MultiQC: New features and flexible data parsing

Phil Ewels

Senior Product Manager for OSS @ Seqera

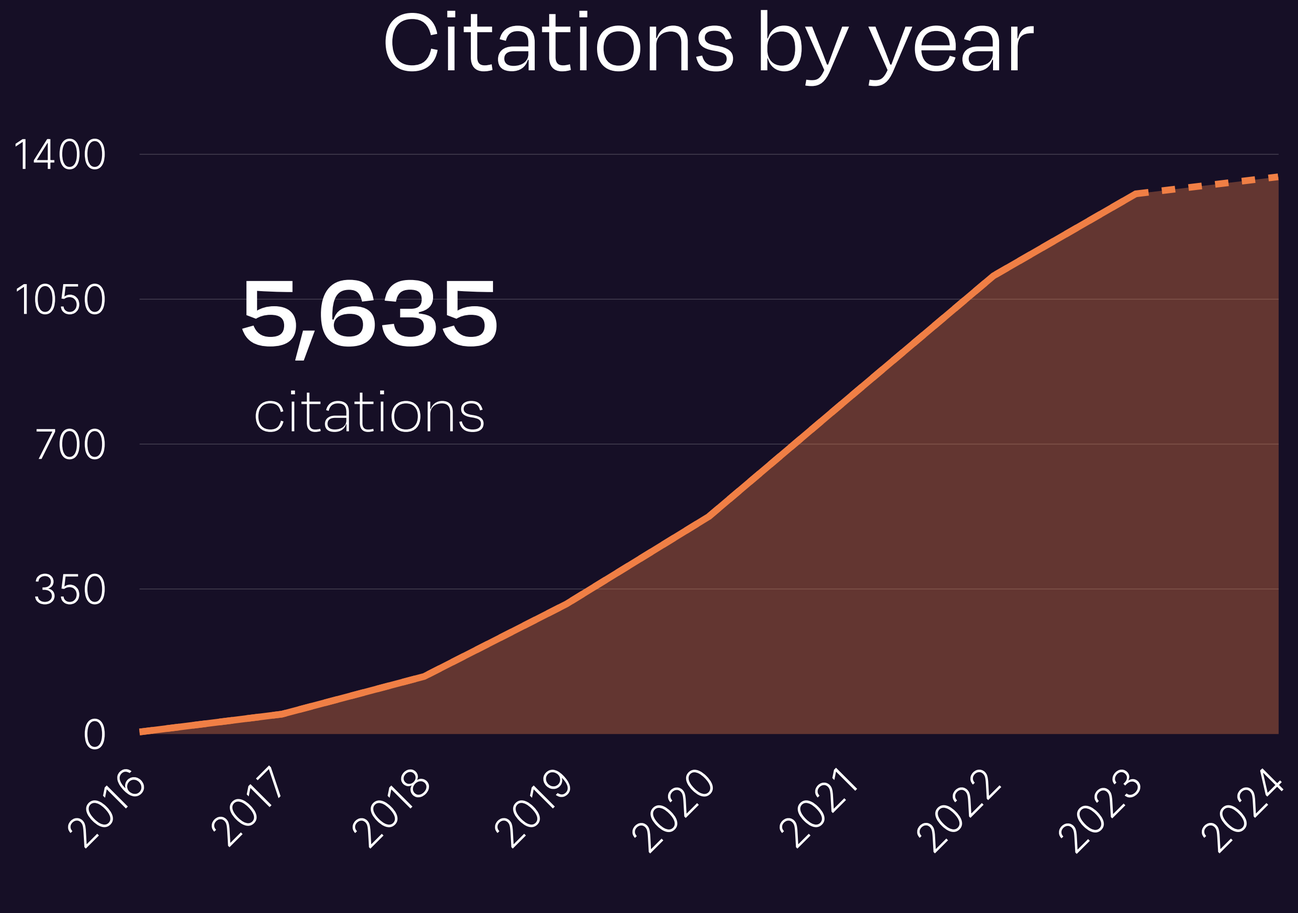
SUMMIT 2024

multiqc

1,216
GitHub Stars

+25K
Runs per day

+1.5M
Downloads



Modern software engineering - for science

1. Static Typing and Unit Tests

2. Config Validation with Pydantic

3. Performance Improvements

