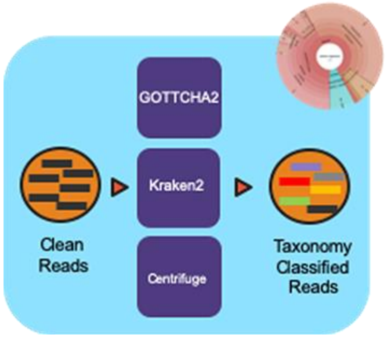
Read-based Taxonomy Classification Workflow (v1.0.1)



Overview

This workflow takes in Illumina sequencing files (single-end or paired-end) and profiles the reads using multiple taxonomic classification tools.

Running the Workflow

Currently, this workflow can be run in NMDC EDGE or from the command line (CLI instructions and requirements are found here).

Tutorial videos on how to run each workflow in NMDC EDGE are found here.

Input

The Metagenome Read-based Taxonomy Classification workflow requires Illumina data and can accept data as an interleaved file or as separate pairs of FASTQ files. Interleaved data will be treated as single-end reads. (It is highly recommended to input clean data from the ReadsQC workflow.)

- **Acceptable file formats:** .fastq, .fq, .fastq.gz, .fq.gz

Details

To create a community profile, this workflow utilizes three taxonomy classification tools: GOTTCHA2, Kraken2, and Centrifuge. These tools vary in levels of specificity and sensitivity. Each tool has a separate reference database. These databases (152 GB) are built into NMDC EDGE. Users can select one, two, or all three of the classification tools to run in the workflow.

Software Versions

- GOTTCHA2 v2.1.6
- Kraken2 v2.0.8
- Centrifuge v1.0.4

Output

Full results are available for each tool at three taxonomic levels (Species, Genus, and Family) in the largest .tsv files for each tool and the top results in the smaller .tsv file. An interactive Krona plot is also generated for the results of each tool.

