



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.
 - o 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.*



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

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

	a_id	read1	read2
Page	23FD-00011_orig_shovill	SB222640375_R1.fastq.gz	SB222640375_R2.fastq.gz
All (5)	23FD-00011_orig_spades	SB222640375_R1.fastq.gz	SB222640375_R2.fastq.gz
None			
<input type="checkbox"/>	2023FD-00019_orig_shovill	SB222760381_R1.fastq.gz	SB222760381_R2.fastq.gz
<input checked="" type="checkbox"/>	2023FD-00019_orig_spades	SB222760381_R1.fastq.gz	SB222760381_R2.fastq.gz
<input type="checkbox"/>	2023FD-00043	SB223240112_R1.fastq.gz	SB223240112_R2.fastq.gz

Figure 1

Select columns

Show: all **none** Sort: alphabetical

r2_mean_readlength_raw

raw_read_screen

read1

read1_clean

read2

read2_clean

resfinder_db_version

resfinder_docker

resfinder_pheno_table

resfinder_pheno_table_species

resfinder_pointfinder_pheno_table

resfinder_pointfinder_results

resfinder_results

resfinder_sens

SAVE THIS COLUMN SELECTION

CANCEL DONE

Figure 2

- Click **export**, **download as tsv** (Figure 1), and **open** the file in excel (Figure 3).
- 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 **entity:a30x_id** to indicate subsampled reads (read#_subsampled) in table a30x will be approximately 30X coverage.(Figure 3 and Figure 4)
 - Save the file** with a new file name, e.g. a30x.



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

- 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

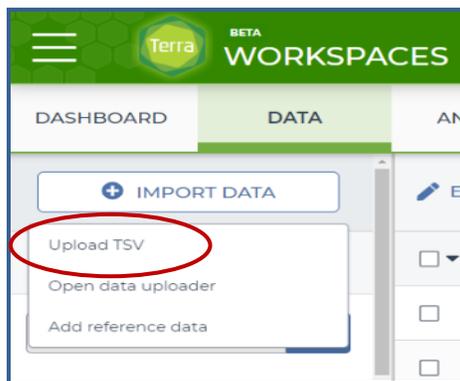


Figure 5

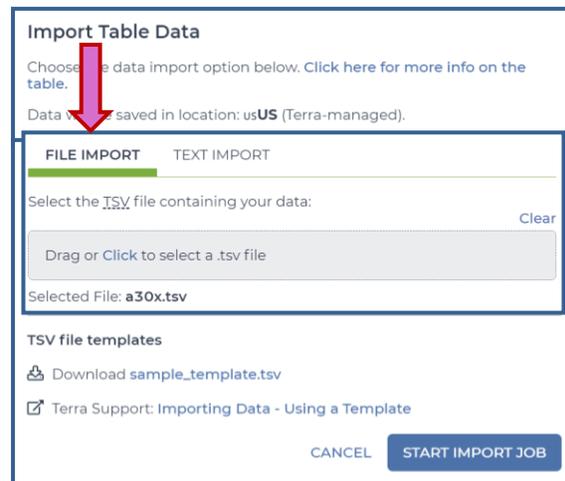


Figure 6

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)