4. New Plotly Plots

5. Sample Grouping

6. MultiQC as a Library

Command Line Interface

Web interface

MultiQC Plugins

Custom Content

Notebooks and Scripts

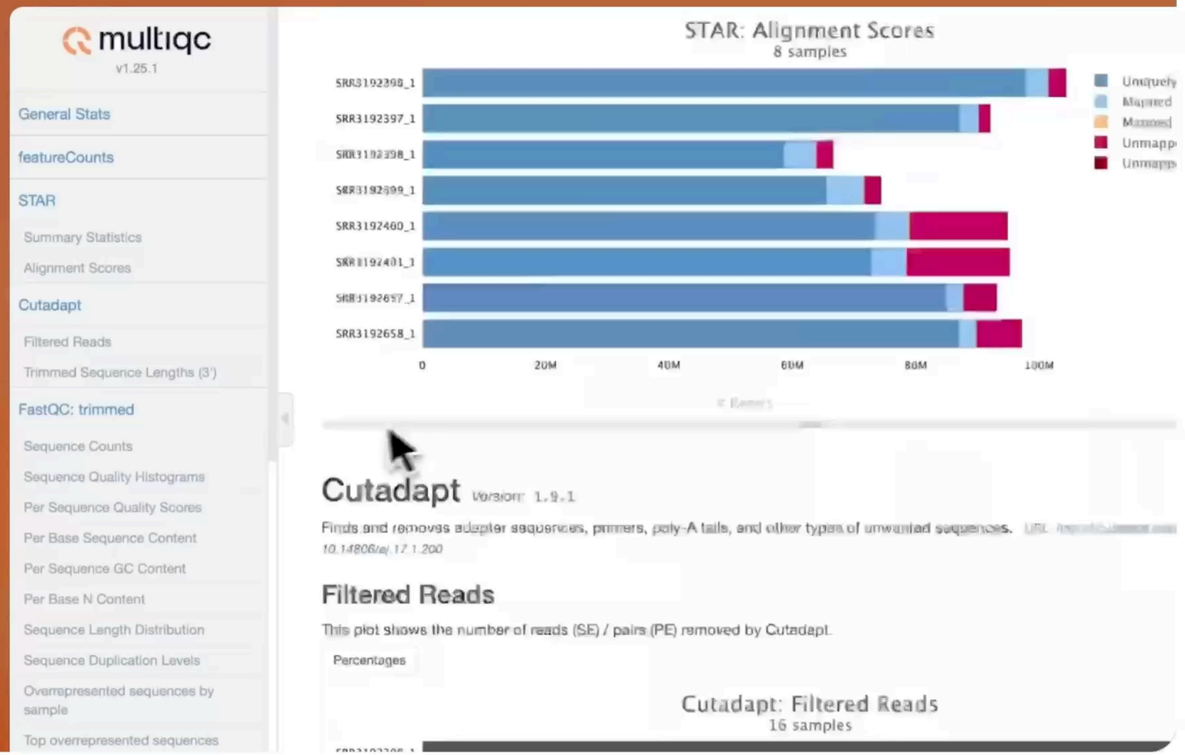
Demo

https://seqera.io/multiqc/

multiqc

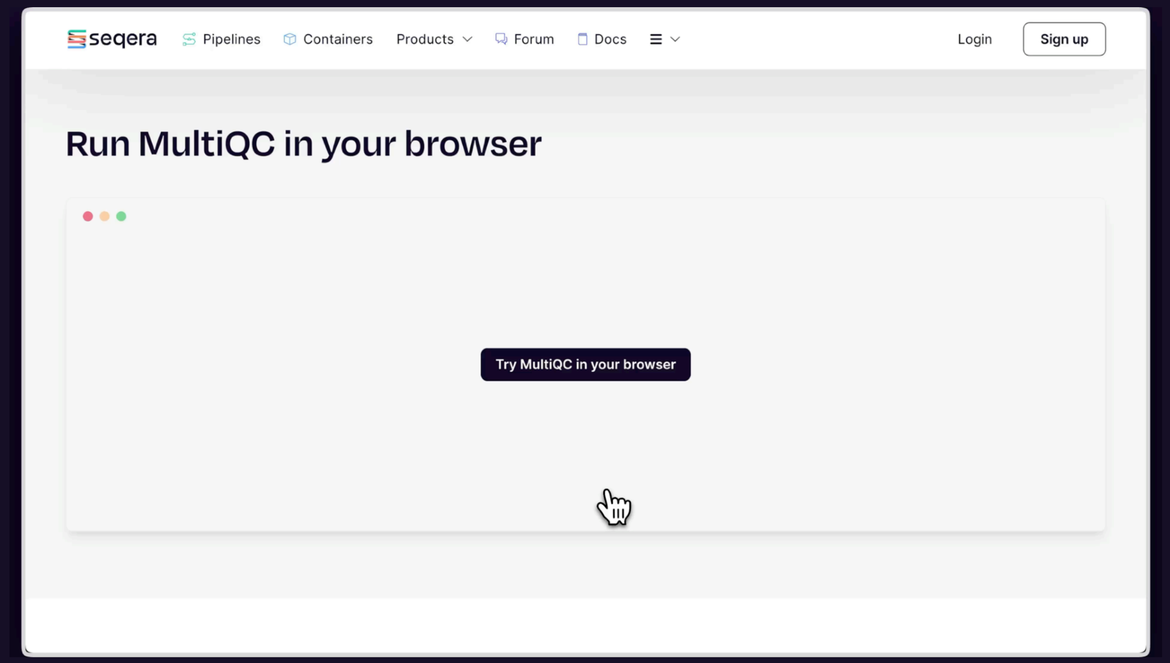
Open-source tool to aggregate bioinformatic analyses results.