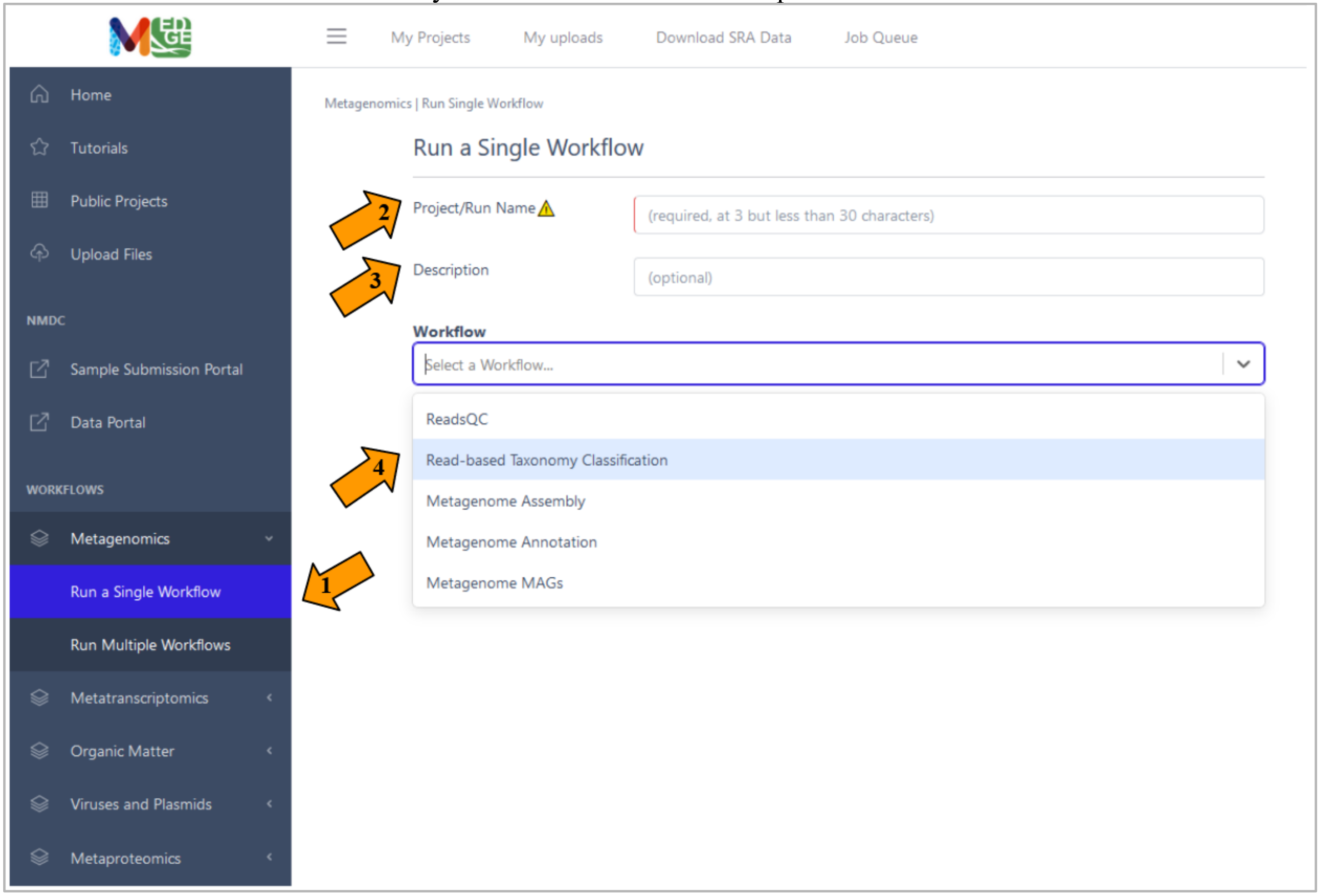
Primary Output Files	Description

Profiling results for each tool	Tabular results of the profile for each tool (.tsv)
Krona plots for each tool	Interactive graphic file (.html)

Running the Read-based Taxonomy Classification Workflow in NMDC EDGE

Select a workflow

1. From the Metagenomics category in the left menu bar, select 'Run a Single Workflow'.
2. Enter a ***unique*** project name with no spaces (underscores are fine).
3. A description is optional, but helpful.
4. Select 'Read-based Taxonomy Classification' from the dropdown menu under Workflow.

