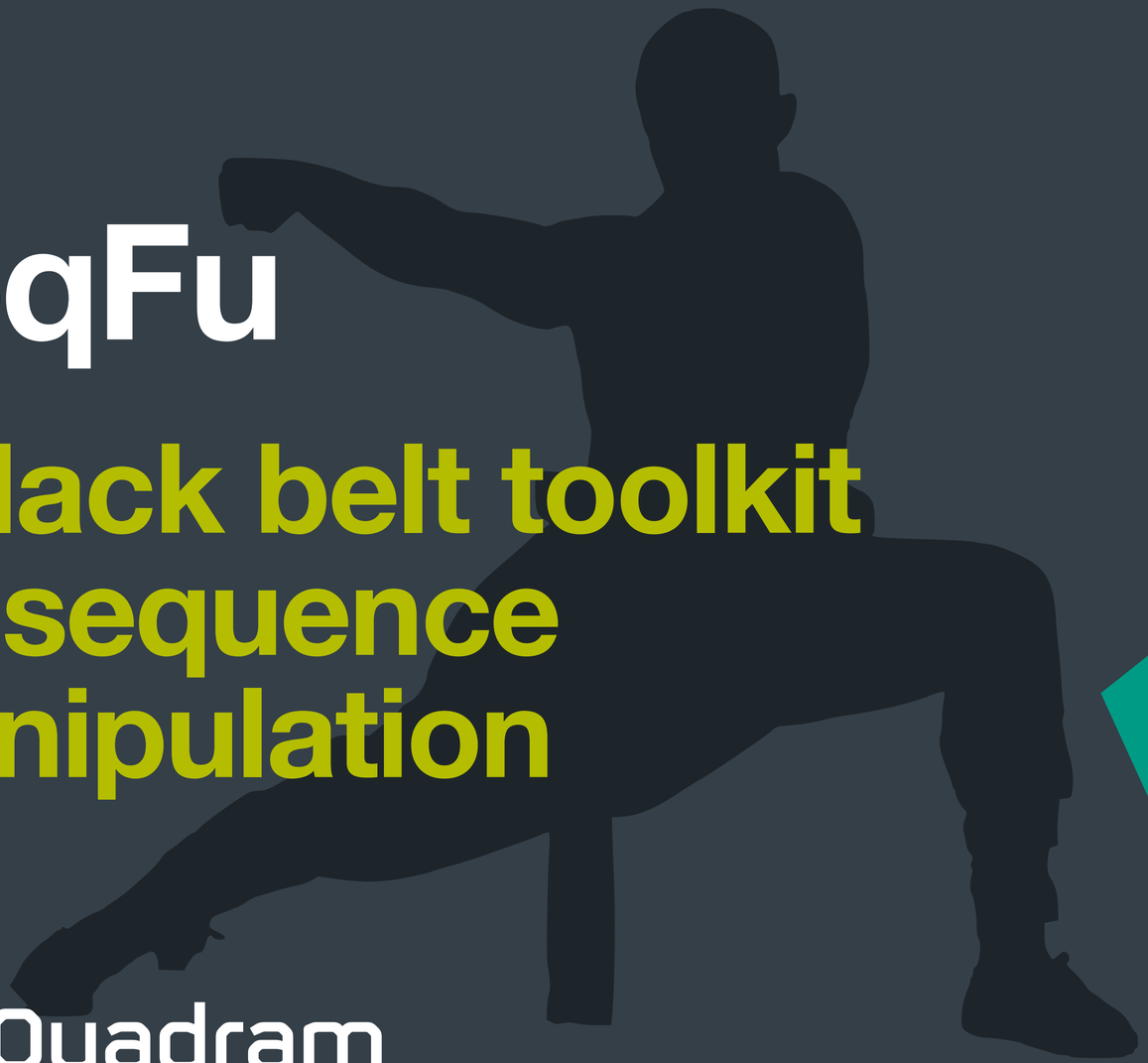
SeqFu

a black belt toolkit for sequence manipulation

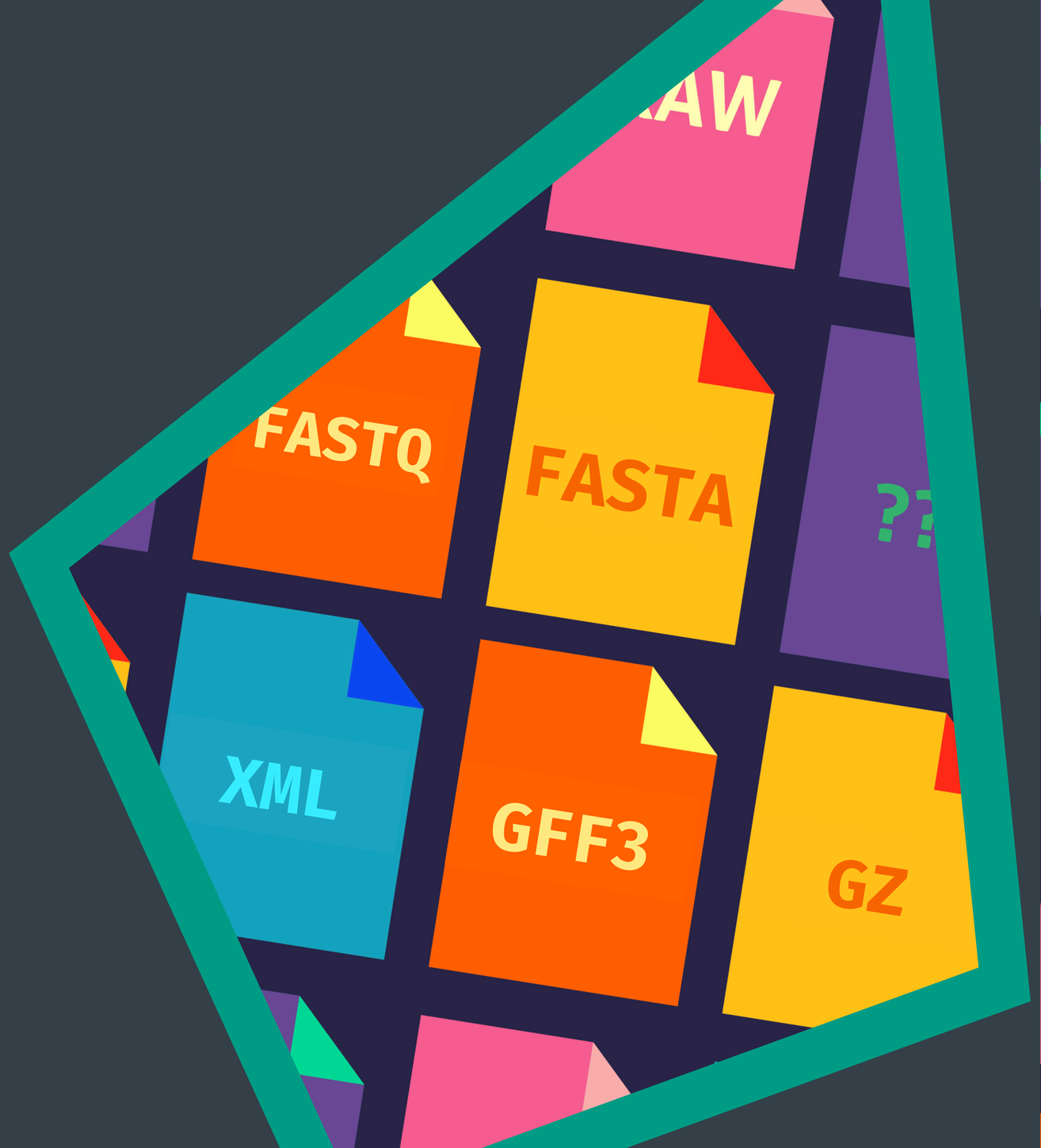


Quadram Institute

Science · Health · Food · Innovation

Andrea Telatin

FASTA.gz and FASTQ.gz



Quadram Institute
Science Health Food Innovation

FASTA

- FASTA to store one or more sequences (DNA, RNA or Protein)
- Determined (assumed correct)

```
>SeqOne
CAGATACGTACGTAGCTTATCTAGA
CGACTGACTGTACGTACGTAGCAGT
ACGTACGTAGCTGATCTACGTACGTAG
ACGT
>SeqFu
CAGATACGTACGTAGCTTATCTAGCAC
CGACTGACTNA
```

FASTQ

- FASTQ to store raw sequencing output
- Associate a quality score to each base

```
@Read1
CAGATACGTACGTAGCTTATCTAGCTACGATCTACTAGC
+
FEFACFFFFFFF,FFF:FFFFFFFFFFFFFFFFFFFF,FFF
@Read2
GCTTATCTAGCTACGATCTACTAGCTGAACACTCGACAGCNT
+
DGGGDGDGDEGGC*;@BGG?EFFECCG<BGGFFFFFDBFF!8
```

Quadram Institute
Science Health Food Innovation

Manipulation from the command line

- There are plenty of tasks we need to perform daily on FASTA or FASTQ files:

Inspection - peeking into files, checking integrity, oligo/primers...

Subsetting - getting some reads from a large dataset

Summary - calculating statistics on the assembly/run

You name it: dereplication, reverse complement...

