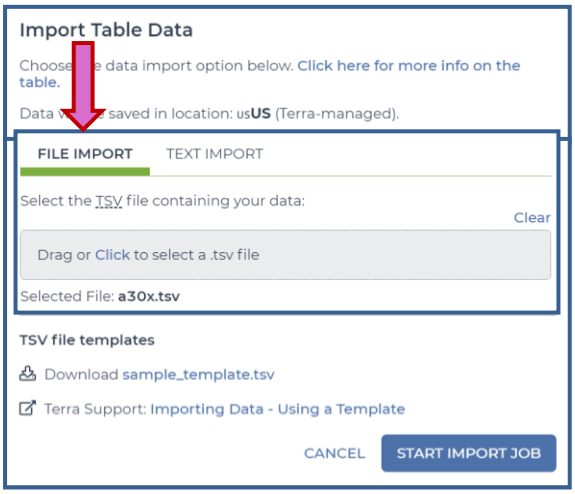

Figure 6

4.2 RUNNING THE RASUSA WOKFLOW

1. In the Terra workspace containing downsampled reads, navigate to the *workflows* tab and open *Rasusa_PHB* (Figure 7).
2. Uncheck *Use call caching* (Figure 8).
3. Choose the latest workflow *version* available (Figure 8a).
4. Select the second bullet to *Run workflow(s) with inputs defined by data table* (Figure 8b)
5. Select the relevant data table name under the *select root* entity type dropdown (Figure 8c)
6. Click *Select data* (Figure 8d)