Read documentation



Ready to get started?

Generate reports in your browser



https://seqera.io/multiqc/

- Generate reports in your browser, no installations necessary
- Point and click, doesn't use the terminal
- Uses WebAssembly

multiqc_config.yml

```yaml
table_sample_merge:
  "Read 1":
    - type: regex
      pattern: "_1$"
  "Read 2":
    - type: regex
      pattern: "_2$"
```

```
❯ multiqc .

/// MultiQC 🎃 v1.26.dev0

            config | Loading config settings from: multiqc_config.yml
       file_search | Search path: ./part_1
         searching | ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 145/145
             fastp | Found 48 reports
            fastqc | Found 96 reports
     write_results | Data         : multiqc_data
     write_results | Report       : multiqc_report.html
           multiqc | MultiQC complete
```

multiqc

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2024-10-27, 18:11 CET based on data in:

- `/data/fastp`
- `/data/fastqc`
- `/data/multiqc_config.yml`

Welcome! Not sure where to start? Watch a tutorial video *(6:06)* don't show again

General Statistics

Copy table | Configure columns | Scatter plot | Violin plot | Showing $^{0}/_{48}$ rows and $^{8}/_{13}$ columns. | Export as CSV

Sample Name	% Duplication	Reads After Filtering	GC content	% PF	% Adapter	Dups	GC	Seqs
SAMPLE_01	17.2 %	1.2 M	51.9 %	57.1 %	9.0 %	61.2 %	52.5 %	2.0 M
SAMPLE_02	46.2 %	1.7 M	38.4 %	78.7 %	5.0 %	91.2 %	38.0 %	2.2 M
SAMPLE_03	48.4 %	1.7 M	39.0 %	77.2 %	5.4 %	90.0 %	39.0 %	2.2 M
SAMPLE_04	44.0 %	1.5 M	38.4 %	78.9 %	4.8 %	90.7 %	38.0 %	1.9 M
SAMPLE_05	46.0 %	1.7 M	38.5 %	78.5 %	5.0 %	91.0 %	38.0 %	2.1 M
SAMPLE_06	45.6 %	1.7 M	38.3 %	77.8 %	4.6 %	90.8 %	38.0 %	2.1 M
SAMPLE_07	48.2 %	2.0 M	38.4 %	79.2 %	5.3 %	91.7 %	38.0 %	2.5 M
SAMPLE_08	48.5 %	1.9 M	38.5 %	79.6 %	5.5 %	91.4 %	38.0 %	2.3 M
SAMPLE_09	37.6 %	1.2 M	42.4 %	66.4 %	10.5 %	86.8 %	44.0 %	1.9 M
SAMPLE_10	45.4 %	1.6 M	38.3 %	78.9 %	5.1 %	90.8 %	38.0 %	2.0 M
SAMPLE_11	49.3 %	2.1 M	38.4 %	79.4 %	5.2 %	91.7 %	38.0 %	2.6 M
SAMPLE_12	45.2 %	1.6 M	38.3 %	77.5 %	4.5 %	90.5 %	38.0 %	2.1 M

