From Bench to Big Data: What's new with MultiQC and Nextflow

Phil Ewels

Senior Product Manager for OSS @ Seqera

Background

PhD in the lab, epigenetics

LAB

Postdoc, core bioinformatics

INFRA

Software development

INDUSTRY

Background

UNIVERSITY OF CAMBRIDGE

Babraham Institute

LAB

Postdoc, core bioinformatics

INFRA

Software development

INDUSTRY

Background

UNIVERSITY OF CAMBRIDGE

Babraham Institute

LAB

SciLifeLab

NATIONAL GENOMICS INFRASTRUCTURE

INFRA

Software development

INDUSTRY

Background

UNIVERSITY OF CAMBRIDGE

Babraham Institute

LAB

SciLifeLab

NATIONAL GENOMICS INFRASTRUCTURE

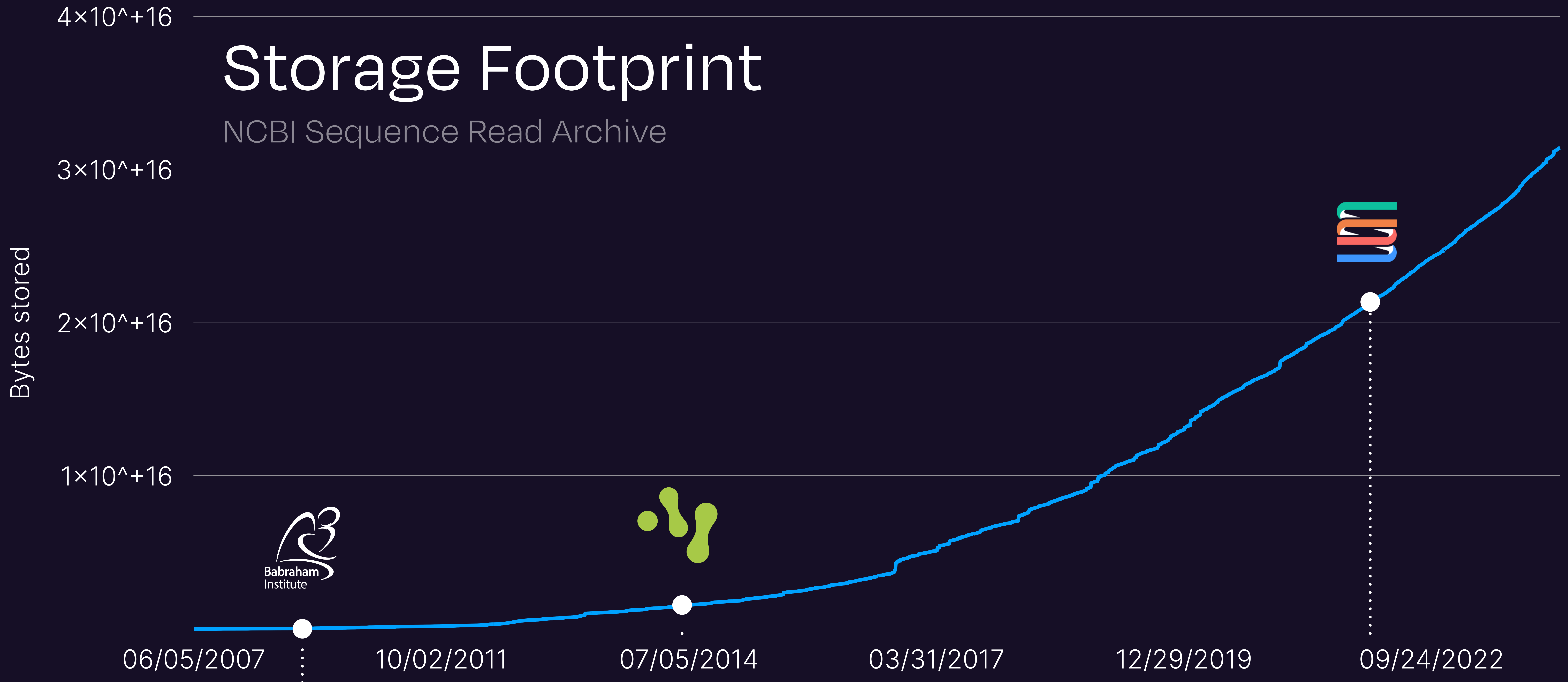
INFRA

seqera

INDUSTRY

Storage Footprint

NCBI Sequence Read Archive



LAB

INFRA

INDUSTRY

seqera

Advancing Science For Everyone Through Software

Open Source Software

</> BUILD

Open Science Resources

DISCOVER