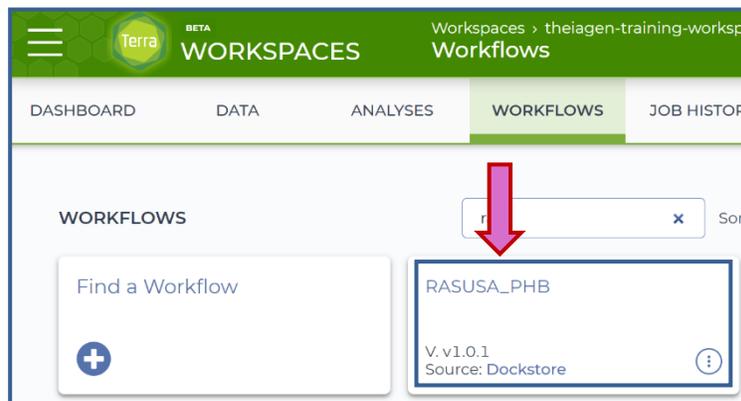


Figure 7

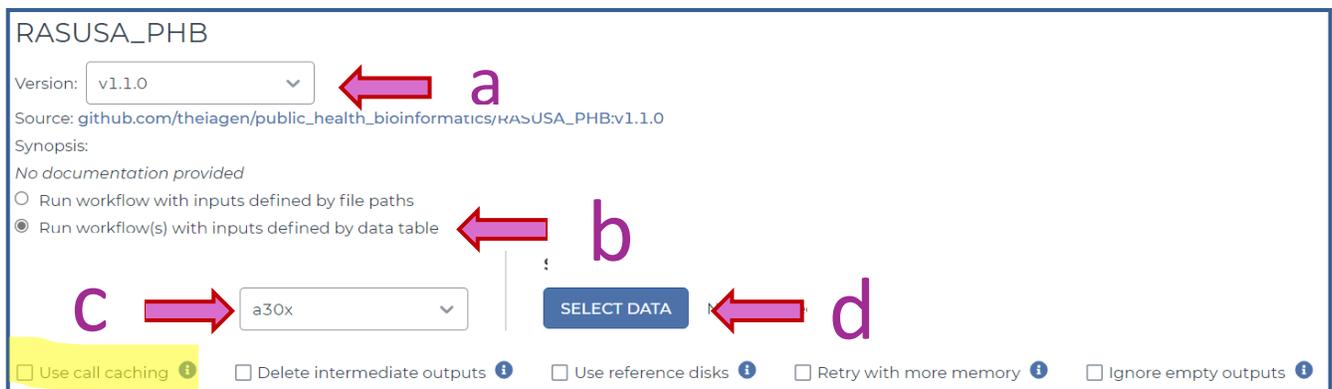


Figure 8

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



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

- 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.
- Optional:** rename the sample set name to include data and analyst initials, as desired (Figure 9b).
- Click `ok`.

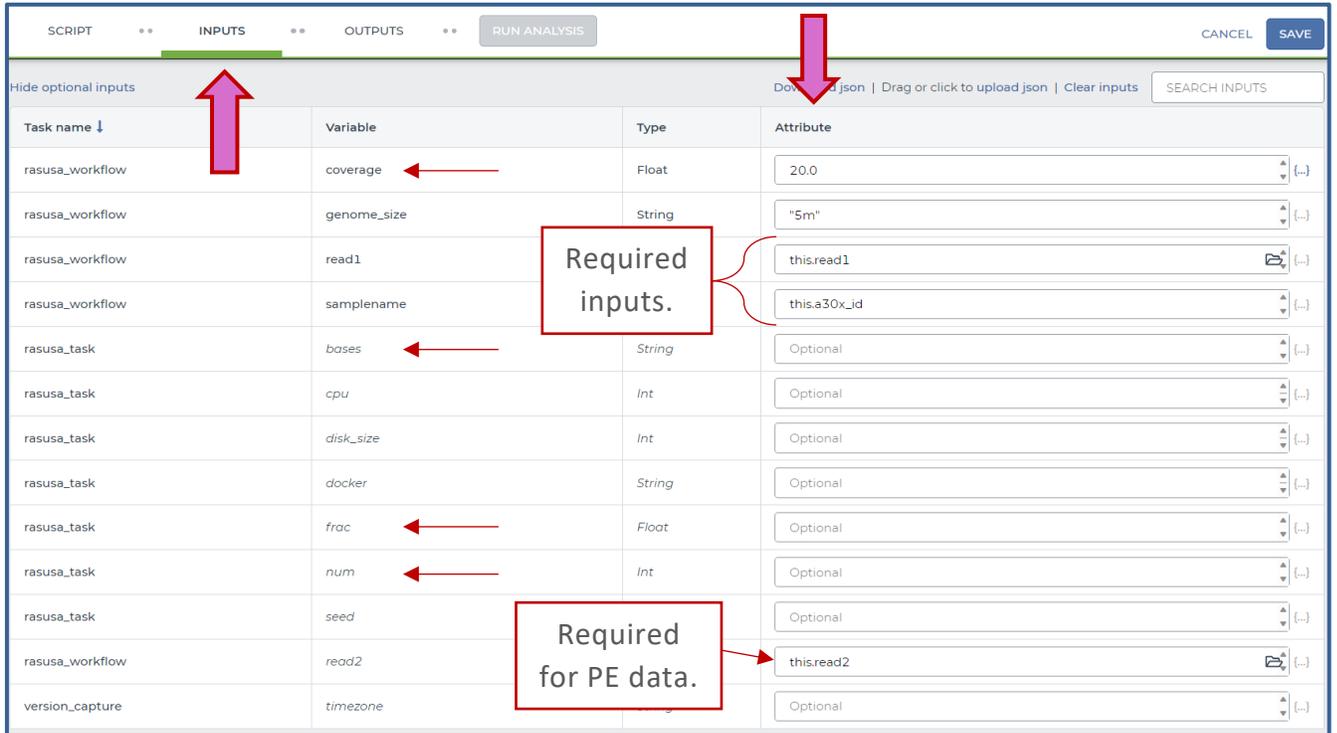
The screenshot shows the 'Select Data' interface. At the top, there are two radio buttons: 'Choose specific a30xs to process' (selected) and 'Choose existing sets of a30xs'. Below this is a table with columns 'a30x_id', 'read1', and 'read2'. The table has 4 rows, with the second and third rows selected. Below the table, there is a text input field for 'Selected a20xs will be saved as a new a20x_set named:' with the value 'RASUSA_PHB_230925kk'. An 'OK' button is visible at the bottom right.