Toolbox

```
❯ multiqc .

/// MultiQC 🎃 v1.26.dev0

      file_search | Search pa
        searching | ━━━━━━━━━━
   custom_content | pct_magic: Found 48 general statistics columns
            fastp | Found 48 reports
     write_results | Data       : multiqc_data
     write_results | Report     : multiqc_report.html
          multiqc | MultiQC complete
```

pct_magic_mqc.tsv

```
# plot_type: generalstats
Sample       % Magic
SAMPLE_01    57.99087052
SAMPLE_02    39.12145114
SAMPLE_03    36.14175885
SAMPLE_04    78.25359712
SAMPLE_05    35.47539651
```

fastp

fastqc

pct_magic_mqc.tsv

```
custom_content | pct_magic: Found 48 General Statistics columns
```



pct_magic_mqc.tsv

multiqc

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2024-10-27, 18:15 CET based on data in: `/Users/ewels/GitHub/ewels/multiqc-demo-summit-2024/part_2`

Welcome! Not sure where to start? Watch a tutorial video *(6:06)* don't show again

General Statistics

Copy table | Configure columns | Scatter plot | Violin plot | Showing 0/48 rows and 6/8 columns. | Export as CSV

Sample Name	% Magic	% Duplication	Reads After Filtering	GC content	% PF	% Adapter
SAMPLE_01	58.0	17.2 %	1.2 M	51.9 %	57.1 %	9.0 %
SAMPLE_02	39.1	46.2 %	1.7 M	38.4 %	78.7 %	5.0 %
SAMPLE_03	36.1	48.4 %	1.7 M	39.0 %	77.2 %	5.4 %
SAMPLE_04	78.3	44.0 %	1.5 M	38.4 %	78.9 %	4.8 %
SAMPLE_05	35.5	46.0 %	1.7 M	38.5 %	78.5 %	5.0 %
SAMPLE_06	1.9	45.6 %	1.7 M	38.3 %	77.8 %	4.6 %
SAMPLE_07	53.1	48.2 %	2.0 M	38.4 %	79.2 %	5.3 %
SAMPLE_08	24.8	48.5 %	1.9 M	38.5 %	79.6 %	5.5 %
SAMPLE_09	0.2	37.6 %	1.2 M	42.4 %	66.4 %	10.5 %
SAMPLE_10	6.3	45.4 %	1.6 M	38.3 %	78.9 %	5.1 %
SAMPLE_11	40.3	49.3 %	2.1 M	38.4 %	79.4 %	5.2 %
SAMPLE_12	82.4	45.2 %	1.6 M	38.3 %	77.5 %	4.5 %
SAMPLE_13	47.3	36.9 %	0.1 M	46.0 %	62.4 %	43.7 %
SAMPLE_14	9.8	33.7 %	0.4 M	38.0 %	27.5 %	4.5 %

run_multiqc.py

```python
import multiqc

# Load data
multiqc.parse_logs('./fastp')

# Write the report
multiqc.write_report()
```

```
❯ python run_multiqc.py
```

run_multiqc.py

```python
# Fetch the custom data
reads = {}
for samp, data in multiqc.get_module_data(module='fastp').items():
    reads[samp] = {
      'Reads Before Filtering': data['summary']['before_filtering']['total_reads']
    }

# Add new column to the General Stats table
fastp_module = multiqc.report.modules[0]
fastp_module.general_stats_addcols(data_by_sample=reads)
```

multiqc

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2024-10-27, 18:48 CET based on data in: `/Users/ewels/GitHub/ewels/multiqc-demo-summit-2024/part_1/fastp`