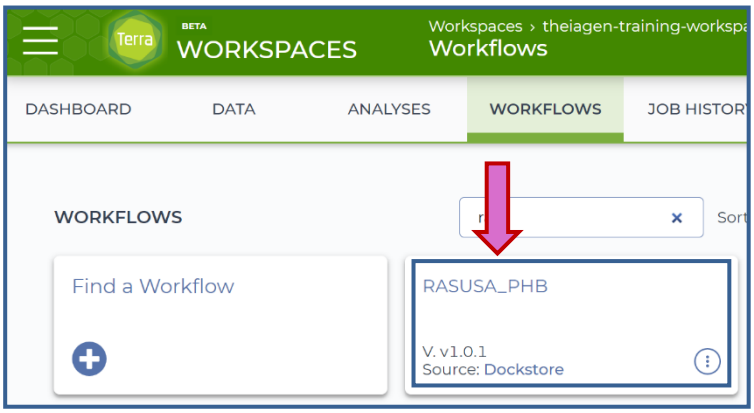

Figure 7

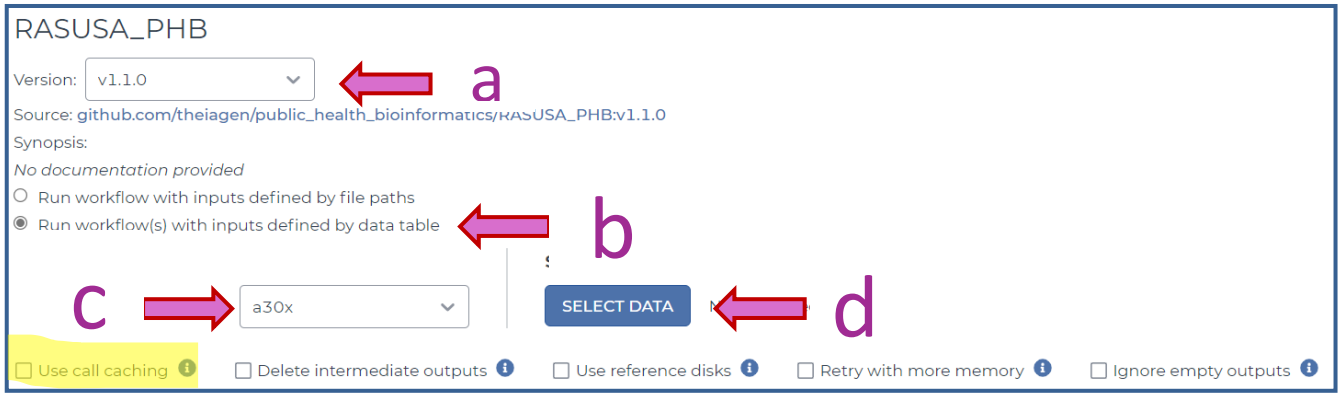

Figure 8

7. In the pop-up window *select the checkbox* for each sample to be included in the analysis (Figure 9a).

a. Click the checkbox dropdown and select *all* to select all samples in the data table. ***Important****: if the checkbox at the top is checked, only the first 100 samples in the data table will be selected.*
b. Optional*:* rename the sample set name to include data and analyst initials, as desired (Figure 9b).
c. Click *ok*.

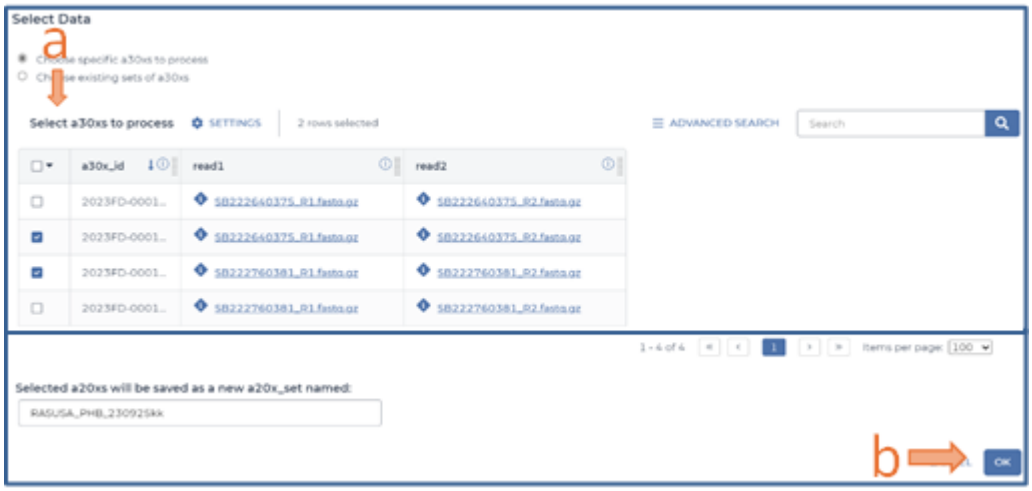

Figure 9

8. In the inputs tab, specify the following variables (Figure 10):
   a. *read 1*: specify the column containing read1 files in the data table (e.g. *this.read1*)
   b. *read2*: specify the column containing read2 files in the data table for paired end data (e.g. *this.read2*).
   c. *samplename*: select the column containing samples ID (e.g. *this.a30x_id*).
   d. only **ONE** of the following:
      i. *coverage*: enter the desired, final coverage (e.g. *30.0*),
         a. If coverage is selected, then *genome_size* is required. Enter the approximate genome size in quotations (e.g "5m"). Acceptable metric suffixes include b, k, m, g and t to indicate base, kilobase, megabase, gigabase and terabase respectively.
      ii. *frac*: enter the final fraction of reads to keep (e.g. *0.5*),

iii. *num*: enter the final number of read pairs (if paired) to keep (e.g. *"5"*)

iv. *Bases*: enter the desired number of final bases (e.g. "5m").