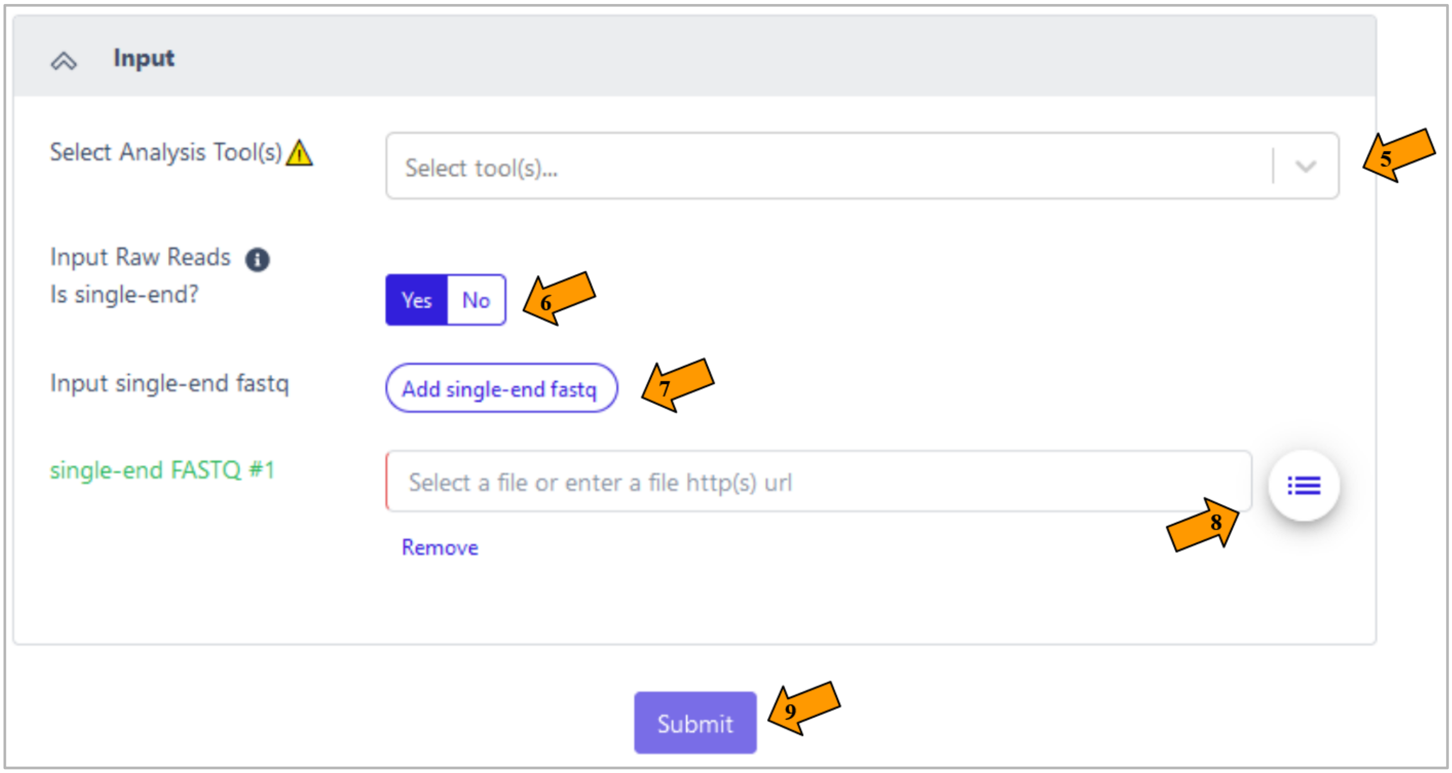


Input

This workflow accepts Illumina data in FASTQ format as the input; the file can be interleaved and can be compressed. This input can be the output from the ReadsQC workflow, and this is recommended. **Acceptable file formats:** .fastq, .fq, .fastq.gz, .fq.gz

5. Select your analysis tool using the drop-down menu. Users can select one, two, or all three of the classification tools to run in the workflow.
6. The default setting is for the raw data to be in an interleaved format (paired reads interleaved into one file). If the raw data is paired reads in separate files (forward and reverse), click 'No'.