Science Infrastructure

SCALE

Product Manager for
Open Source Software

nextflow

multiqc

wave

fusion

Steering committee,
Core team

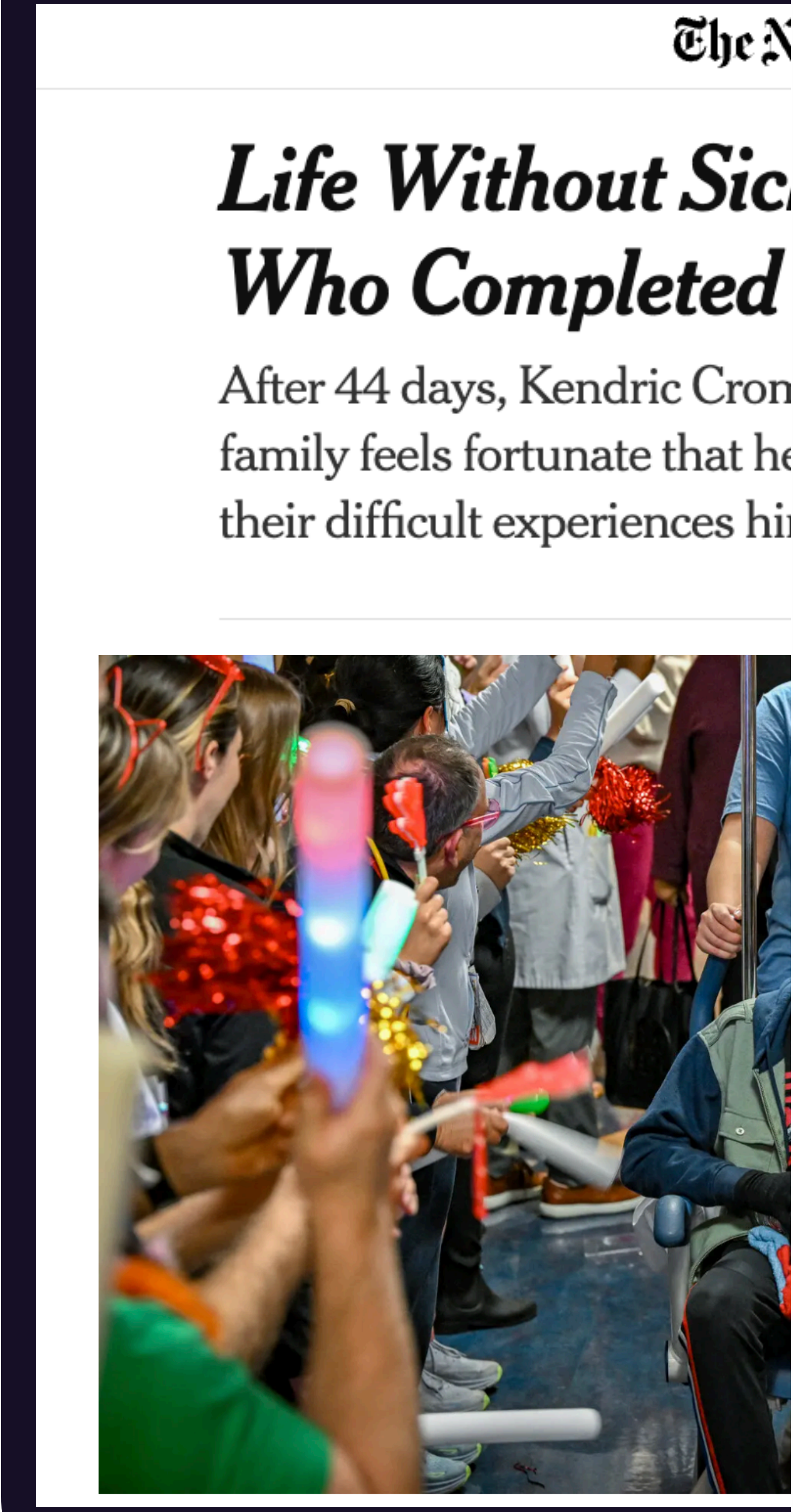
nf-core

Trustee

ossf

VERTEX

OUR WHY



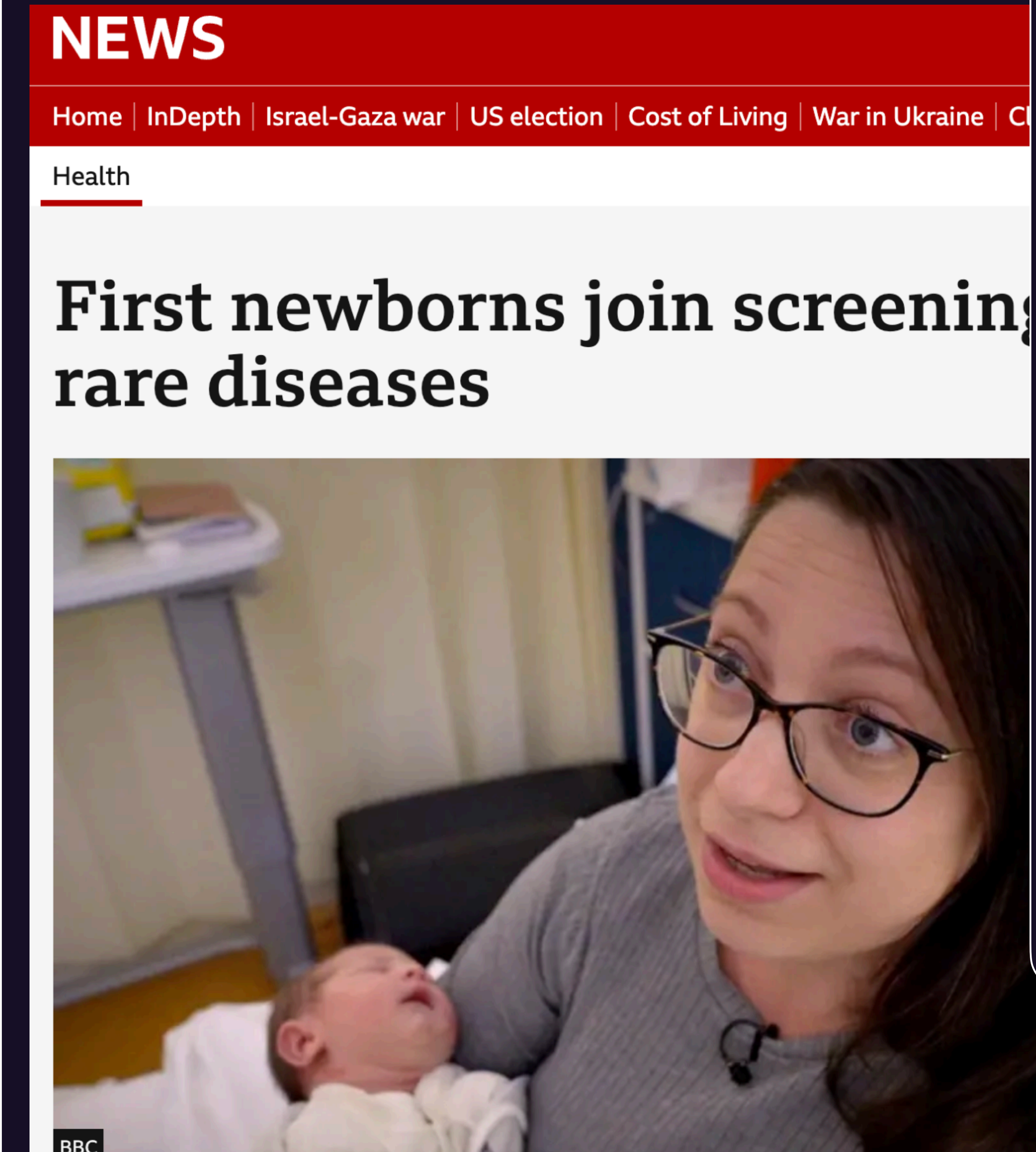
The N

Life Without Sic
Who Completed

After 44 days, Kendric Cron
family feels fortunate that he
their difficult experiences hi

Genomics England

OU



NEWS

Home | InDepth | Israel-Gaza war | US election | Cost of Living | War in Ukraine | Cl

Health

First newborns join screening
rare diseases

BBC

moderna

OUR WHY

‘Real hope’ for cancer cure as personal mRNA vaccine for melanoma trialled

Excitement among patients and researchers as custom-built jabs enter phase 3 trial

A nurse prepares to give Steve Young, one of the first patients in the trial, his first jab at UCLH in London. Photograph: Jordan Pettitt/PA