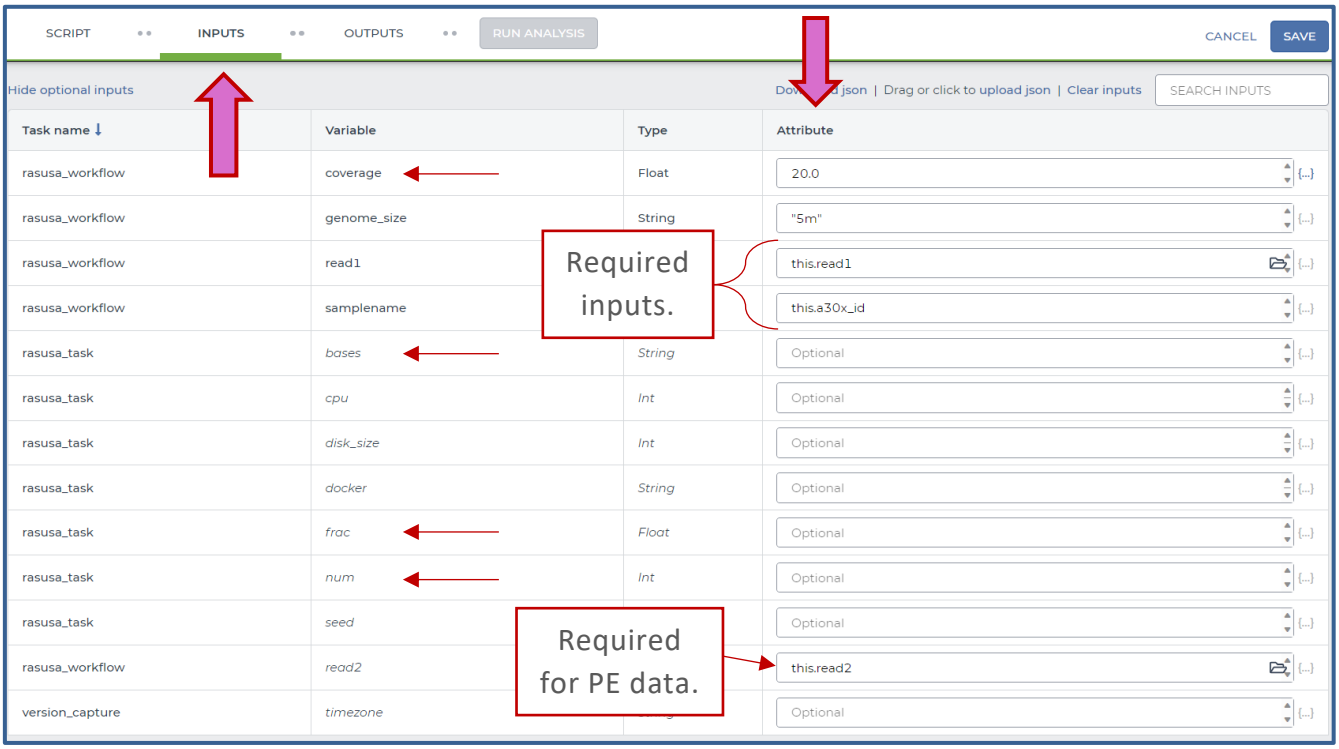

Figure 10

9. Specify outputs by clicking on the *outputs* tab and *use defaults* (Figure 11)
10. Click *save.*
11. Launch the workflow by clicking *Run analysis* enter desired comments and click *launch*.