How it started...

Quadram Institute

Science Health Food Innovation

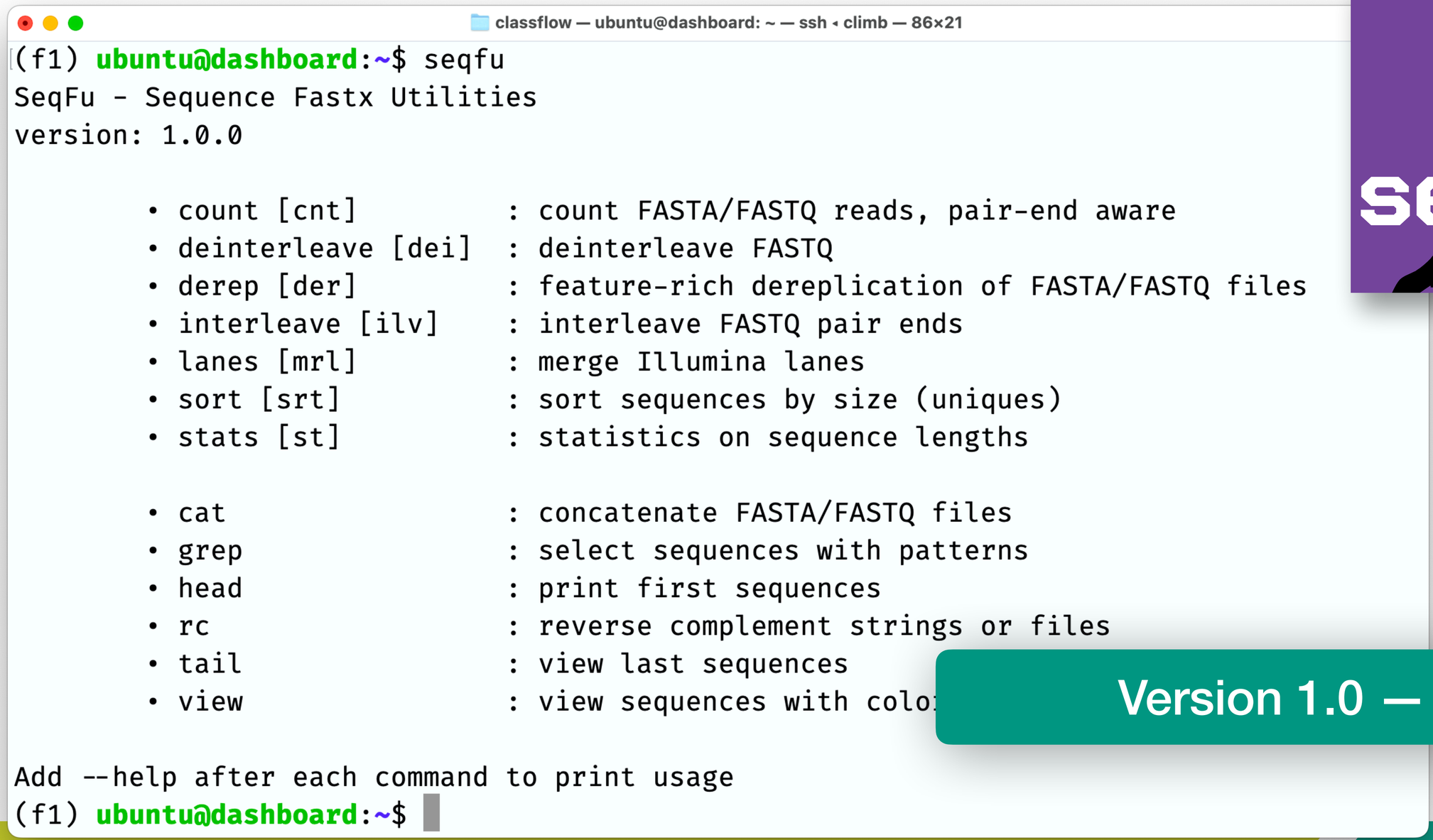
Daily issues: an example

A run has been **demultiplexed** with the **wrong** sample sheet:

- Each sample received the wrong filename
- It was needed to extract the index from the file content