multiqc

Open-source reporting and analytics

x

multiqc

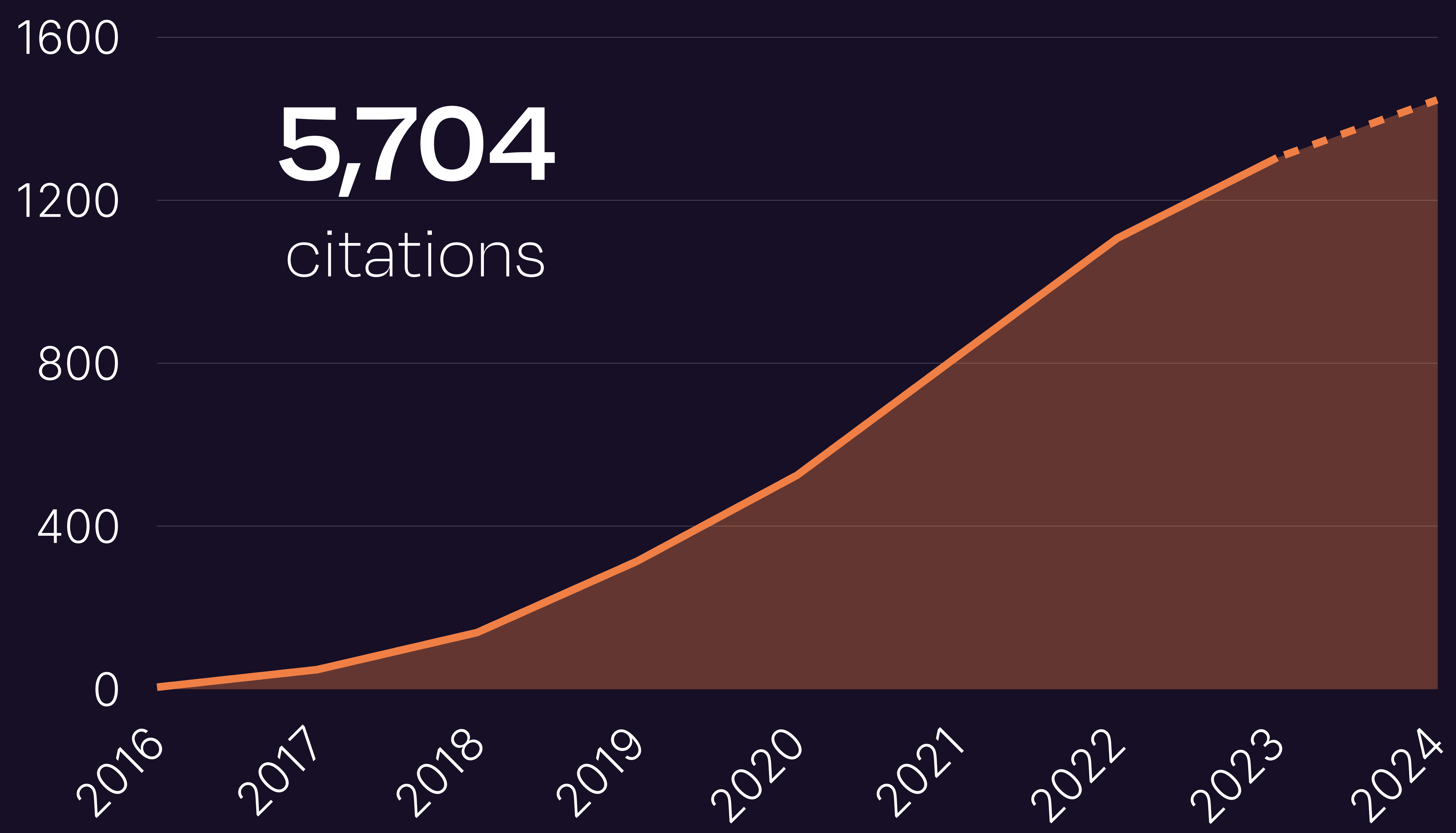
1,216
GitHub Stars

+25K
Runs per day

+1.5M
Downloads

x

Citations by year



Command Line Interface

Web interface

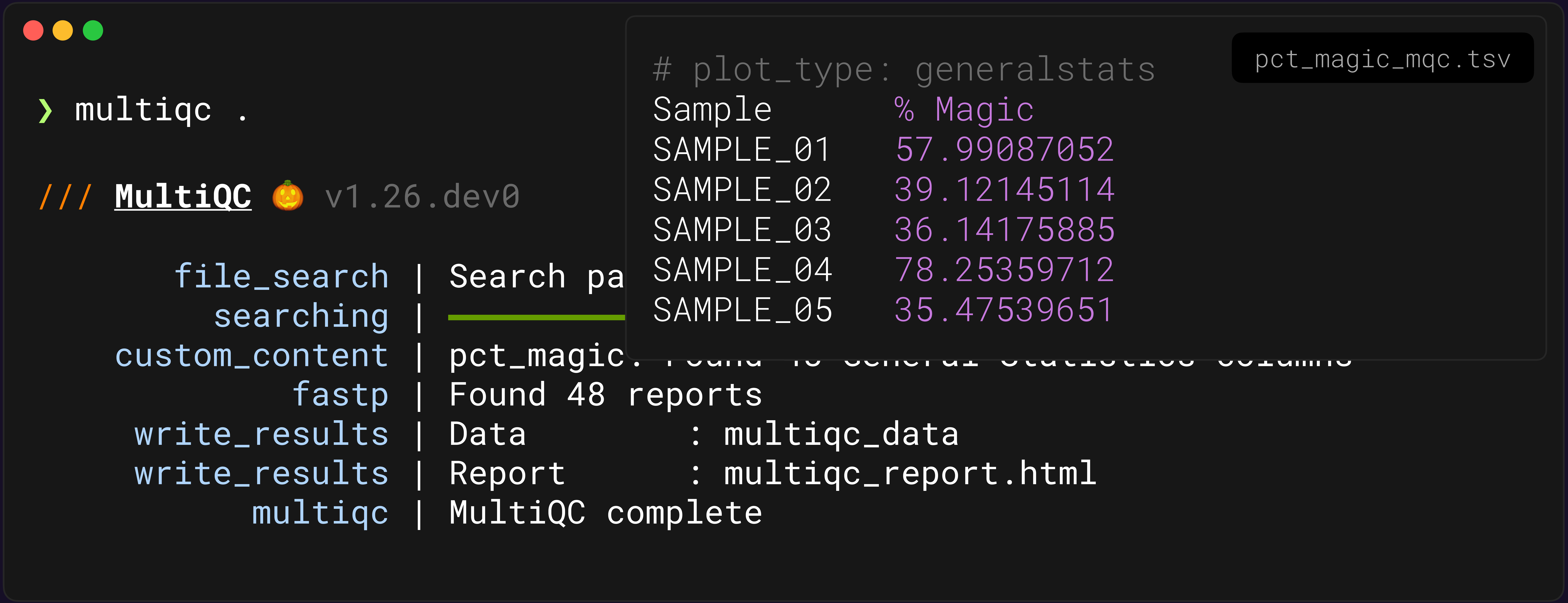
MultiQC Plugins