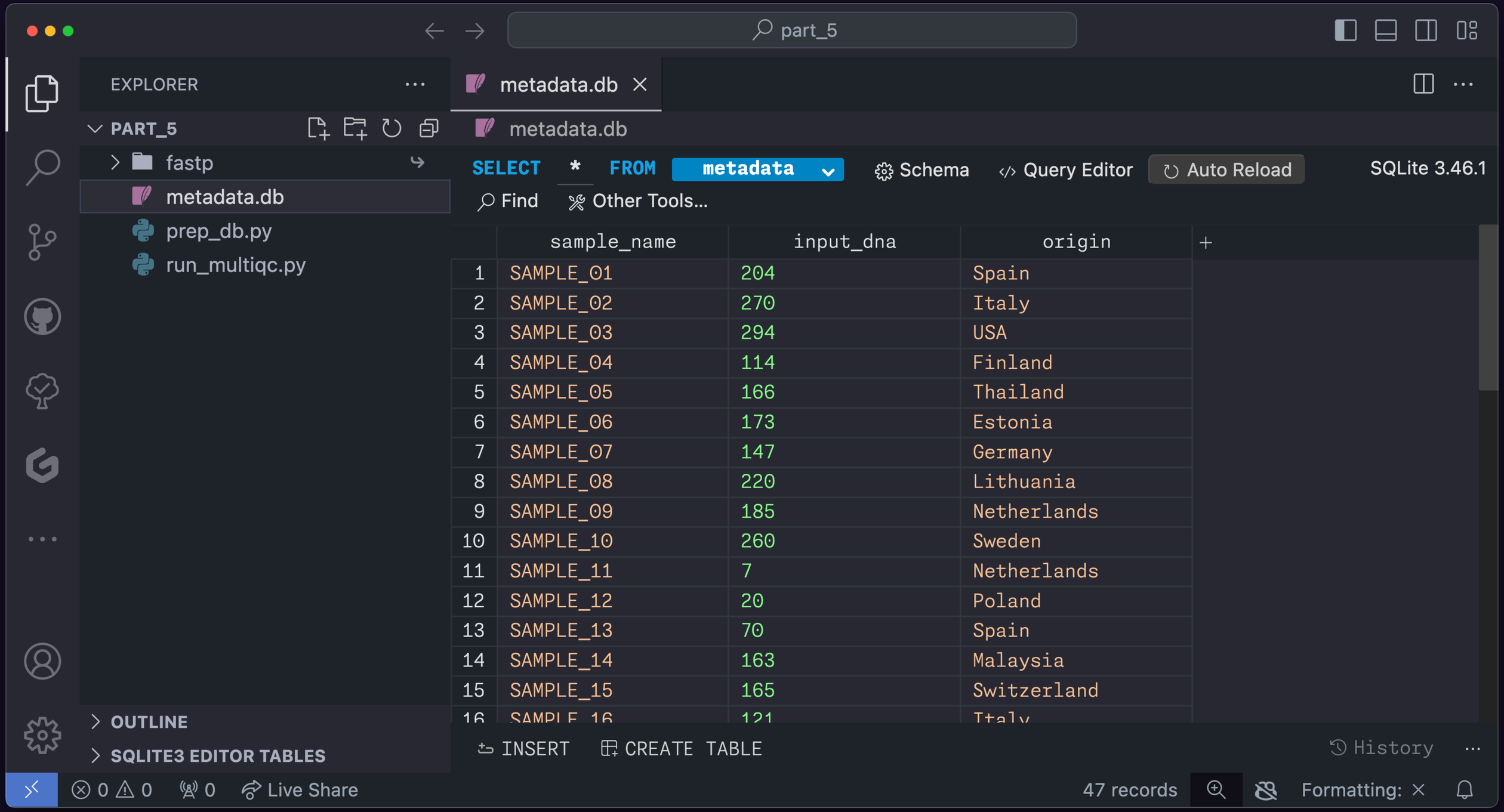
Welcome! Not sure where to start? Watch a tutorial video *(6:06)* don't show again

General Statistics

Copy table | Configure columns | Scatter plot | Violin plot | Showing 0/48 rows and 6/8 columns. | Export as CSV

Sample Name	% Duplication	Reads After Filtering	GC content	% PF	% Adapter	Reads Before Filtering
SAMPLE_01	17.2 %	1.2 M	51.9 %	57.1 %	9.0 %	2 028 184
SAMPLE_02	46.2 %	1.7 M	38.4 %	78.7 %	5.0 %	2 204 200
SAMPLE_03	48.4 %	1.7 M	39.0 %	77.2 %	5.4 %	2 235 448
SAMPLE_04	44.0 %	1.5 M	38.4 %	78.9 %	4.8 %	1 873 764
SAMPLE_05	46.0 %	1.7 M	38.5 %	78.5 %	5.0 %	2 118 860
SAMPLE_06	45.6 %	1.7 M	38.3 %	77.8 %	4.6 %	2 137 680
SAMPLE_07	48.2 %	2.0 M	38.4 %	79.2 %	5.3 %	2 497 804
SAMPLE_08	48.5 %	1.9 M	38.5 %	79.6 %	5.5 %	2 328 994
SAMPLE_09	37.6 %	1.2 M	42.4 %	66.4 %	10.5 %	1 874 658
SAMPLE_10	45.4 %	1.6 M	38.3 %	78.9 %	5.1 %	1 986 068
SAMPLE_11	49.3 %	2.1 M	38.4 %	79.4 %	5.2 %	2 599 022
SAMPLE_12	45.2 %	1.6 M	38.3 %	77.5 %	4.5 %	2 073 494
SAMPLE_13	36.9 %	0.1 M	46.0 %	62.4 %	43.7 %	136 210
SAMPLE_14	33.7 %	0.4 M	38.0 %	27.5 %	4.5 %	1 444 594