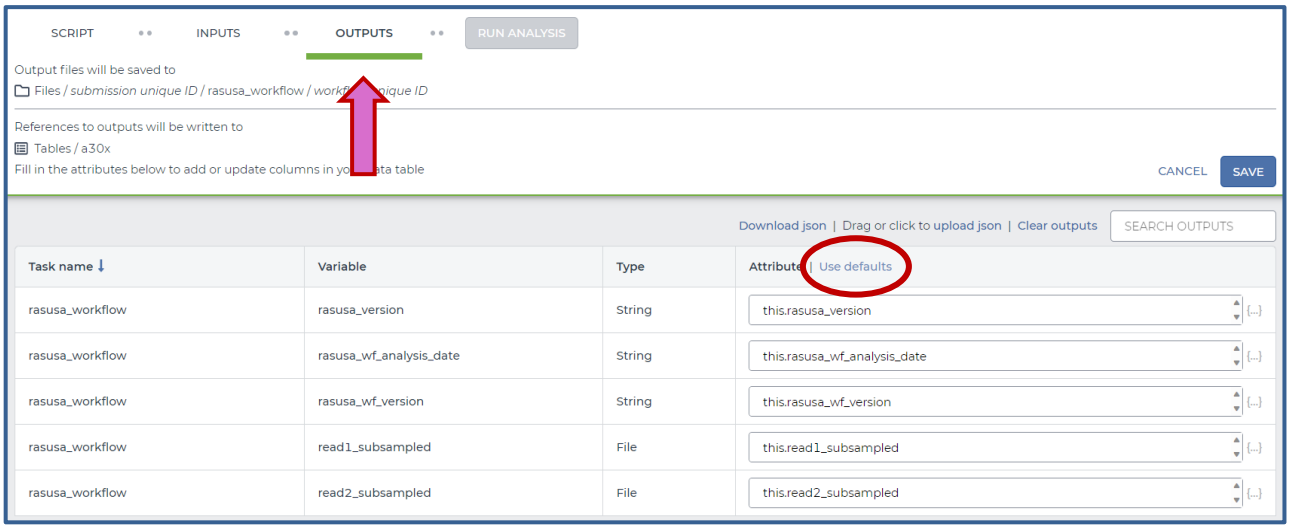

Figure 11

4.3 RASUSA DOWNSAMPLING VERIFICATION

1. In the *data* tab, navigate to the Terra data table containing downsampled reads.
2. Click *settings* (Figure 12) and select *none* to deselect all output columns (Figure 2).

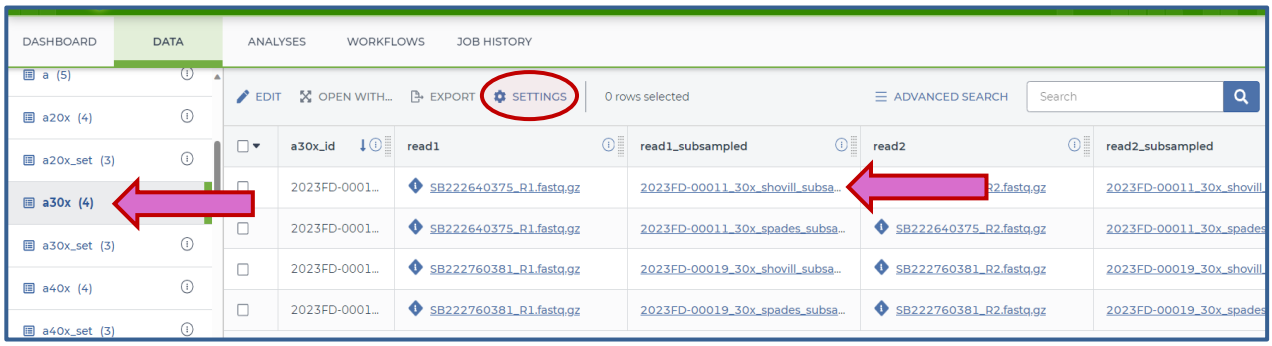

Figure 12

3. To simplify the table. Select the following outputs:
    a. *read1*
    b. *read2*
    c. *read1_subsampled*
    d. *read2_subsampled*

4. Verify downsampling was successful:
    a. Click on the *read1* and *read1_subsampled files* for the first sample (Figure 13), compare file sizes.
        i. *The sampled file should be less than that of the original read file*.
5. Remember to use the downsampled reads for downstream analyses (e.g. *this.read1_subsampled*).