Custom Content

Notebooks and Scripts

Demo

https://seqera.io/multiqc/

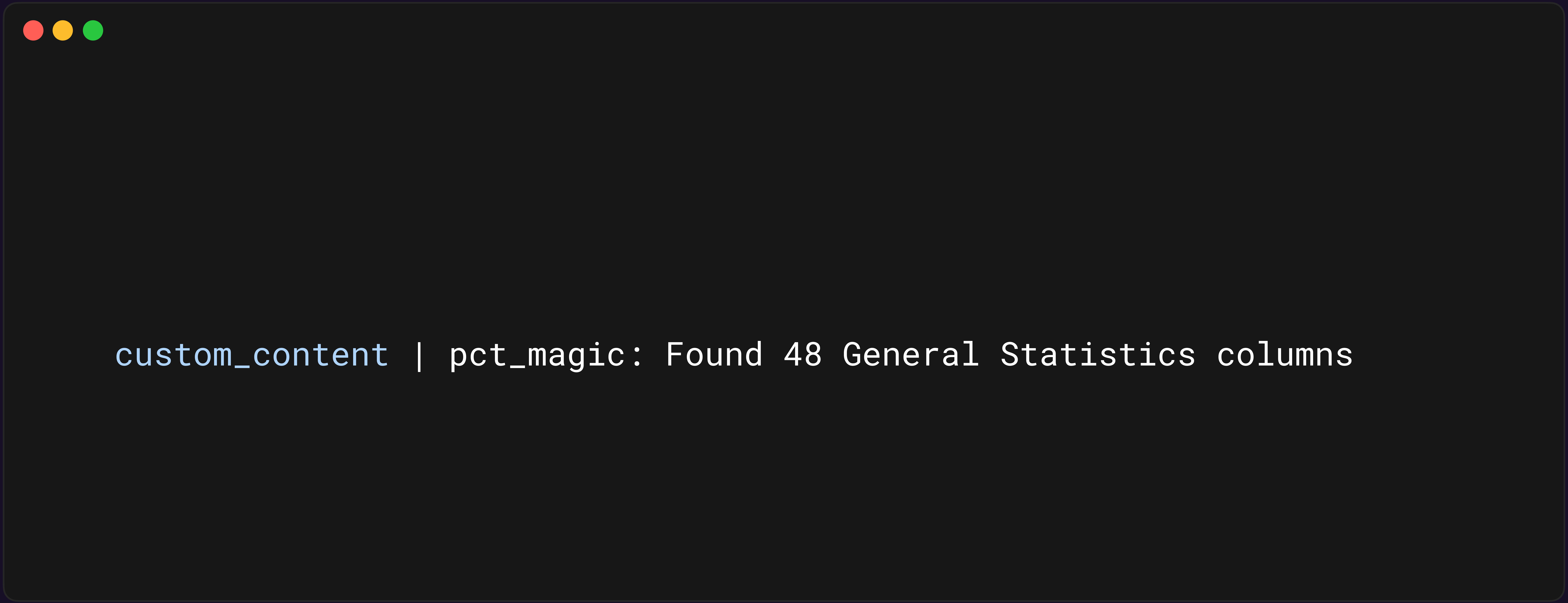


```
❯ multiqc .

/// MultiQC 🎃 v1.26.dev0

       file_search | Search pa
         searching | ━━━━━━━━━
    custom_content | pct_magic: Found 10 