Toolbox


part_5
EXPLORER
metadata.db
PART_5
fastp
metadata.db
prep_db.py
run_multiqc.py
metadata.db
SELECT * FROM metadata
Schema
Query Editor
Auto Reload
SQLite 3.46.1
Find
Other Tools...
sample_name input_dna origin
1 SAMPLE_01 204 Spain
2 SAMPLE_02 270 Italy
3 SAMPLE_03 294 USA
4 SAMPLE_04 114 Finland
5 SAMPLE_05 166 Thailand
6 SAMPLE_06 173 Estonia
7 SAMPLE_07 147 Germany
8 SAMPLE_08 220 Lithuania
9 SAMPLE_09 185 Netherlands
10 SAMPLE_10 260 Sweden
11 SAMPLE_11 7 Netherlands
12 SAMPLE_12 20 Poland
13 SAMPLE_13 70 Spain
14 SAMPLE_14 163 Malaysia
15 SAMPLE_15 165 Switzerland
16 SAMPLE_16 121 Italy
OUTLINE
SQLITE3 EDITOR TABLES
INSERT
CREATE TABLE
History
0 0
0
Live Share
47 records
Formatting:

run_multiqc.py

```
# Fetch from database
metadata = {}
cx = sqlite3.connect('metadata.db')
for row in cx.cursor().execute('SELECT * FROM metadata'):
    metadata[row[0]] = {
        'Input DNA (ng)': row[1],
        'Sample Origin': row[2]
    }

# Add data to report
metadata_module = multiqc.BaseMultiqcModule()
metadata_module.general_stats_addcols(data_by_sample=metadata)
multiqc.report.modules.append(metadata_module)
```

Fetch data

Add to report

multiqc

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2024-10-27, 19:03 CET based on data in: `/Users/ewels/GitHub/ewels/multiqc-demo-summit-2024/part_1/fastp`

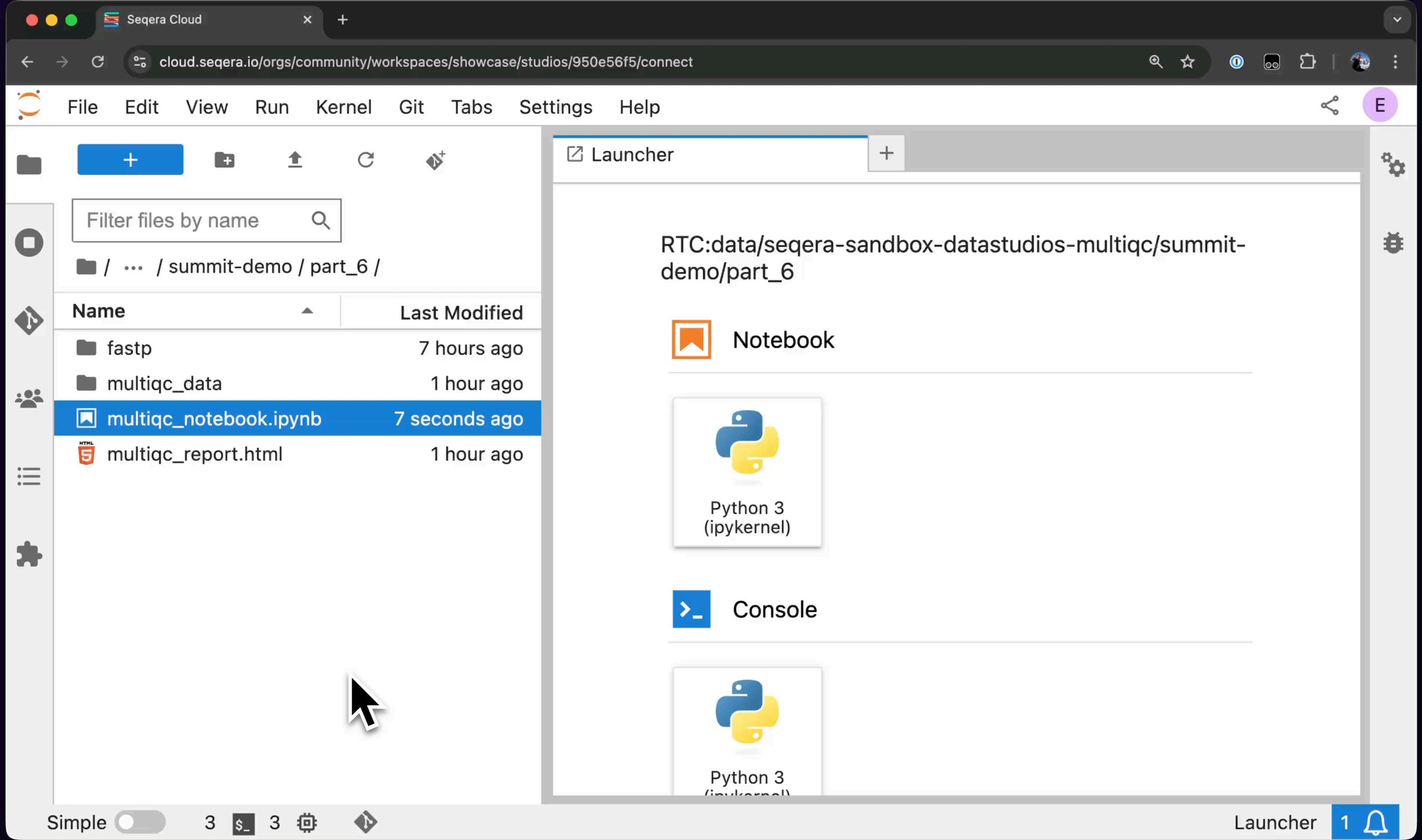
Welcome! Not sure where to start? Watch a tutorial video *(6:06)* don't show again