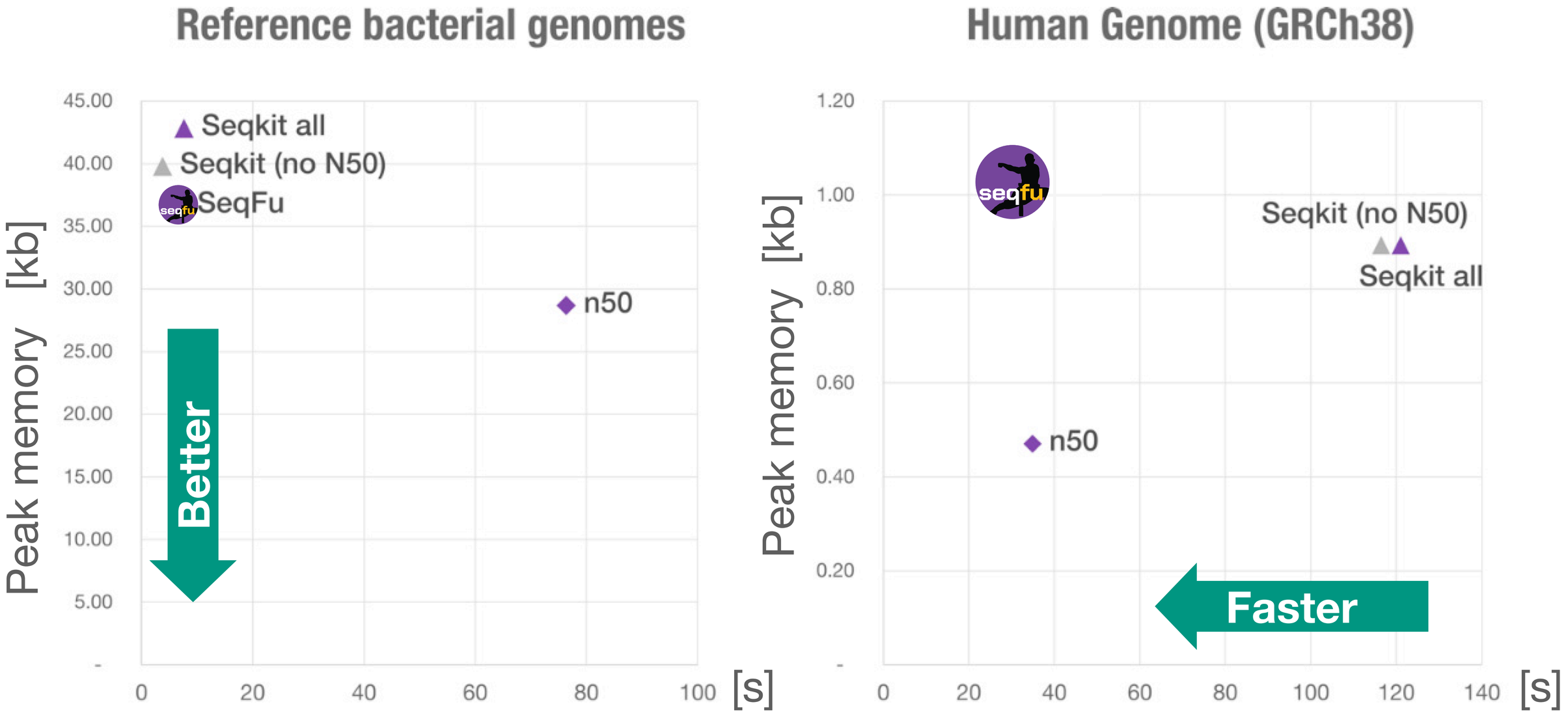
And many more atomic tasks...