Figure 9

- In the inputs tab, specify the following variables (Figure 10):
 - `read1`: specify the column containing read1 files in the data table (e.g. `this.read1`)
 - `read2`: specify the column containing read2 files in the data table for paired end data (e.g. `this.read2`).
 - `samplename`: select the column containing samples ID (e.g. `this.a30x_id`).
 - only **ONE** of the following:
 - `coverage`: enter the desired, final coverage (e.g. `30.0`),
 - 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.
 - `frac`: enter the final fraction of reads to keep (e.g. `0.5`),

	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

- 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").



Task name ↓	Variable	Type	Attribute
rasusa_workflow	coverage	Float	20.0
rasusa_workflow	genome_size	String	"5m"
rasusa_workflow	read1		this.read1
rasusa_workflow	samplename		this.a30x_id
rasusa_task	bases	String	Optional
rasusa_task	cpu	Int	Optional
rasusa_task	disk_size	Int	Optional
rasusa_task	docker	String	Optional
rasusa_task	frac	Float	Optional
rasusa_task	num	Int	Optional
rasusa_task	seed		Optional
rasusa_workflow	read2		this.read2
version_capture	timezone		Optional

Figure 10

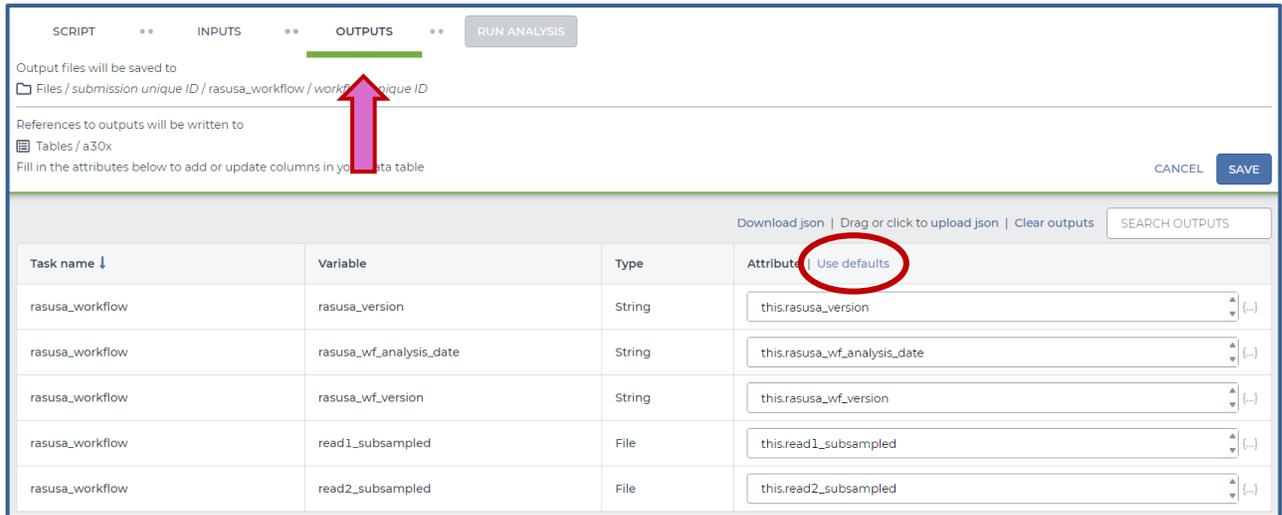
9. Specify outputs by clicking on the `outputs` tab and `use defaults` (Figure 11)
10. Click `save`.
11. Launch the workflow by clicking `Run analysis` enter desired comments and click `launch`.