7. Additional data files (of the same type–interleaved or separate) can be added with the button below.
8. Click the button to the right of the input blank for data to select the data file for the analysis. (If there are separate files, there will be two input blanks.) A box called ‘Select a File’ will open to allow the user to find the desired file(s) from previously run projects, the public data folder, or files uploaded by the user.
9. Then click ‘Submit’.



Input

Select Analysis Tool(s) ⚠ — Select tool(s)...

Input Raw Reads
Is single-end? — [✓ Yes] No

Input single-end fastq — Add single-end fastq

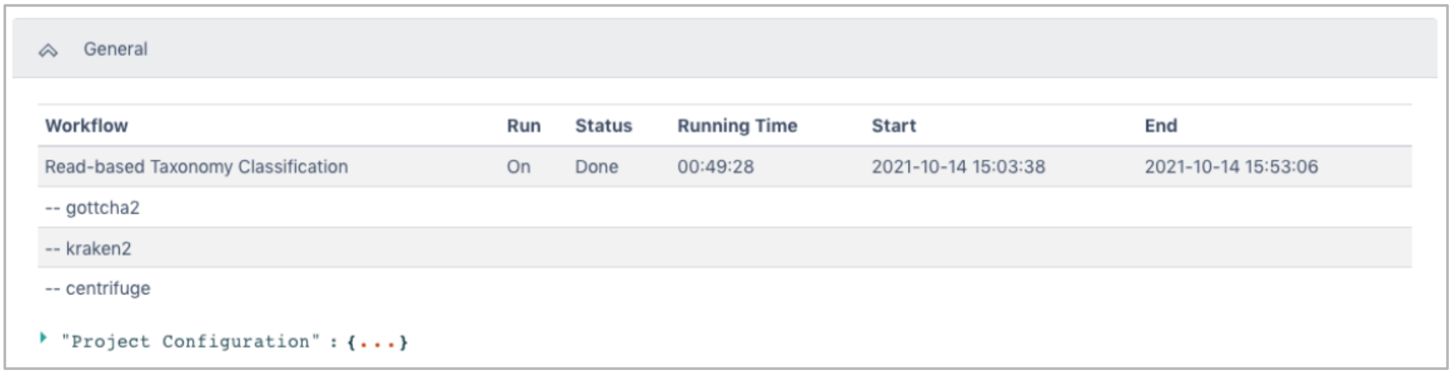
single-end FASTQ #1 — Select a file or enter a file http(s) url