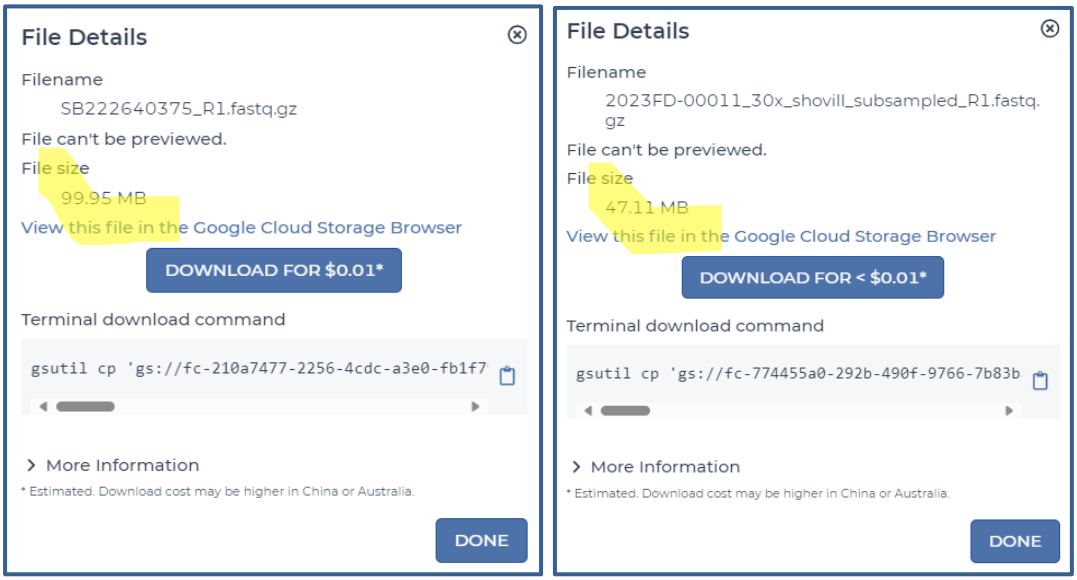

Figure 13

5. QUALITY RECORDS

- Raw read files.
- Subsampled read files.
- Workflow version and input parameters (e.g. Figure 8 and Figure 10).

6. TROUBLESHOOTING

- Consult with internal staff familiar with this procedure or contact support@theiagen.com for troubleshooting inquires.
- For documentation edut requests, contact support@theaigen.com

7. LIMITATIONS

1. Raw read files must be in fasta or fastq file formats.
2. Actual end coverage of subsampled reads may be higher or lower than requested, always check the actual coverage values of subsampled reads. Due to randomness of subsampling, try re-running the workflow again from the original reads for a slightly different coverage result.
3. Attempting to downsample by coverage across species or from a dataset variable in assembly lengths will result in
4. Output read file format will match that of the input format, file formats cannot be converted between fasta or fastq.

8. REFERENCES

1. Hall, M. B., (2022). Rasusa: Randomly subsample sequencing reads to a specified coverage. Journal of Open Source Software, 7(69), 3941, https://doi.org/10.21105/joss.03941

9. REVISION HISTORY

Revision	Version	Date
Document Creation	1	12/2023
Formatting (reference and cross-reference check), updated limitations section, added appendix 10.1	2	4/2024

10. APPENDICES

10.1 IMPORTING THE RASUSA WORKFLOW FROM DOCKSTORE

1. In the Terra workspace of interest, open the workflows tab and click find a workflow (Figure 14)
2. In the pop-up window, click dockstore (Figure 15).

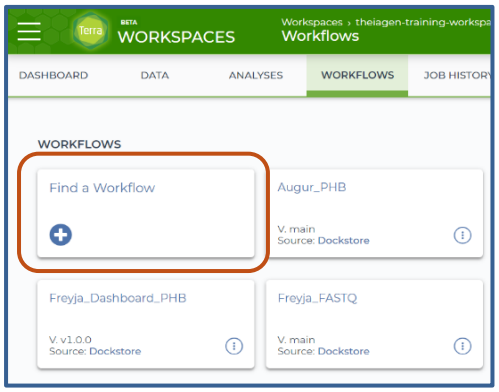

Figure 14

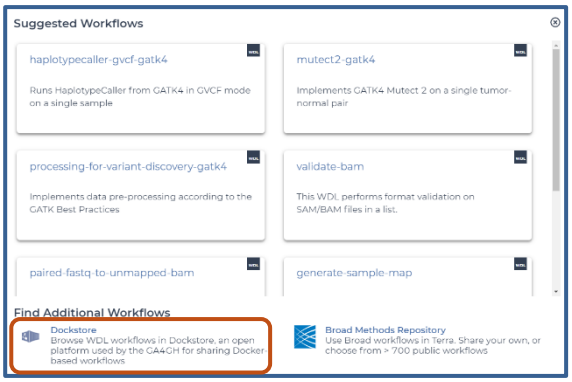

Figure 15

3. Workflows may be found through the search bar or by navigating through the organization if it is known.
4. To find Theiagen's Rasusa PHB Workflow, for example, click *organizations* (Figure 16)
5. In the search bar type *Theiagen* (Figure 17).
6. Click on *view* (Figure 17).and then in the *collection* of interest, see all available workflows(Figure 18).