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



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

References to outputs will be written to
Tables / a30x

Fill in the attributes below to add or update columns in your data table

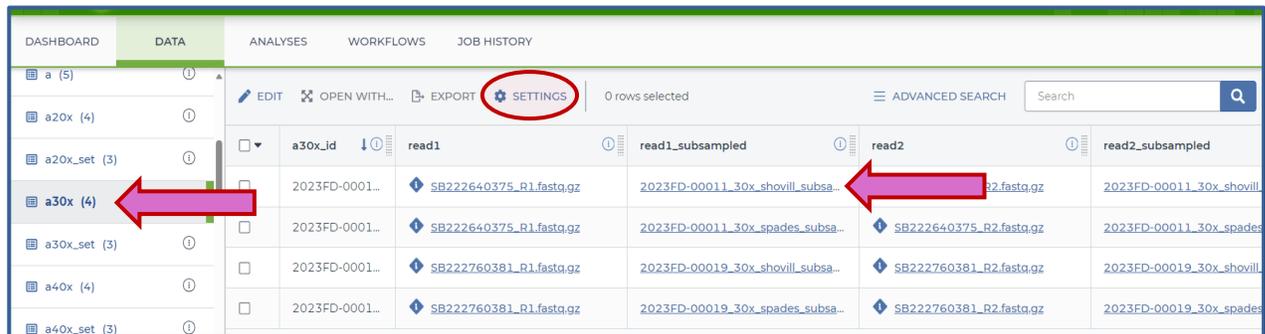
Download json | Drag or click to upload json | Clear outputs SEARCH OUTPUTS

Task name ↓	Variable	Type	Attributes Use defaults
rasusa_workflow	rasusa_version	String	this.rasusa_version
rasusa_workflow	rasusa_wf_analysis_date	String	this.rasusa_wf_analysis_date
rasusa_workflow	rasusa_wf_version	String	this.rasusa_wf_version
rasusa_workflow	read1_subsampled	File	this.read1_subsampled
rasusa_workflow	read2_subsampled	File	this.read2_subsampled

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).



a30x_id	read1	read1_subsampled	read2	read2_subsampled
2023FD-0001...	SB222640375_R1.fastq.gz	2023FD-00011_30x_showill_subsa...	SB222640375_R2.fastq.gz	2023FD-00011_30x_showill...
2023FD-0001...	SB222640375_R1.fastq.gz	2023FD-00011_30x_spades_subsa...	SB222640375_R2.fastq.gz	2023FD-00011_30x_spades...
2023FD-0001...	SB222760381_R1.fastq.gz	2023FD-00019_30x_showill_subsa...	SB222760381_R2.fastq.gz	2023FD-00019_30x_showill...
2023FD-0001...	SB222760381_R1.fastq.gz	2023FD-00019_30x_spades_subsa...	SB222760381_R2.fastq.gz	2023FD-00019_30x_spades...

Figure 12

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

	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

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`).