General Statistics

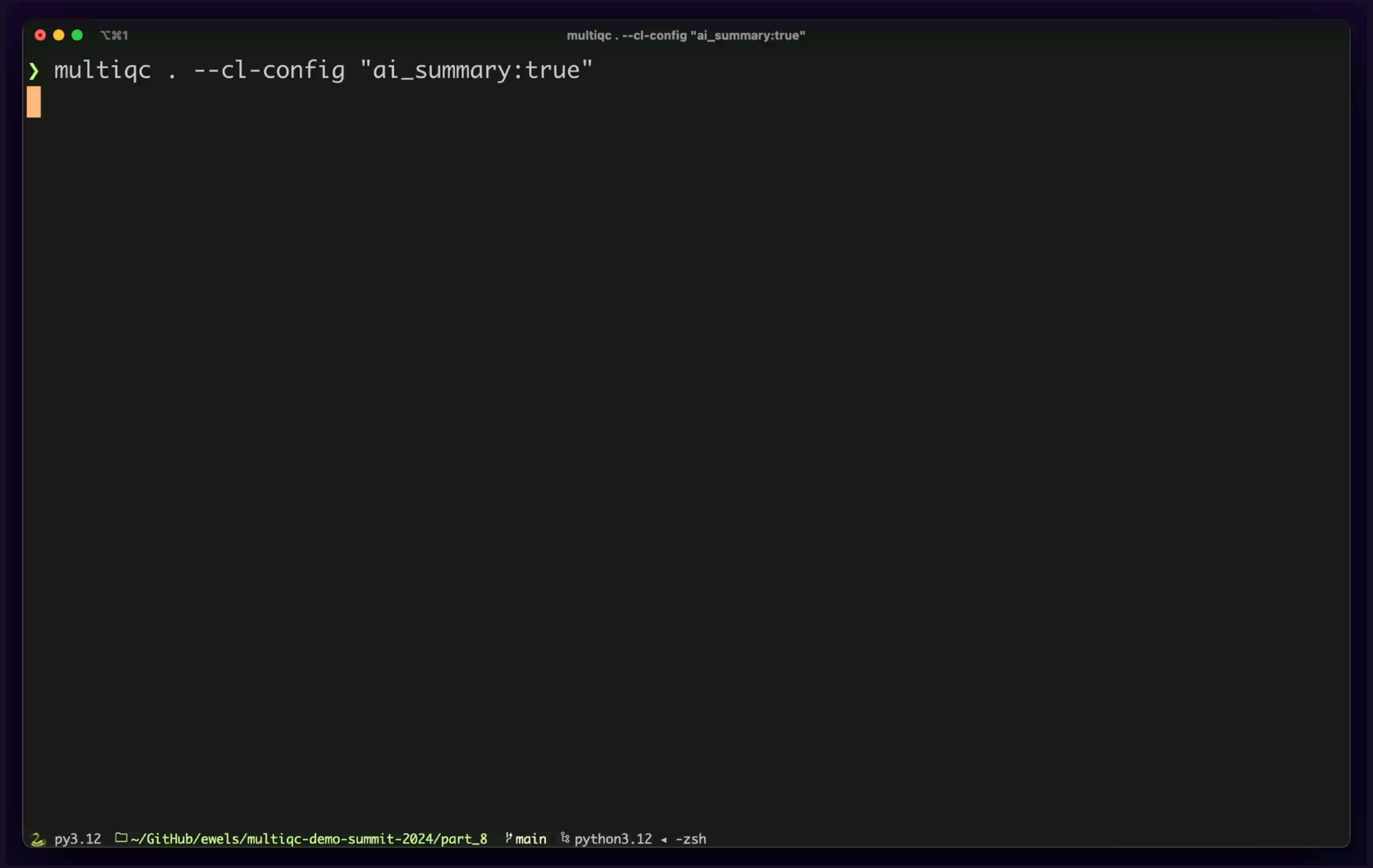
Copy table | Configure columns | Scatter plot | Violin plot | Showing 0/48 rows and 7/9 columns. | Export as CSV