SeqKit didn't always have the answer



Version 1.0 — 2021

Quadram Institute
Science Health Food Innovation



Calculation of N50 and other statistics using two datasets

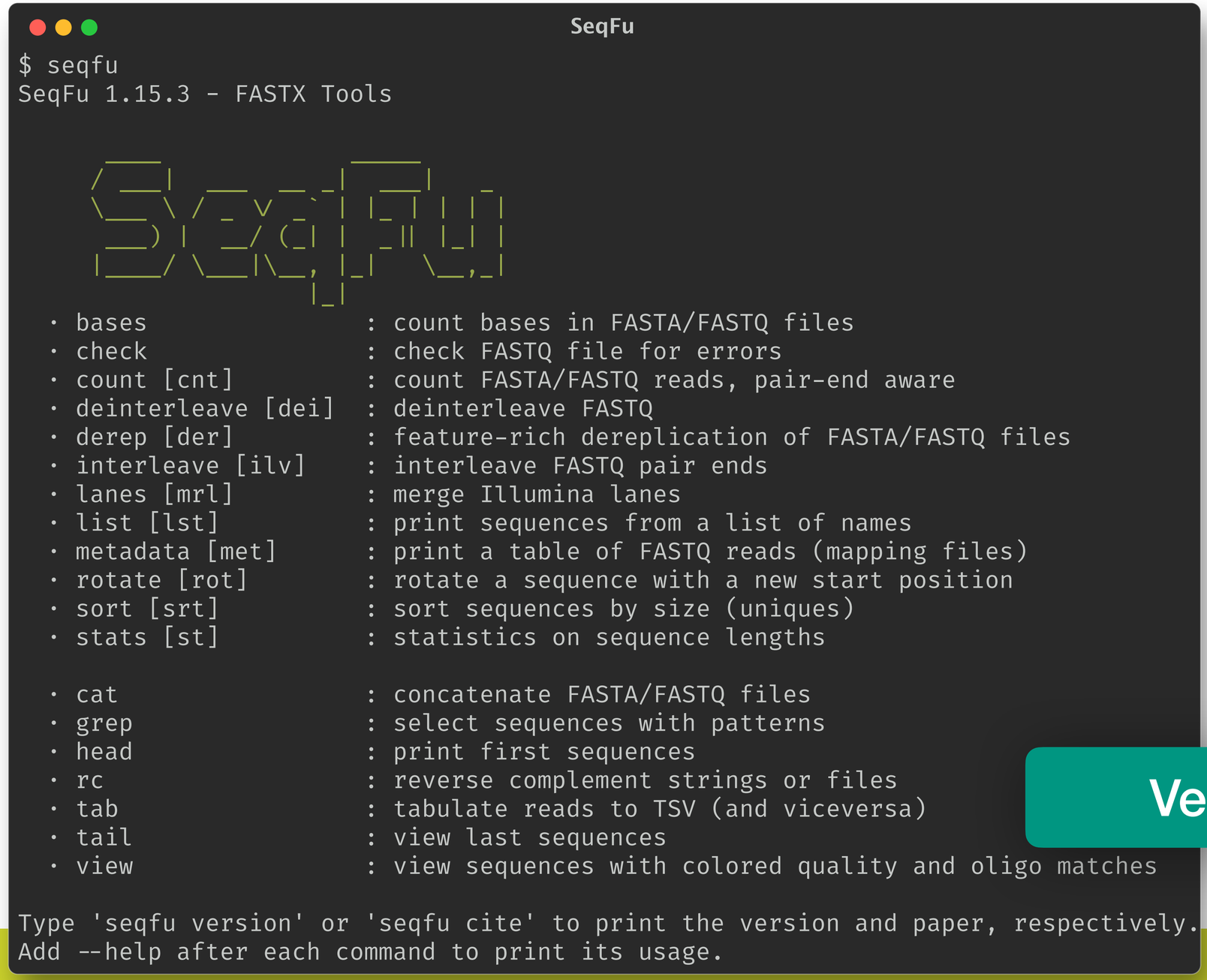
Telatin *et al.* 2021

Quadram Institute
Science Health Food Innovation

How it is going...