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

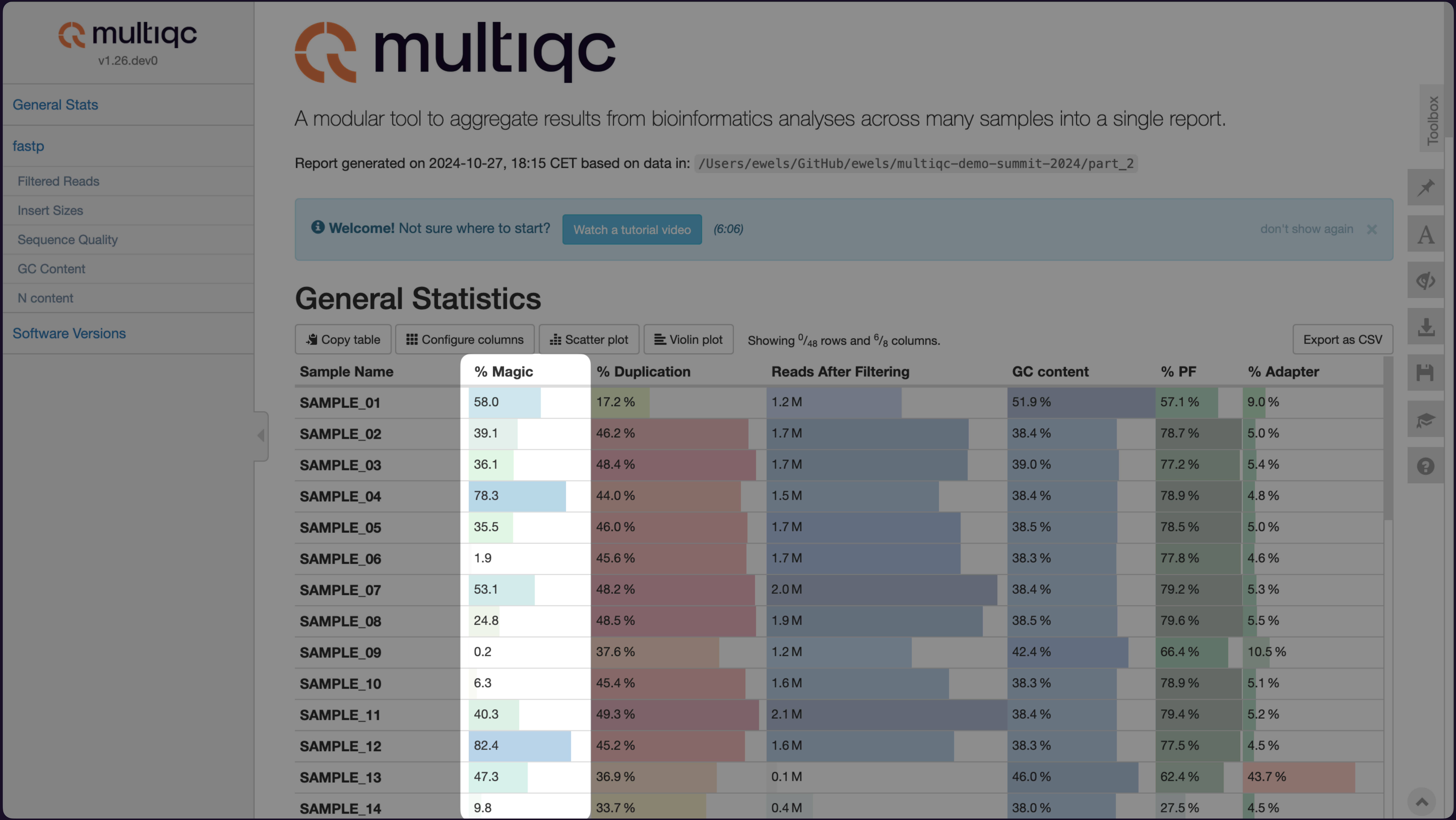
```
# plot_type: generalstats
Sample       % Magic
SAMPLE_01    57.99087052
SAMPLE_02    39.12145114
SAMPLE_03    36.14175885
SAMPLE_04    78.25359712
SAMPLE_05    35.47539651
```