Sample Name	% Duplication	Reads After Filtering	GC content	% PF	% Adapter	Input DNA (ng)	Sample Origin
SAMPLE_01	17.2 %	1.2 M	51.9 %	57.1 %	9.0 %	204	Spain
SAMPLE_02	46.2 %	1.7 M	38.4 %	78.7 %	5.0 %	270	Italy
SAMPLE_03	48.4 %	1.7 M	39.0 %	77.2 %	5.4 %	294	USA
SAMPLE_04	44.0 %	1.5 M	38.4 %	78.9 %	4.8 %	114	Finland
SAMPLE_05	46.0 %	1.7 M	38.5 %	78.5 %	5.0 %	166	Thailand
SAMPLE_06	45.6 %	1.7 M	38.3 %	77.8 %	4.6 %	173	Estonia
SAMPLE_07	48.2 %	2.0 M	38.4 %	79.2 %	5.3 %	147	Germany
SAMPLE_08	48.5 %	1.9 M	38.5 %	79.6 %	5.5 %	220	Lithuania
SAMPLE_09	37.6 %	1.2 M	42.4 %	66.4 %	10.5 %	185	Netherlands
SAMPLE_10	45.4 %	1.6 M	38.3 %	78.9 %	5.1 %	260	Sweden
SAMPLE_11	49.3 %	2.1 M	38.4 %	79.4 %	5.2 %	7	Netherlands
SAMPLE_12	45.2 %	1.6 M	38.3 %	77.5 %	4.5 %	20	Poland
SAMPLE_13	36.9 %	0.1 M	46.0 %	62.4 %	43.7 %	70	Spain
SAMPLE_14	33.7 %	0.4 M	38.0 %	27.5 %	4.5 %	163	Malaysia

Demo


Seqera Cloud
cloud.seqera.io/orgs/community/workspaces/showcase/studios/950e56f5/connect
File Edit View Run Kernel Git Tabs Settings Help
E
Filter files by name
/ ... / summit-demo / part_6 /
Name
Last Modified
fastp 7 hours ago
multiqc_data 1 hour ago
multiqc_notebook.ipynb 7 seconds ago
multiqc_report.html 1 hour ago
Launcher
RTC:data/seqera-sandbox-datastudios-multiqc/summit-demo/part_6
Notebook
Python 3 (ipykernel)
Console
Python 3
Simple
3
3
Launcher
1

MultiQC x ✨


⌥⌘1
multiqc . --cl-config "ai_summary:true"
❯ multiqc . --cl-config "ai_summary:true"
py3.12 ~/GitHub/ewels/multiqc-demo-summit-2024/part_8 main python3.12 -zsh

Thank you

Phil Ewels
phil.ewels@seqera.io

Vlad Savelyev
vladislav.savelyev@seqera.io

Contributors 229

Special thanks to Josh Chorlton and Ruben Vorderman

SUMMIT 2024