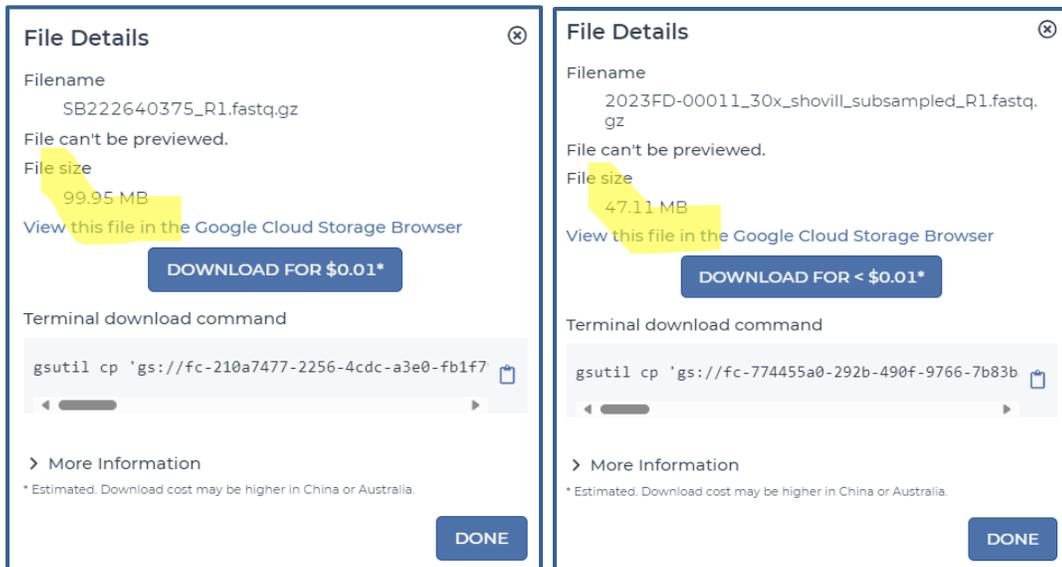


Figure 13

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 inquiries.
- For documentation edut requests, contact support@theaigen.com



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

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



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

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).

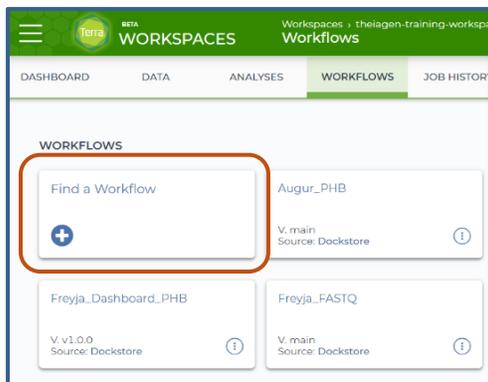


Figure 14

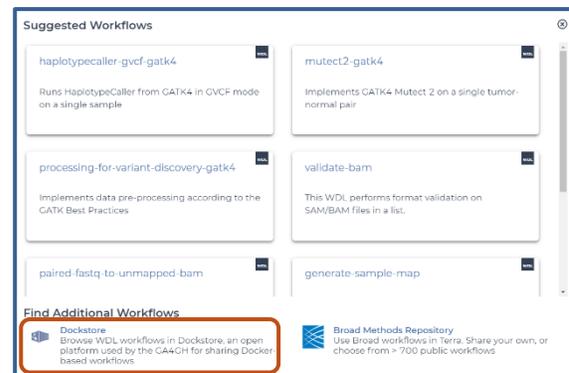


Figure 15

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).