```
fastp
fastqc
pct_magic_mqc.tsv
```



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

pct_magic_mqc.tsv



General Stats

fastp

- Filtered Reads
- Insert Sizes
- Sequence Quality
- GC Content
- N content

Software Versions

multiqc

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

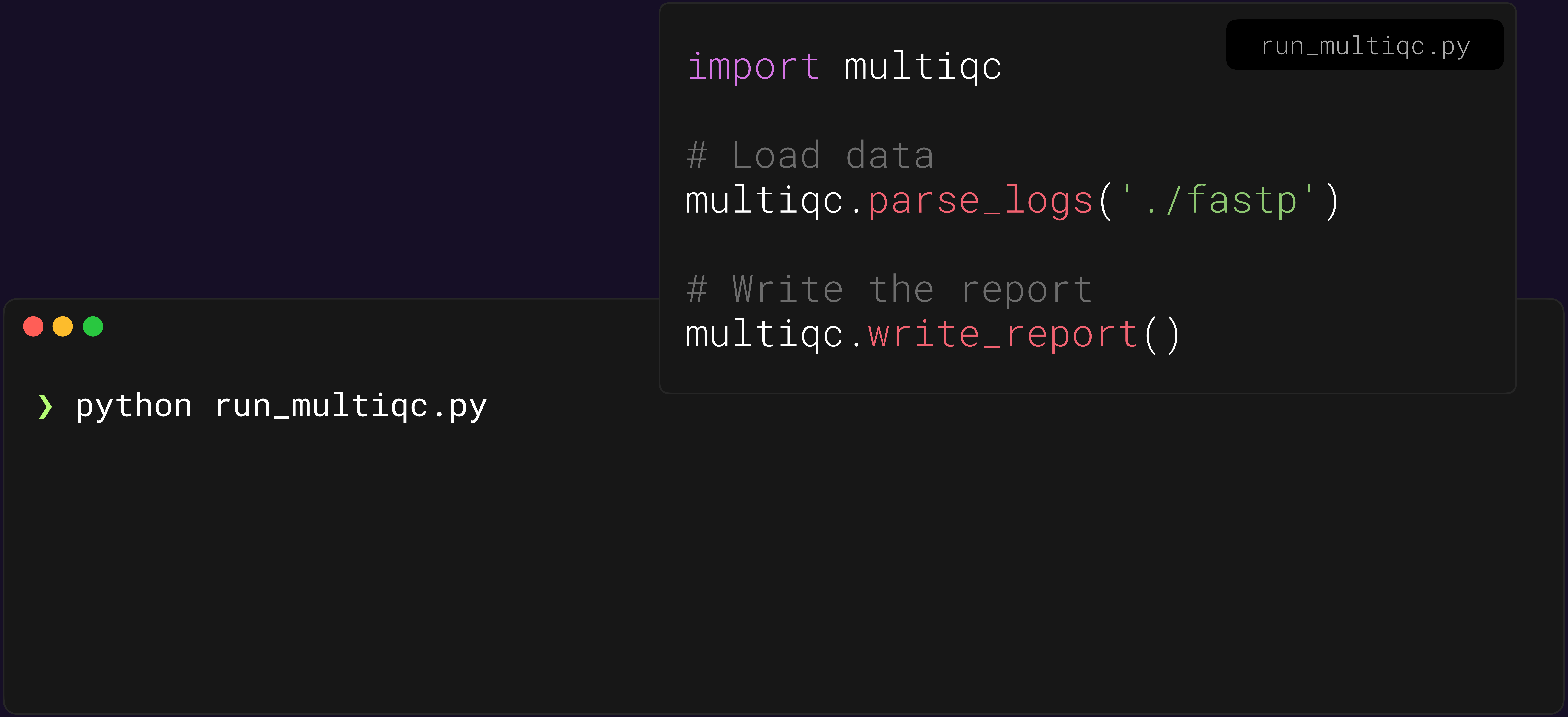
Report generated on 2024-10-27, 18:15 CET based on data in: `/Users/ewels/GitHub/ewels/multiqc-demo-summit-2024/part_2`

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 %



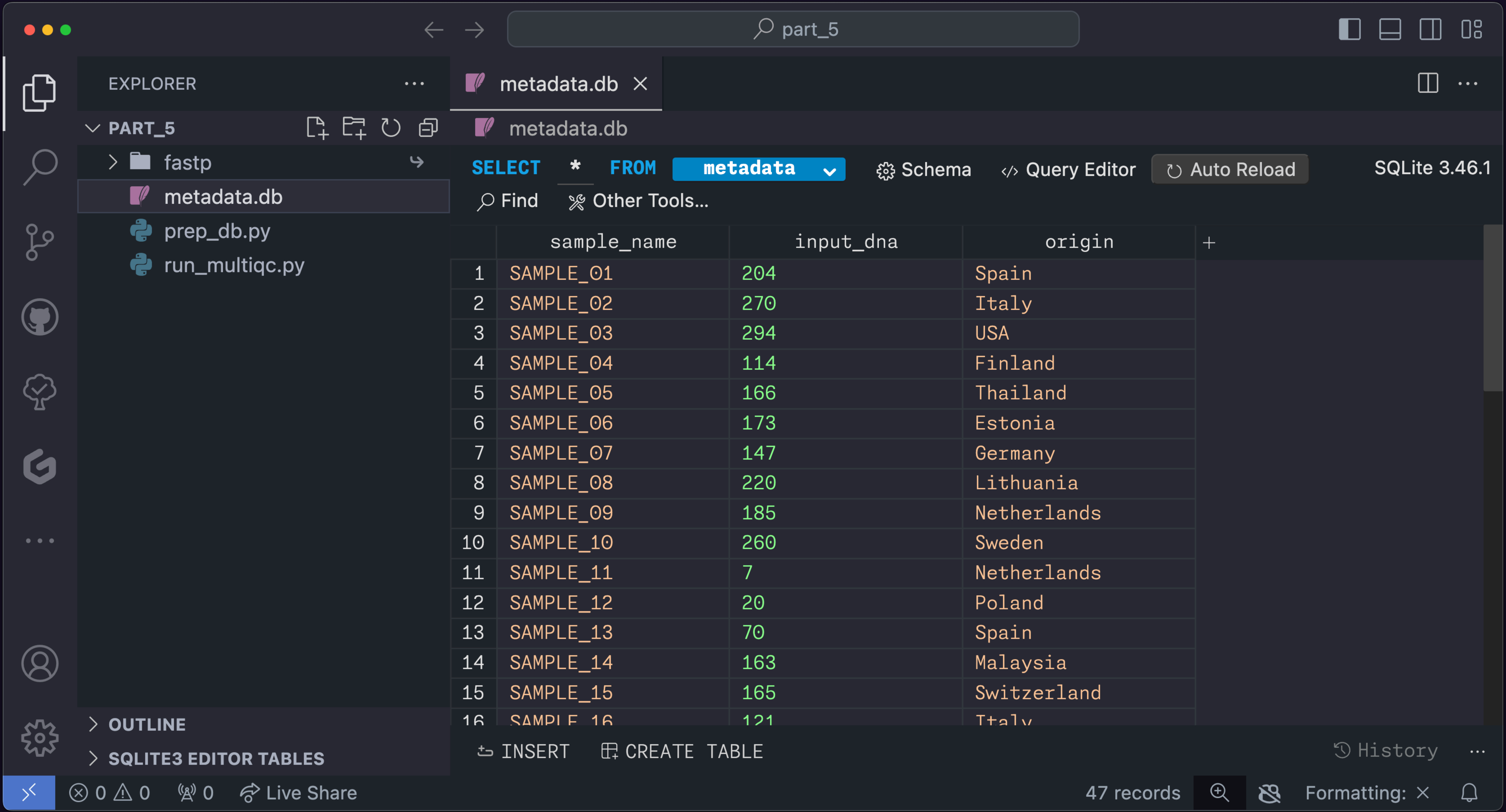
run_multiqc.py