Quadram Institute

Science Health Food Innovation



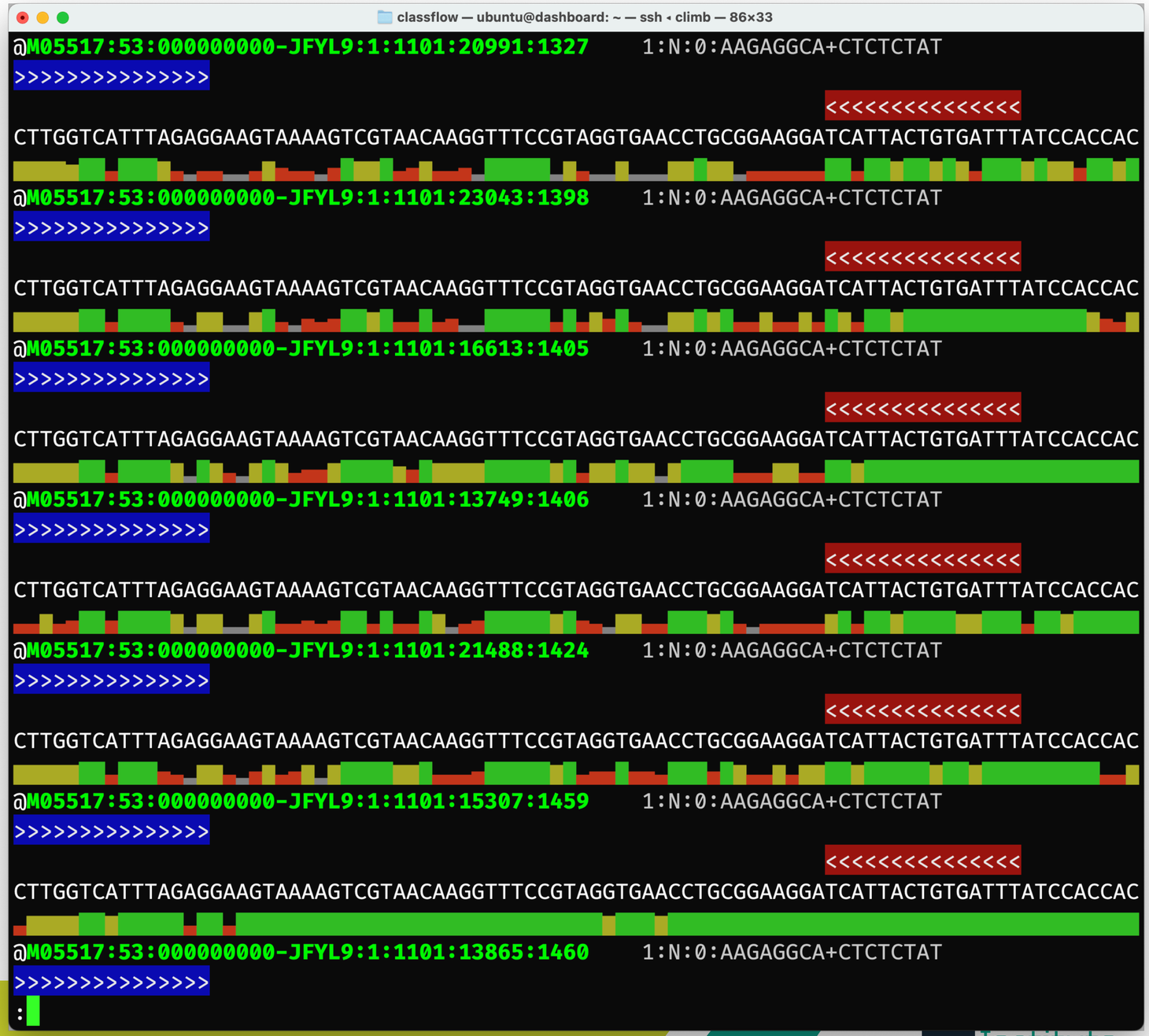
- **19 core functions** in SeqFu 1.17
- **14 utilities** shipped as independent binaries
- ⭐ **>70 stars** on GitHub, and good user feedback

Version 1.17 — Dec 2022

Quadram Institute
Science Health Food Innovation

seqfu view

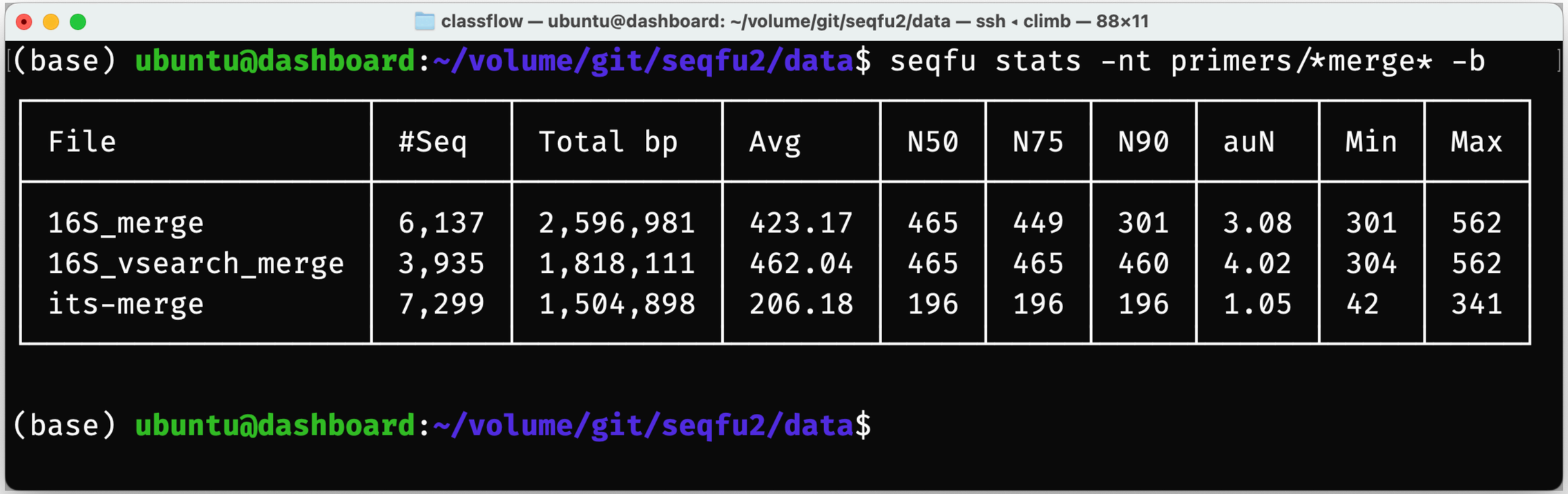
- quickly inspect a FASTQ file, with a coloured representation of qualities
- Degenerated oligo match in both strands supported