Remove

Submit

Output

The General section of the output shows which workflow and which tools were run and the run time information.

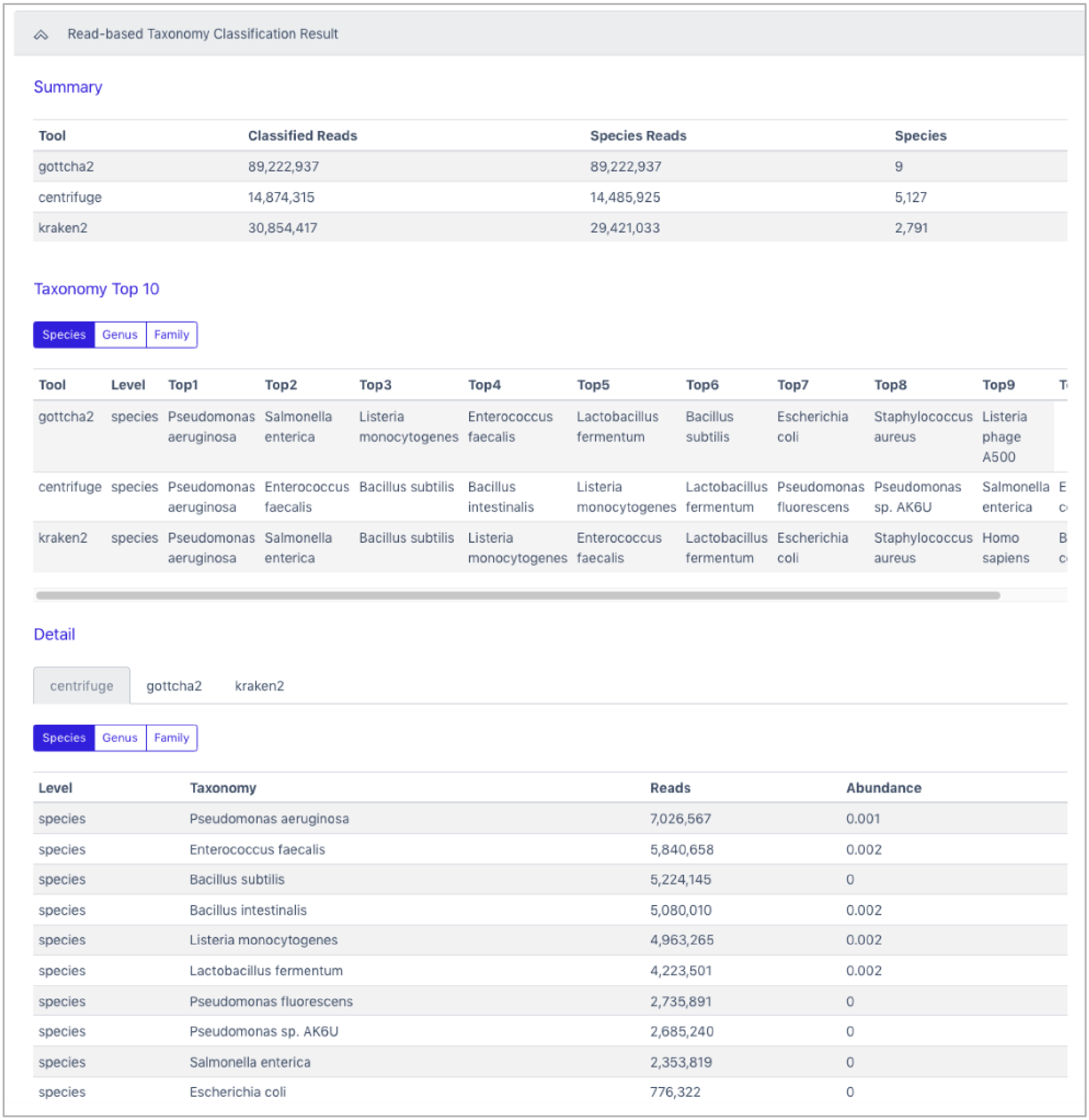


General

Workflow	Run	Status	Running Time	Start	End
Read-based Taxonomy Classification	On	Done	00:49:28	2021-10-14 15:03:38	2021-10-14 15:53:06
-- gottcha2					
-- kraken2					
-- centrifuge					

```
"Project Configuration" : {...}
```

The Read-based Taxonomy Classification Results section has a summary section at the top and results for each tool at three levels of taxonomy in the Taxonomy Top 10 section. The Detail section has classified reads results and relative abundance results for each tool at three levels of taxonomy.



Read-based Taxonomy Classification Result

Summary

Tool	Classified Reads	Species Reads	Species
gottcha2	89,222,937	89,222,937	9
centrifuge	14,874,315	14,485,925	5,127
kraken2	30,854,417	29,421,033	2,791