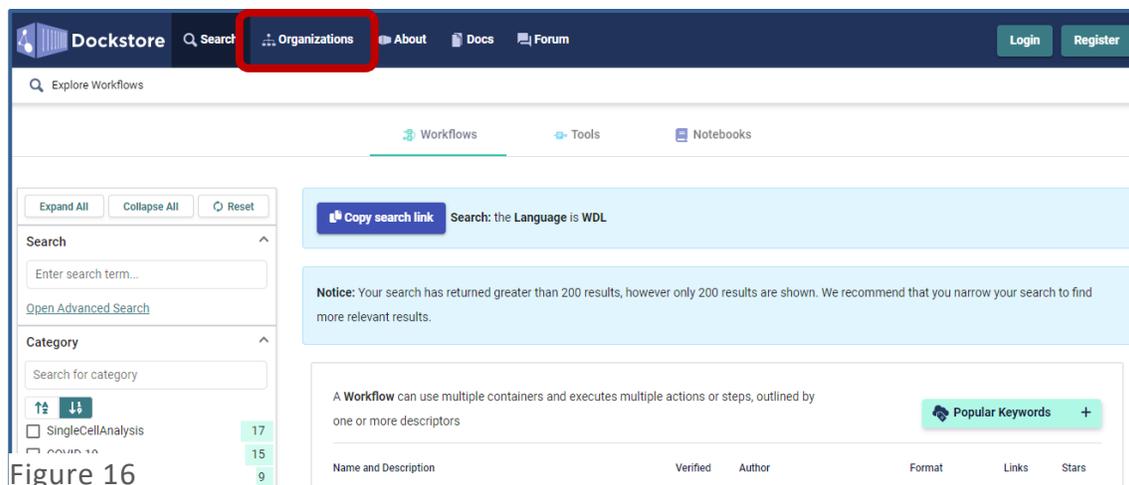


Figure 16



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

The screenshot shows the Dockstore Organizations page. The header includes the Dockstore logo, search bar, and navigation links for Organizations, About, Docs, and Forum. There are Login and Register buttons in the top right. The main content area displays a list of organizations. Theiagen Genomics is the second organization in the list, with a red box around its 'View' button. Theiagen Genomics profile includes its logo, name, a star rating of 24, a description 'Public health bioinformatics for pathogen surveillance', a website link, location 'Denver, CO USA', and email 'joel.sevinsky@theiagen.com'. It also shows '4 Collections' and a 'View' button.

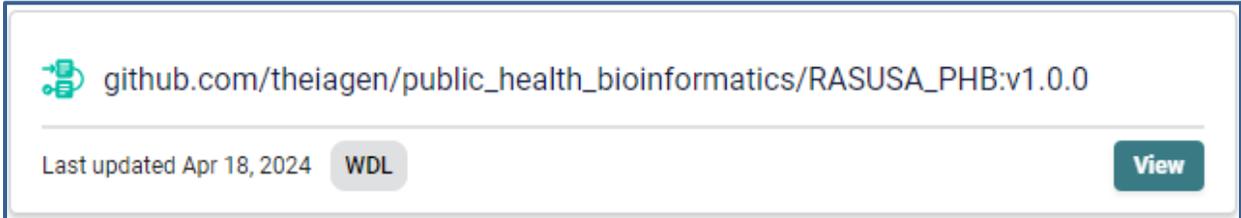
Figure 17

The screenshot shows the Dockstore profile page for Theiagen Genomics. The header includes the Dockstore logo, search bar, and navigation links for Organizations, About, Docs, and Forum. There are Login and Register buttons in the top right. The main content area displays the organization's profile. Theiagen Genomics is the organization, with a star rating of 24. The profile includes its logo, name, a description 'Public health bioinformatics for pathogen surveillance', a website link, location 'Denver, CO USA', and a list of statistics: 'Collections 4', 'Members 2', and 'Updates 10'. The 'Public Health Bioinformatics (PHB)' collection is highlighted with a red box around its 'View' button. The collection description is 'Terra-accessible workflows for public health pathogen genomics' and it shows '52 Workflows'.

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).

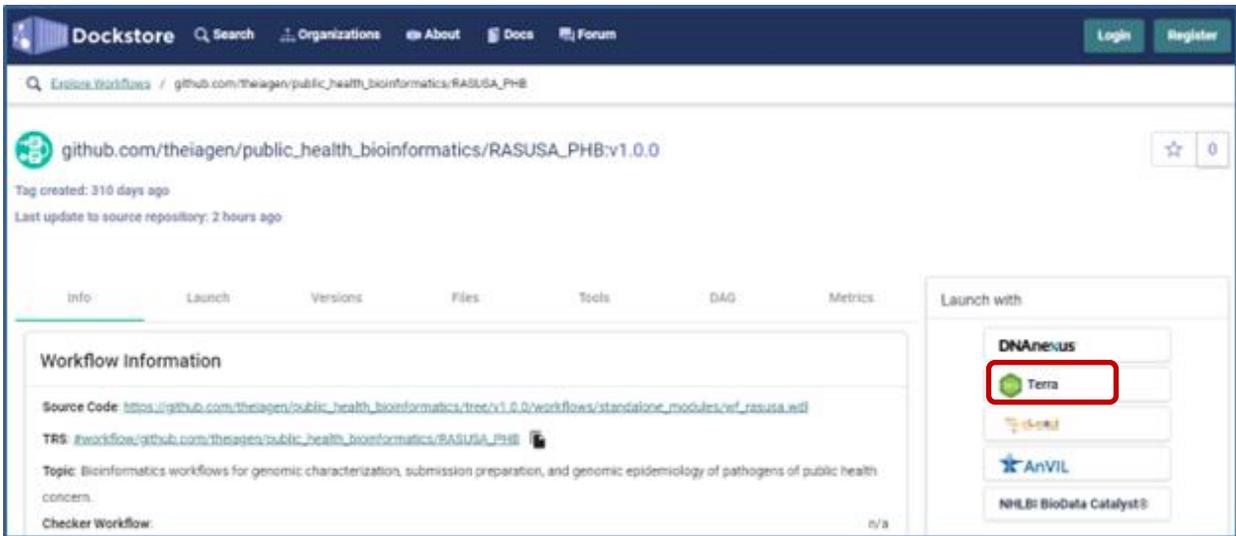
	<h2>Running Theiagen's Rasusa Workflow in Terra to Randomly Downsample Reads</h2>	
	<p>Document TG-RASUSA-01, Version 2</p>	
	<p>Date: 04/20/2024</p>	<p>Workflow Versions: PHB v1.3.0 and PHB v2</p>