Institute
Science Health Food Innovation

seqfu stats

- Get statistics including N50 from FASTA/FASTQ files
- Nice terminal tables are supported
- Can export in MultiQC format



Quadram Institute
Science Health Food Innovation

Good practices

- Documentation
- Streams support
- Automatic tests
- Distribution

Quadram Institute
Science Health Food Innovation

Documentation

- Extensive online documentation
- Clear help screens from the commands

telatin.github.io/seqfu2

seqfu.it

SeqFu docs

SeqFu 1.16.0 — crayon

Search docs...

1. Installation
2. Overview
3. Short guide
4. About SeqFu

CORE TOOLS

seqfu bases
seqfu cat
seqfu check
seqfu count
seqfu deinterleave
seqfu derep
seqfu grep
seqfu head
seqfu interleave
seqfu lanes
seqfu list
seqfu merge
seqfu metadata
seqfu qual
seqfu rc
seqfu rotate
seqfu sort

seqfu view

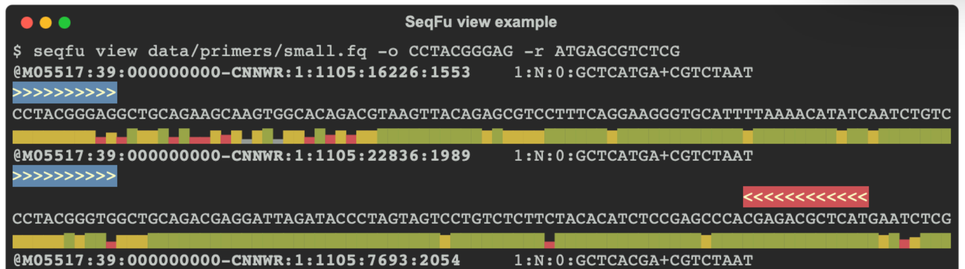
view is one of the core subprograms of *SeqFu*.

It can be used to visually inspect a FASTQ file printing colored bars for quality scores and highlighting oligonucleotide matches.

```text
Usage: view [options] <inputfile> [<input_reverse>]

View a FASTA/FASTQ file for manual inspection, allowing to search for
an oligonucleotide.

Options:
  -o, --oligo1 OLIGO     Match oligo, with ambiguous IUPAC chars allowed
                         (rev. compl. search is performed), color blue
  -r, --oligo2 OLIGO     Second oligo to be scanned for, color red
  -q, --qual-scale STR   Quality thresholds, seven values
                         separated by columns [default: 3:15:25:28:30:35:40]

  --match-ths FLOAT      Oligo matching threshold [default: 0.75]
  --min-matches INT      Oligo minimum matches [default: 5]
  --max-mismatches INT   Oligo maxmimum mismataches [default: 2]
  --ascii                Use simple ASCII chars instead of UNICODE to
                         render the quality values
  -Q, --qual-chars       Show quality characters instead of bars
  -n, --nocolor          Disable colored output
  --verbose              Show extra information
  -h, --help             Show this help
```

Example output

The quality scores are rendered as colored bars (grey, red, yellow, green) of differen
Matching oligos are rendered as blue arrows (forward) or red arrows (reverse).



Quadram Institute
Science Health Food Innovation

data — telatina@N121515: ~/git/seqfu2/data — bash — 91×12

```
telatina@VM:~/git/seqfu2/data$ seqfu cat --min-len 100 --trim-front 10 dataset.fastq.gz | \
     seqfu head -n 100 | \
     seqfu stats -n
[seqfu head] Waiting for STDIN ... [Ctrl-C to quit, type with --help for info].
[seqfu stats] Waiting for STDIN ... [Ctrl-C to quit, type with --help for info].
```

File	#Seq	Total bp	Avg	N50	N75	N90	auN	Min	Max
-	100	24100	241.00	241	241	241	2.41	241	241

```
telatina@VM:~/git/seqfu2/data$
```

Pipes

- Input/Output streams support to combine multiple tools together

Quadram Institute
Science Health Food Innovation

Automated tests

- Unit tests (164)
- **Continuous integration**: all tests run on the cloud at every change in the code