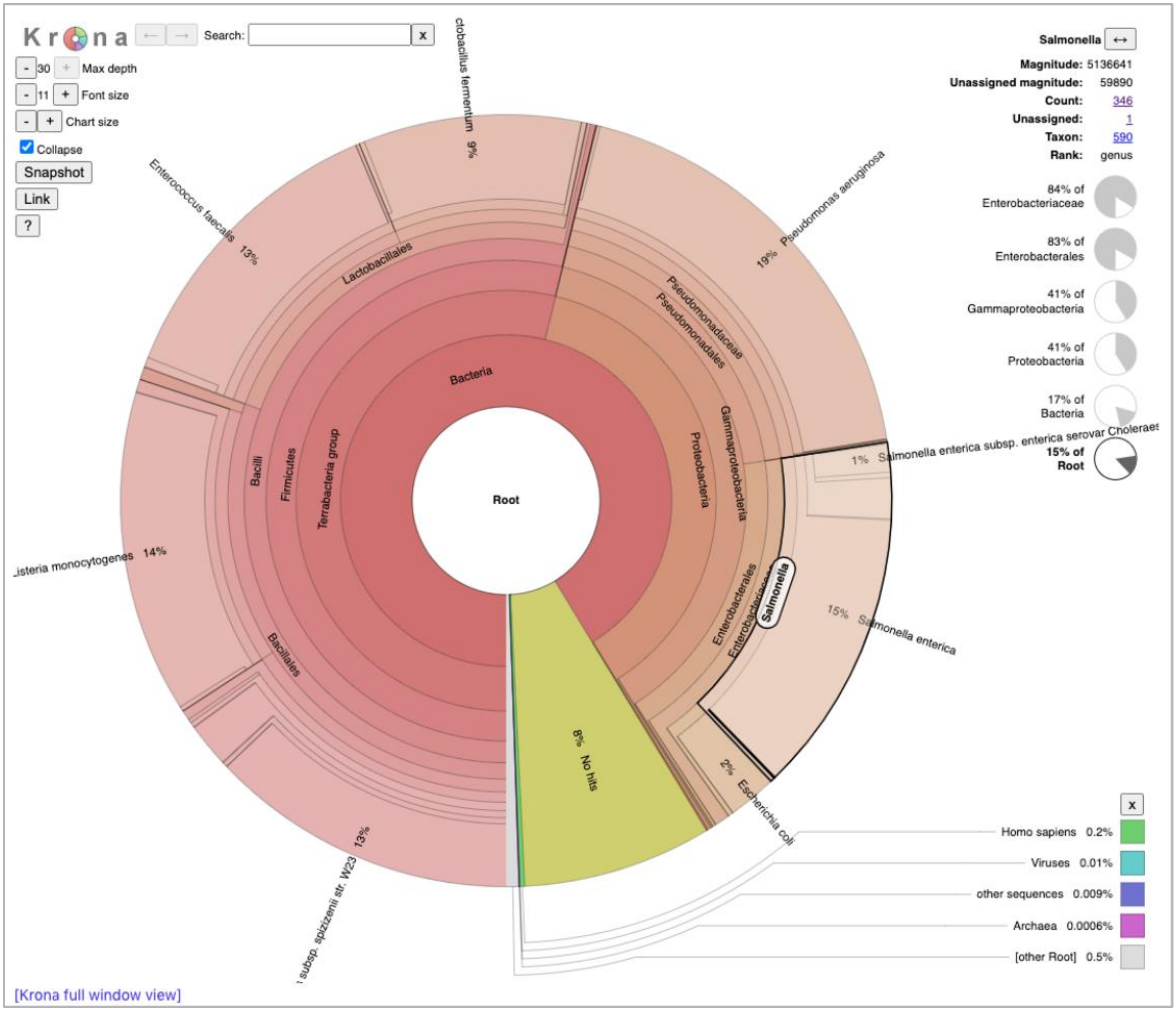
Taxonomy Top 10

Tool	Level	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	Top9	T
gottcha2	species	Pseudomonas aeruginosa	Salmonella enterica	Listeria monocytogenes	Enterococcus faecalis	Lactobacillus fermentum	Bacillus subtilis	Escherichia coli	Staphylococcus aureus	Listeria phage A500	
centrifuge	species	Pseudomonas aeruginosa	Enterococcus faecalis	Bacillus subtilis	Bacillus intestinalis	Listeria monocytogenes	Lactobacillus fermentum	Pseudomonas fluorescens	Pseudomonas sp. AK6U	Salmonella enterica	E c
kraken2	species	Pseudomonas aeruginosa	Salmonella enterica	Bacillus subtilis	Listeria monocytogenes	Enterococcus faecalis	Lactobacillus fermentum	Escherichia coli	Staphylococcus aureus	Homo sapiens	B c

Detail

Level	Taxonomy	Reads	Abundance
species	Pseudomonas aeruginosa	7,026,567	0.001
species	Enterococcus faecalis	5,840,658	0.002
species	Bacillus subtilis	5,224,145	0
species	Bacillus intestinalis	5,080,010	0.002
species	Listeria monocytogenes	4,963,265	0.002
species	Lactobacillus fermentum	4,223,501	0.002
species	Pseudomonas fluorescens	2,735,891	0
species	Pseudomonas sp. AK6U	2,685,240	0
species	Salmonella enterica	2,353,819	0
species	Escherichia coli	776,322	0

The Detail section also provides an interactive Krona plot for each tool.