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

Last updated Apr 18, 2024 **WDL** **View**

Figure 19



Dockstore Search Organizations About Docs Forum Login Register

github.com/theiagen/public_health_bioinformatics/RASUSA_PHB

Tag created: 310 days ago
Last update to source repository: 2 hours ago

Info Launch Versions Files Tools DAG Metrics

Workflow Information

Source Code: https://github.com/theiagen/public_health_bioinformatics/tree/v1.0.0/workflows/standalone_modules/wf_rasusa.wdl

TRS: [#workflow/github.com/theiagen/public_health_bioinformatics/RASUSA_PHB](#)

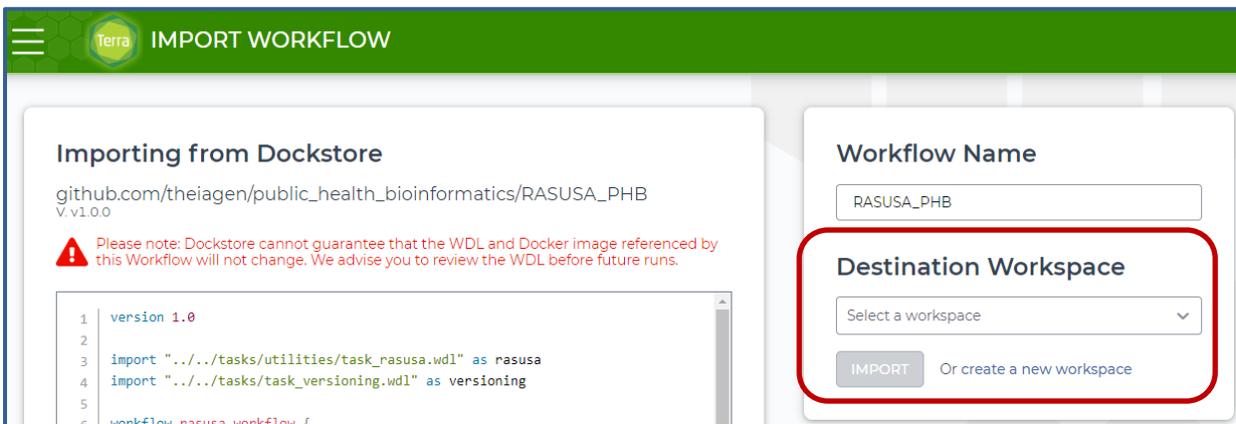
Topic: Bioinformatics workflows for genomic characterization, submission preparation, and genomic epidemiology of pathogens of public health concern.

Checker Workflow: n/a

Launch with

- Terra** (highlighted with a red box)
- DNAnexus
- AnVIL
- NHLBI BioData Catalyst

Figure 21



Terra IMPORT WORKFLOW

Importing from Dockstore

github.com/theiagen/public_health_bioinformatics/RASUSA_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.

```

1 version 1.0
2
3 import "../tasks/utilities/task_rasusa.wdl" as rasusa
4 import "../tasks/task_versioning.wdl" as versioning
5
6 workflow_rasusa_workflow {

```

Workflow Name: RASUSA_PHB

Destination Workspace (highlighted with a red box)

Select a workspace

IMPORT Or create a new workspace

Figure 20