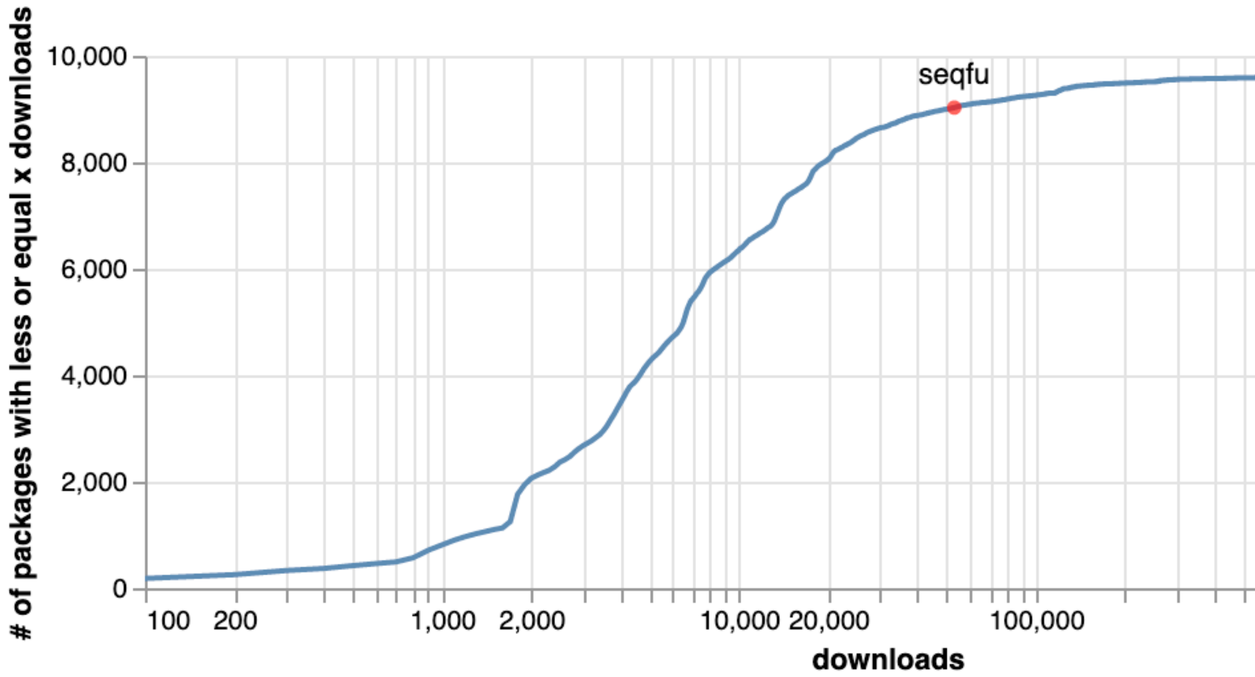
classflow — ubuntu@dashboard: ~/volume/git/seqfu2 — ssh ◂ climb — 104×26

```
  Testing module: stats
──────────────────────────────────────────────────────────────────────────────
OK: Checking normal output expecting 2 lines: <2>
OK: Checking normal output expecting total seqs 78730: <78730>
OK: Checking normal output expecting total bases 24299931: <24299931>
OK: Checking normal N50 to be 316: <316>
OK: Checking CSV output N50 is 316, got: <316>
OK: Checking nice output expecting 5 lines: <5>
OK: Checking MultiQC output expecting 39 lines: <39>
OK: Experimental JSON output on 1 line: <1>
OK: Checking default starting  by 'filt': <filt>
OK: Checking default N50 starting by 'sort': <sort>
OK: Check absolute paths
OK: Check sort orded when not specified
OK: Check sort: tot seq sorted at 3300
OK: Check reverse sort: tot seq sorted at 3300
OK: %GC check at 1.00: <1.00>
OK: %GC check at 0.00: <0.00>
OK: %GC check at 0.50: <0.50>
# Finished with 17 passed, 0 failed

  Testing module: tab
──────────────────────────────────────────────────────────────────────────────
OK: FASTQ sequence tabulated / detabulated
OK: FASTA sequence tabulated / detabulated
OK: FQ interleaved sequence tabulated / detabulated
```

Quadram Institute
Science Health Food Innovation

BioConda

- Available for **easy** installation from Bioconda
- On of the top downloaded packages in the BioConda repository




```
conda install -c conda-forge -c bioconda seqfu
```

Quadram Institute
Science Health Food Innovation

Acknowledgments

Andrew Page team (Thanh, Sumeet,...)

QIB users for feedback
(Evelien, Stefano, Rebecca,...)

Thank you!

for your attention

Quadram Institute
Science Health Food Innovation

Twitter
@SeqFu2

Links

- SeqFu Paper mdpi.com/2306-5354/8/5/59
- SeqFu repository github.com/telatin/seqfu2
- Info on Core Bioinformatics qintranet.com/bioinformatics/