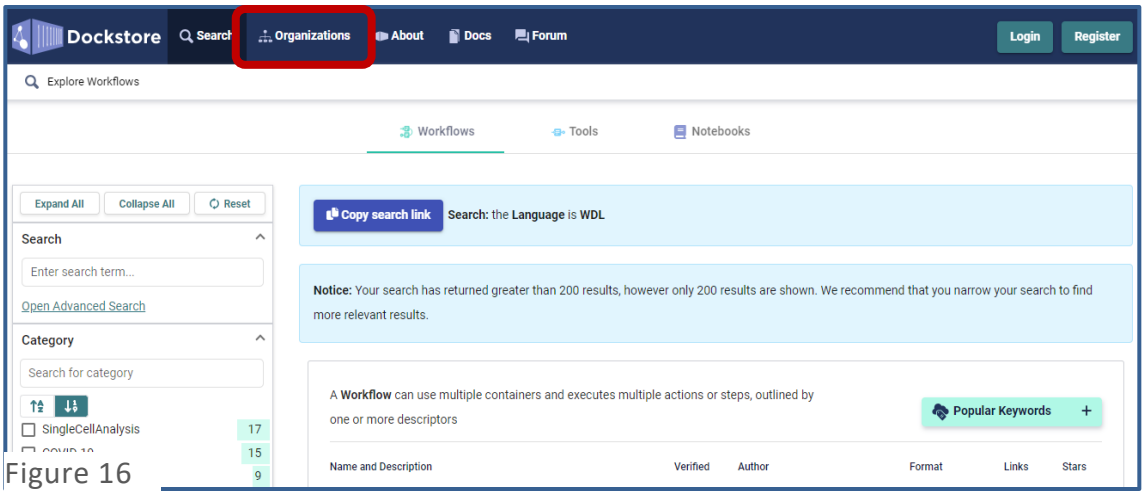

Figure 16

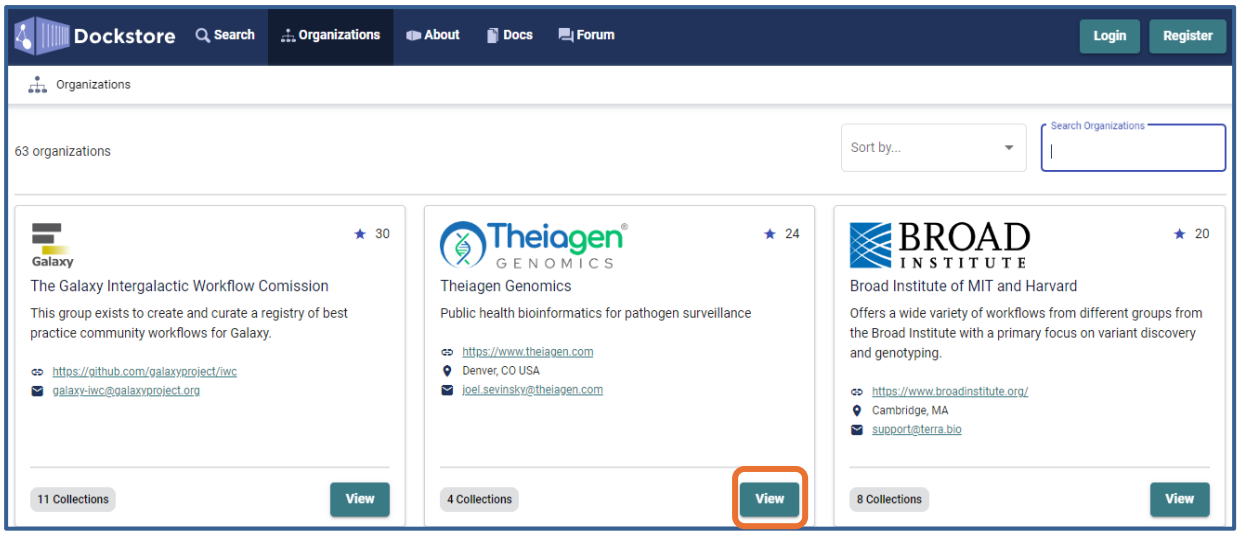

Figure 17

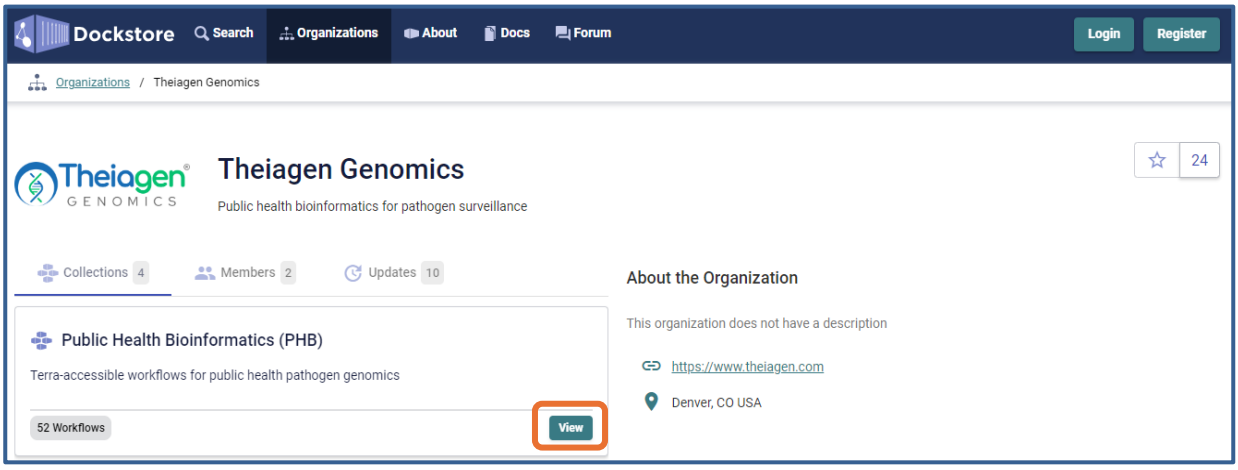

Figure 18

7. Find the *Theiagen's Rasusa PHB Workflow* and click *view* (Figure 19).
8. On the right-hand side, click *Terra* to launch the workflow in Terra (Figure 21).
9. Choose the *destination workspace* in the dropdown and click *import* or *create a new workspace* (Figure 20).