```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)
```



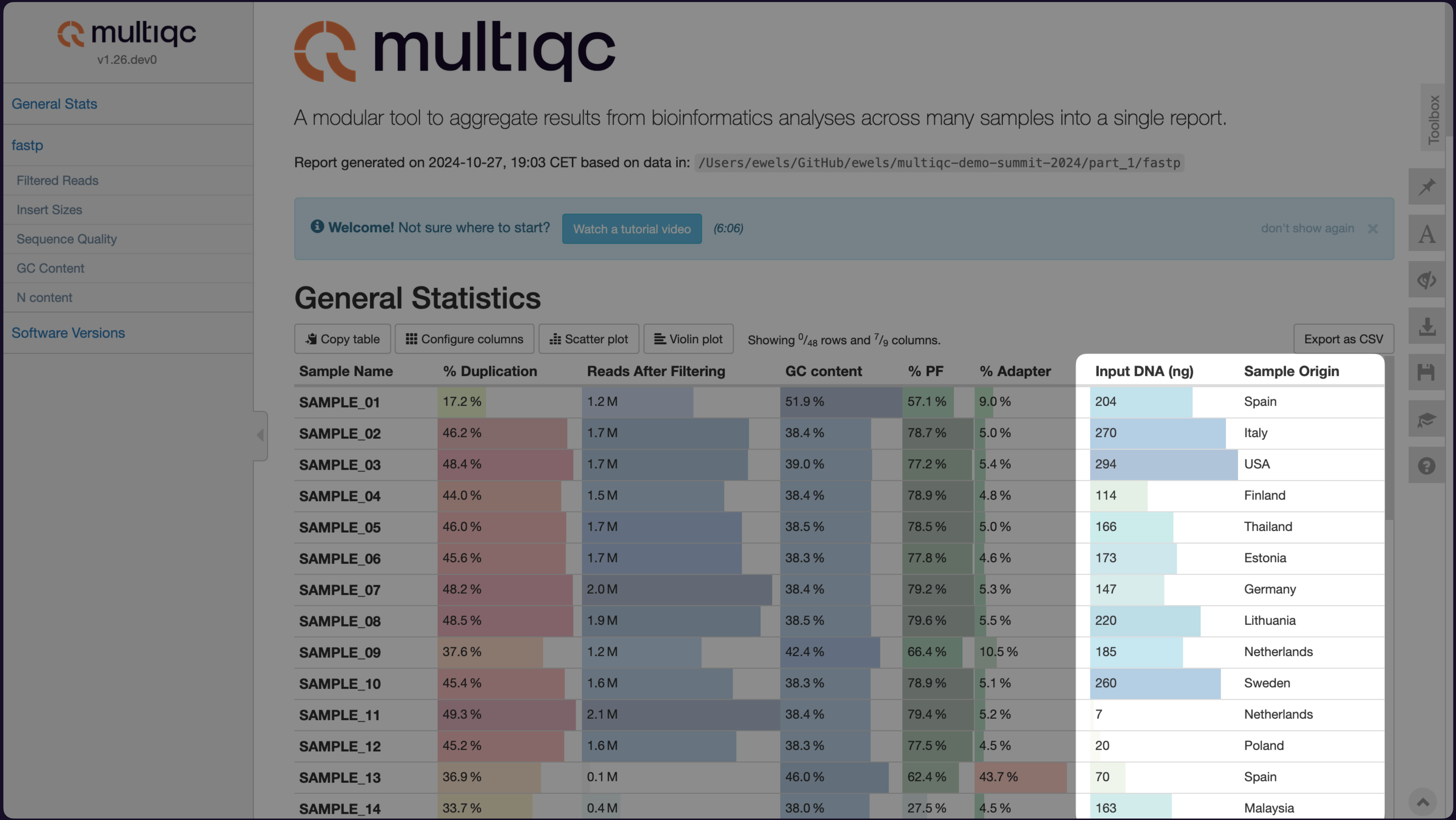
	sample_name	input_dna	origin
1	SAMPLE_01	204	Spain
2	SAMPLE_02	270	Italy
3	SAMPLE_03	294	USA
4	SAMPLE_04	114	Finland
5	SAMPLE_05	166	Thailand
6	SAMPLE_06	173	Estonia
7	SAMPLE_07	147	Germany
8	SAMPLE_08	220	Lithuania
9	SAMPLE_09	185	Netherlands
10	SAMPLE_10	260	Sweden
11	SAMPLE_11	7	Netherlands
12	SAMPLE_12	20	Poland
13	SAMPLE_13	70	Spain
14	SAMPLE_14	163	Malaysia
15	SAMPLE_15	165	Switzerland
16	SAMPLE_16	121	Italy

run_multiqc.py

```python
# 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? (6:06)

General Statistics

Showing $^{0}/_{48}$ rows and $^{7}/_{9}$ columns.

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

nextflow

Open-source orchestrator for deploying workflows

Nextflow DSL limitations

Nextflow was created as an extension of the Groovy programming language.

Missing a formal grammar and syntax parser.

Too fragile. Poor syntax error detection and reporting. Lack of tooling.

X

Introducing: Language server & VS Code integration for Nextflow

Roadmap

Bring new parsers into Nextflow CLI

- Better error messages
- Improve the `nextflow inspect` command
- New commands for linting, formatting

Move beyond Groovy syntax

- Type annotations
- Static type checking
- Simpler dataflow syntax

X

The future of code development is AI-driven

How can we generate code that works in bioinformatics?

Seqera AI

Seqera AI

Available today at

seqera.io/ask-ai

Convert any bash / language script to Nextflow

AI error debugging and code-testing

Rooted in Nextflow and best practices

Discover

Nextflow Summit 2024

(You missed it, sorry)

https://summit.nextflow.io

https://youtube.com/@Nextflow

Thank you

Phil Ewels

phil.ewels@seqera.io

SUMMIT 2024

https://summit.nextflow.io

https://seqera.io