- 30 + Max depth

- 11 + Font size

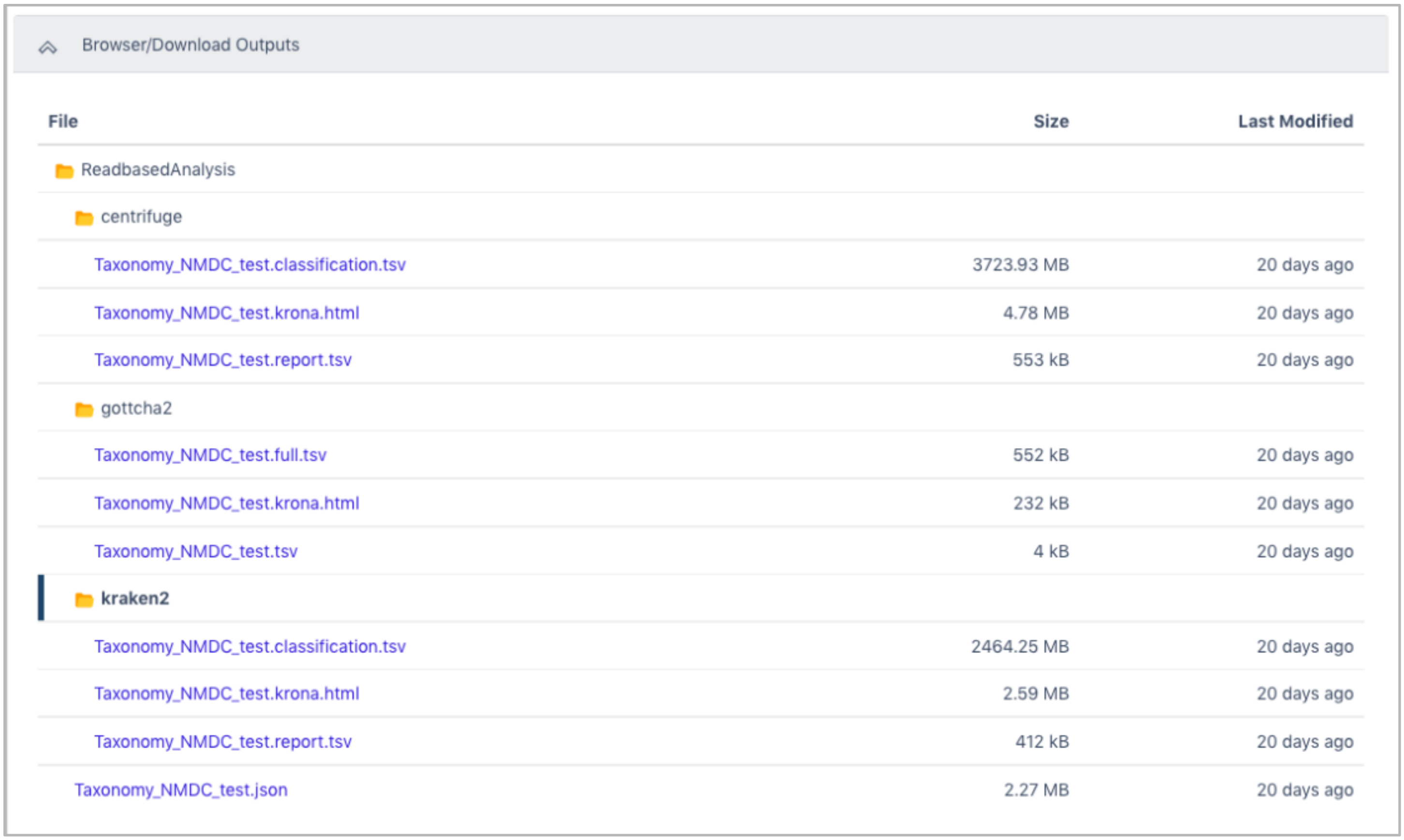
- + Chart size

☑ Collapse

Salmonella ↔

Magnitude: 5136641
Unassigned magnitude: 59890
Count: 346
Unassigned: 1
Taxon: 590
Rank: genus

The Browser/Download Output section provides output files available to download. Each tool has a separate folder for the results from that tool. Full tabular results are in the largest .tsv file and the interactive Krona plots (.html files) open in a separate browser window.



Browser/Download Outputs

File	Size	Last Modified
ReadbasedAnalysis		
centrifuge		
Taxonomy_NMDC_test.classification.tsv	3723.93 MB	20 days ago
Taxonomy_NMDC_test.krona.html	4.78 MB	20 days ago
Taxonomy_NMDC_test.report.tsv	553 kB	20 days ago
gottcha2		
Taxonomy_NMDC_test.full.tsv	552 kB	20 days ago
Taxonomy_NMDC_test.krona.html	232 kB	20 days ago
Taxonomy_NMDC_test.tsv	4 kB	20 days ago
kraken2		
Taxonomy_NMDC_test.classification.tsv	2464.25 MB	20 days ago
Taxonomy_NMDC_test.krona.html	2.59 MB	20 days ago
Taxonomy_NMDC_test.report.tsv	412 kB	20 days ago
Taxonomy_NMDC_test.json	2.27 MB	20 days ago