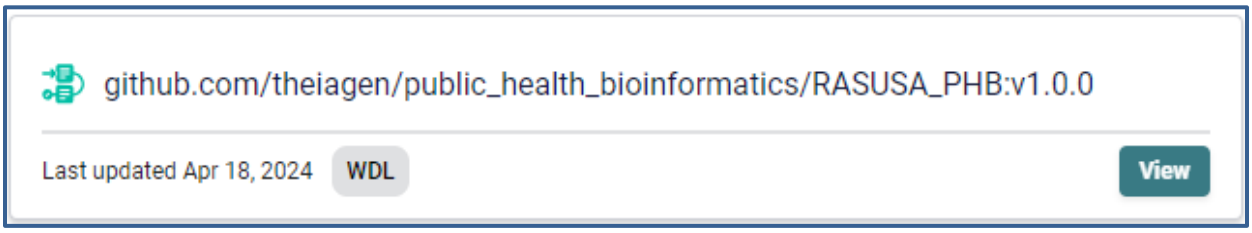

Figure 19

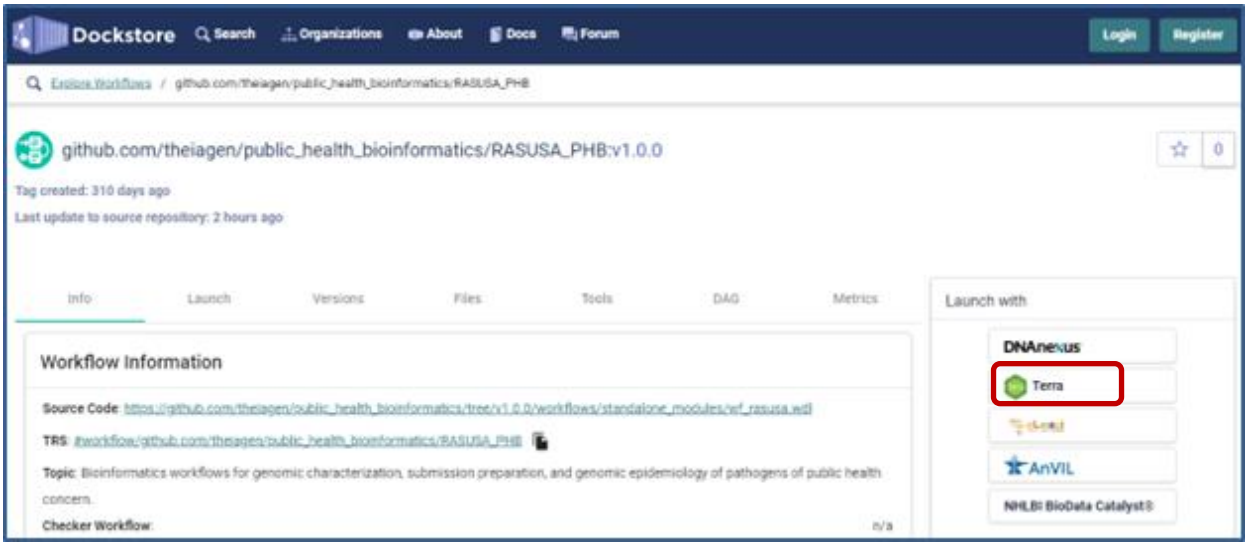

Figure 21

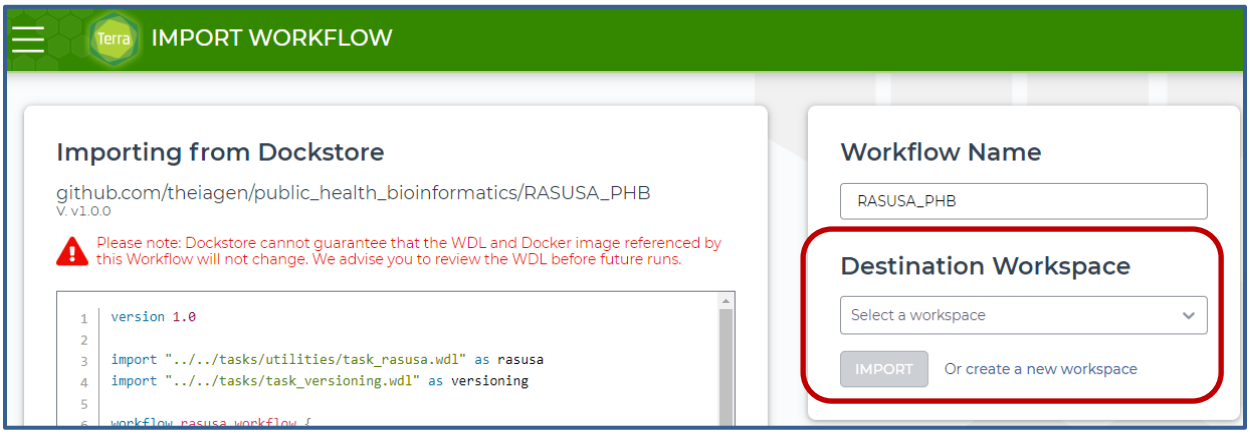

Figure 20